PORTABLE TEXT TO SPEECH DEVICE FOR VISUALLY IMPAIRED USING RASPBERRY PI AND WEBCAM

[1]Siva Rama Krishna, [2]Sourabh Kumar, [3]Jayendra Kumar
[1]UG research scholar, [2]UG research scholar, [3]Assistant Professor
[1,2,2] Department of Electronics and Communication Engineering,
[1,2,3]National Institute of Technology Jamshedpur, Jamshedpur, India

Abstract: **Communication plays a major role in human life. Vision has one of the utmost importance in making the communication complete. A person needs vision to access information from a text or image. Visually impaired people gather information from voice only. In many situations, information access for those people is very difficult due to their disability. The proposed paper here implements an idea for image conversion to speech with the help of Raspberry Pi, a Web cam and a speaker or an earphone. The proposed idea uses the concept of Tesseract OCR (Optical Character Recognition) to produce speech from image captured by the camera. This helps the visually impaired people to access the information of text in printed materials i.e. books as well as hand-written notes. The implemented prototype will be a portable device since a power bank is used for its battery backup and gives speech output in multiple languages with the help of Speech API and Microsoft translator. This idea can help millions of visually impaired people to get vision experience and helps them a lot for reading, shopping, indoor and outdoor navigation.**

Keywords – **Raspberry Pi, Tesseract OCR, Google Speech API, Microsoft translator, webcam.**

I. Introduction

Reading is an integral part of our daily life. Person who are visually impaired suffer the most in this aspect as their disability restricts them for reading purposes. Although Braille language is quite helpful for them, but with the development of technology and cameras, it is possible to help the blind by developing applications that can use vision tools from computer and Optical Character Recognition algorithm embedded on a portable device and the central core being Raspberry Pi.

The paper proposes a system which consists of Raspberry Pi and a camera to serve the purpose of conversion of text to speech. The system uses Tesseract OCR Optical Character Recognition engine for text extraction. The image processing task is done by the OpenCV library. The proposed system is different from many traditional systems since it can give output voice in any many languages as per the user's interest. The text to speech engine is used for speech synthesis. eSpeak is a speech synthesizer which is used here. eSpeak produces speech in English which can be translated into other languages using google text to speech engine and Microsoft translator.

II. Literature survey

There are many works done already in this field. In [1] explains a text reading system which is camera based for blind people. In the proposed paper, a binary image is produced with the help of global or local thresholding which can be decided from Fisher's Discriminant Rate (FDR). The technique is essentially based on OTSU's binarization method. It is an automatic threshold selection region-based segmentation method. In this method when the characters are present on a frame, then-the local histogram has two peaks and this is reflected as a high value for the FDR. For quasi-uniform frames the value of the FDR is small and the histogram has only one peak. In the case of complex areas, the histogram is dispersed resulting in higher FDR values, which are still lower than in the case of text areas. With a bimodal gray-level histogram the FDR is used to detect the image frames. When the image frames are of high FDR values, the local OTSU threshold is used for binarizing the image, frames with low FDR values Maintaining the Integrity of the Specifications.

The paper [2] presents a prototype for extracting text from images using Raspberry Pi. The images are captured using a web cam and are processed using Open CV and OTSU's algorithm. Initially the captured images are converted to gray scale colour mode. The images are rescaled and cosine transformations are applied by setting vertical and horizontal ratio. After applying some morphological transformations OTSU's thresholding is applied to images which is adaptive thresholding algorithm. After thresholding, contours for the images are generated using special functions in Open CV. Using these contours, bounding boxes are drawn around the objects and text in the images. Using these drawn bounding boxes each and every character present in the image is extracted which is then applied to the OCR engine to recognize the text present in the image.

In [3], a camera based assistive text reading framework to help visually impaired persons read text labels and product packaging from hand-held objects in daily life. The proposed system isolates objects from unclean backgrounds or other surrounding objects in the camera vision. Motion based methods is implemented to define ROI Region of Interest. The object region in motion is extracted using a mixture of Gaussians-based background subtraction technique. Text localization and recognition are implemented to obtain details of text from the ROI and the text regions from the ROI of object are focused simultaneously. In an Adaboost model the ramp features of pixel edge distributions and orientations of stroke are carried out by Novel Text Localization algorithm. Off-the-shelf optical character identification software recognizes the characters present in the text of localized text regions by performing binarization.

Using SWT [4] the text in natural scenes are detected by performing the merging of pixels with similar strike width into connected components which is done by the bottom-up integration Across wide range of scales in the same image, letters are detected by performing this method. It recognizes text lines, strokes of any direction as it won't use a filter bank of orientations which are discrete. Information is well carried out for executing the text segmentation accurately. So, the detected text will have a good mask readily available. The limitations of this method are filter orientations, inherent attenuation to horizontal texts, necessity for integration over scales. The linear features which are related with the stroke definition are used in medical imaging, remote sensing domain. The road width range in a satellite or aerial photo is limited and known in road detection process whereas the text present in an image can vary in scale extremely. Text won't have standard extended linear structures with low curvature as in roads.

III. System architecture

Raspberry Pi 3, webcam, power bank of 10000 mAh capacity, and Bluetooth earphones are the hardware components used for the design of hardware. The webcam captures the images and the captured images are sent to the Raspberry Pi 3 where they are processed. The voice output produced is sent to a wire-less earphone via Bluetooth. The power supply is provided by using a power bank which is rechargeable one.

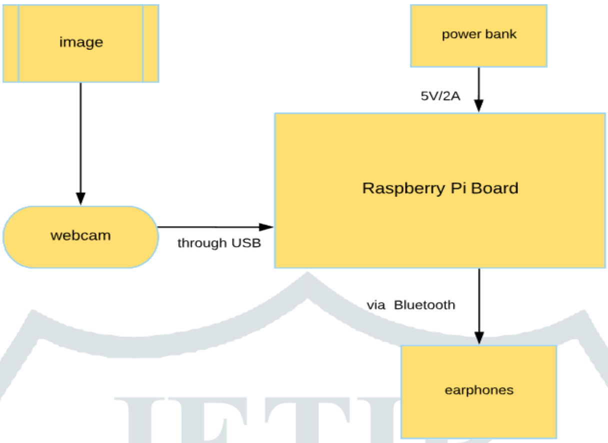


fig.1: Block Diagram of System implementation

A. Raspberry Pi 3 Model

Raspberry Pi is a versatile single board computer of size as small as a credit card which is developed by the Raspberry Pi Foundation [7]. It features a quad-core 64-bit ARM Cortex A53 operating at 1.2GHz. It contains 1 GB of RAM and also supports graphics, provided by Video Core IV GPU. It includes an onboard 802.11n Wi-Fi, Bluetooth 4.0 and has 4 USB Ports and an Ethernet Port. It has got an option for increasing the storage capacity through a microSD card slot and for display using HDMI port. It includes a 40-pin GPIO header and a composite audio/video output slot. Initially, it was developed with the aim to promote the teaching of computer science in educational institutes but the Original model gained much popularity than expected, selling beyond its target market for users such as Robotics. Raspberry Pi needs to be programmed with LINUX programming language to communicate with the environment. Since it is based on LINUX, optimum performance of Raspberry Pi can be achieved if it is therefore operated in the same environment. The most convenient operating system for Raspberry Pi is RASPBIAN as it has over 35,000 packages, pre-compiled software packed together for easy installation. RASPBIAN is the most popular operating system to be used with raspberry pi, as it comes with numerous pre-installed software for educational, programming and general purposes which includes Java, Scratch, Sonic Pi, Python, Mathematica etc. Another commonly used operating System for Raspberry Pi is Units

fig.2(a): Raspberry Pi 3 Model B

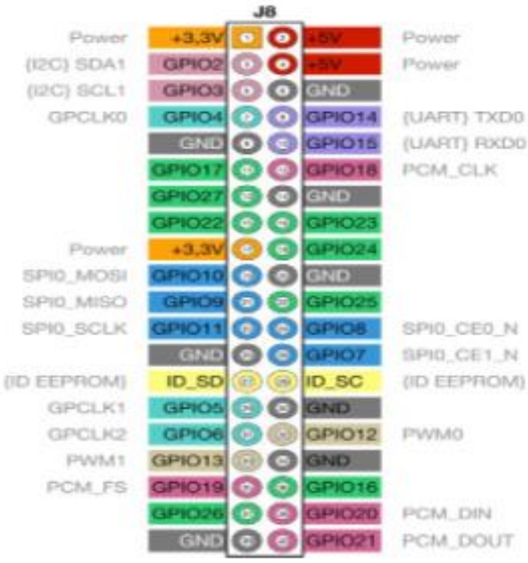


fig.2 (b): pin description of Raspberry Pi 3
[Source: raspberrypi-spy.co.uk]

B. Webcam

A webcam [6] is a video camera which streams real time video or images with the help of a computer or computer network. It is connected via a USB cable. It is a cheap and reliable form of video-telephony because of its low-cost and high flexibility. Despite the low-cost, the resolution is impressive. Webcams are mostly used for establishing links for video, enabling computers to serve as video-phones or video-conferencing nodes. Additional applications involve Security Surveillance [5], Computer Vision & Video Broadcasting. In this project, Webcam is used as Smart Security Camera [6] which is used to stream Live video over the Internet and also to click the image of Intruder i.e. unauthorized person. The Webcam we have used over here is Logitech C310 HD Webcam which features 5 MP snapshot, built-in mic with noise reduction, automatic light correction, 60-degree field of vision with fix focus type specialty and a maximum resolution of 720 p and 30 frames per second.

fig.3: webcam

C. Powerbank

A power bank (capacity of 10,000 mAh) is used for making the model portable. This power bank could keep the system live for a day with low power consumption by disabling the supply to webcam and other components and can be recharged. It provides a steady voltage of 5 V, 2 A for a long time. It makes the system compact and flexible.

fig.4: power bank

D. Bluetooth earphones

Any wired earphone or wireless earphone with Bluetooth connectivity can be used for the purpose. Here, wireless earphone has been used for the ease of convenience.

IV. System software design

The various processes in System Design are: Text extraction from image, Text to speech and Microsoft translator so that the machine understands the text of the image and gives voice output. Firstly, camera takes the image and is stored as an image file of .jpg extension. Then, the OCR engine [9] converts it from image file to text file by extracting the numbers and characters of English alphabet provided the text is printed. It can also recognize handwritten texts but with less accuracy as it depends upon the clarity of the handwriting. eSpeak [9] is a text to speech engine which is then used to convert the file with .flac extension. Then a python program which is pre-written is executed and output is generated as translated speech by using Google text to speech engine in response to the flac file generated.

A. *Text extraction from image*

The image is captured by the webcam used in place of Camera module to get good resolution images which eases further processing like reading from the image i.e. text extraction. Tesseract [3] [10] software is installed in the raspberry pi by a command which is used for the purpose.

Command used to install Tesseract in Raspberry Pi is:

```
sudo apt-get install tesseract-ocr
```

Tesseract converts image to text and stores it in .txt format. The process of text extraction involves the OCR technology, character recognition and extraction.

(a) OCR TECHNOLOGY

OCR stands for Optical character recognition is a technology that can detect text or characters present in an image. Apart from recognising text from any scanned documents, it serves many purposes. OCR processes an image in digital form and detect characters such as numbers, letters, symbols and is widely used as a form of data entry from printed paper data records, any suitable documentation. OCR is a common digitalization method to extract printed texts. The recognized text can be edited electronically and stored which are used in many machine processes such as text to speech, text mining, machine translation etc.

Pre-processing-

Various pre-processing techniques are used to increase the accuracy: -

- De-skew- To align the image properly, tilting the image clockwise or anti-clockwise so that text lines become horizontal or vertical to make it easy to recognize.
- Despeckle – The process of smoothening of edges.
- Binarization –The task of binarization is performed as a simple way of separating the text from the background[11].
- Line removal – Cleans up non-glyph boxes and lines
- Layout analysis – Captions, columns and paragraphs are identified as distinct blocks which are mainly important in multi-column layouts and tables.
- Line and word detection – Baseline for word and character shapes and separate words is established as per necessity..
- Script recognition – Right OCR is invoked to handle specific scripts in multilingual documents.
- Ssegmentation – For per-character OCR, single characters that are broken into multiple pieces due to artifacts must be connected while multiple characters that are connected due to image artifacts must be separated.
- Normalize aspect ratio and scale-fixed-pitch fonts is segmented by simple alignment of image to a uniform grid in which vertical grid lines will least often intersect black areas. whitespace between letters may be larger than that between words. Hence, some other technique might be used.

(b) Character recognition

Character Recognition is done by pattern matching, pattern recognition or image correlation.

(c) Character detection

OpenCV library is used to extract the letters during the processing of images. The frame captured by the camera is sent for processing and object of interest is extracted using cascade classifier. Further processing is done in following steps: -

- Gray scale Conversion: It is done by generating a black-and-white version of the colour or grayscale scanned page. It is the first step in processing by OCR which is a binary process. For original image, generally any black region in it is part of a character that is to be recognised while white regions are portion of the background. This process might induce some error during the conversion process.
- OCR grayscale conversion process: The main functioning of all OCR algorithms is the same i.e. they process the image taken by webcam to recognize the text in the image line by line, word by word and character by character.
- Basic error correction: Optical Character Recognition Algorithm uses near-neighbour analysis to automatically detect and correct error.
- Layout analysis: Good OCR algorithm can detect complex page layouts like images, tables and multiple text columns and can handle such complexities automatically.For example- Images are automatically turned into graphics and table columns are split, this results in differentiation in text of first line of first column and the first line of the second column.

(d) TESSERACT

Tesseract is an OCR Optical Character Recognition engine that works with Page Layout Analysis Technology.Tesseract takes input image in form of a binary image. Tesseract performs both, the traditional- white on black text and also inverse is possible. After the outlines of components are stored, nesting of outline process is done. By gathering the outlines together Blobs are formed which are organized into text lines. Text lines are analysed with respect to their pitch and are then broken into words by analysing the spacing of characters. By fixed spaces and fussy spaces, the proportional text is broken into words and fixed pitch is sliced in character cells.

Tesseract recognizes words in two steps. In the first phase it tries to recognize words and the recognized words will be sent to the adaptive classifier as training data which recognizes the text accurately and in the second step, the partially recognised in the first step are recognized well through running over the page again. page.

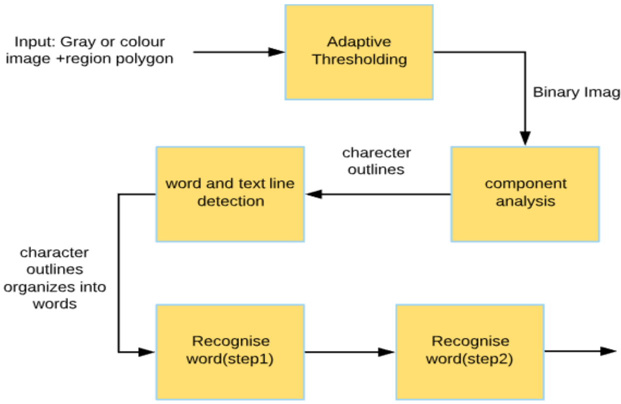


fig.5: Tesseract Architecture

B. Text to speech synthesizer

It is an software used to synthesize speech from text [9]. A TTS Engine converts written text to a phonemic representation, the phonemic representation is converted to waveforms that can give a sound output. eSpeak [9] is a software which can be easily used in a raspberry pi by installing eSpeak engine. Here it is used for converting the text file into an audio file using .flac extension file. Flac stands for free lossless audio codec. It is an audio coding format for lossless compression of digital audio. At the end of this phase an audio file is created.

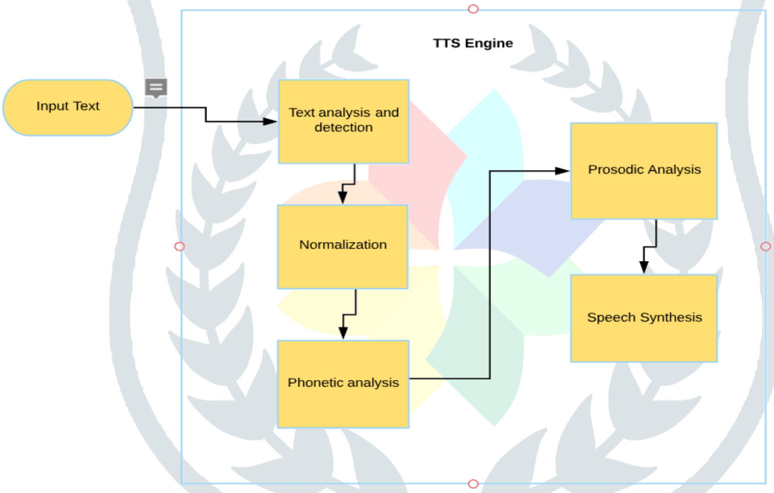


fig.6: Text-To-Speech Engine

- Analysis and detection of text: Analysis of text involves the analysing and organising the input text into a list of words which contains abbreviations, acronyms and numbers.
- Normalization: This process involves the conversion of all letters both upper and lower case and stopping common word and removing punctuations.
- Phonetic Analysis: Grapheme to Phoneme conversion process is done here where orthographical symbols are converted into phonological symbols.
- Prosodic Analysis: It is highly concerned process which implements the analysis of language by observing the stress patterns and intonation in various contexts.
- Speech synthesis: Finally, waveforms are created from the phonetic representation to produce sound output.

C. Translation And Speech Output

Output speech with desired language is obtained by providing .flac file to the written python program which consists of Google TTS engine, MP layer and various authorizations for translator facilities within it. An extensive varieties of media formats, any format reinforced by FFmpeg libraries can be played and saved to a local file by the MPlayer. MEncoder, a program that can take the input file converts into various output formats by applying many transformations. Text to speech conversion is later on done by the Google text to speech engine. Microsoft translator which is a translation service by Microsoft[6] is used for translation process of speech in required language of user. Typical language code has to be given in the command for execution for getting output in user required language.

A sample command for getting Spanish as the destination language is: sudo nano pitranslate.py –o en –d es “filename”.

The speech in Spanish language is then audible via the earphones connected to the 3.5mm jack on Raspberry Pi or via Bluetooth in case of wireless earphones.

V. Implementation

The hardware is setup properly as shown in fig 7 and power supply will be provide d to the board using a power bank.

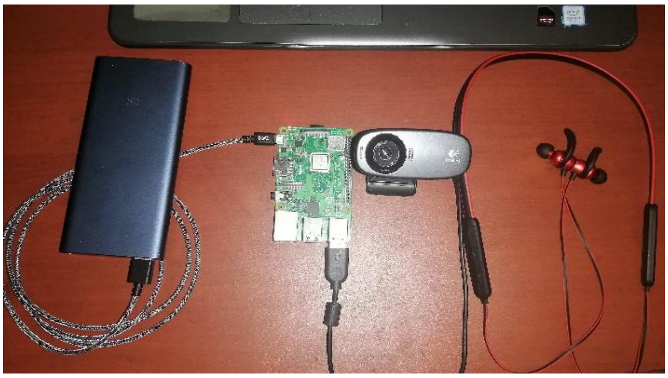


fig.7: Hardware setup

.

The implementation process starts with the insertion of SD card having Raspbian OS installed into the Raspberry Pi 3 board’s slot. After the insertion of SD card, the Raspberry Pi 3B model is booted up. The Raspbian OS operating system provides a graphic user interface with several inbuild functions and libraries. The commands for performing several functions in the pi are Linux based. For security purpose the username and password of the Raspberry Pi are changed. Various permissions for ssh and camera are enabled before starting the implementation process. Since Raspberry Pi3B is an advanced version the Raspbian OS has to be updated and upgraded. The following commands are used for the purpose.

Command 1: sudo apt-get update

Command 2: sudo apt-get upgrade

This two-command execution downloads and installs new packages and libraries into the Raspberry Pi.

The next step of implementation begins with the image capturing process. An image is captured by the webcam and is given as input to the Raspberry Pi 3B board.

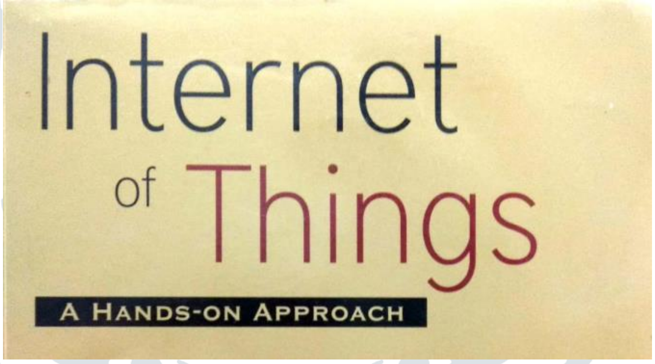


fig.8: Image captured by Webcam

The text extraction process needs the installation of Tesseract OCR Optical Character Recognition engine by executing the following command,

Command 3: sudo apt-get install tesseract-ocr

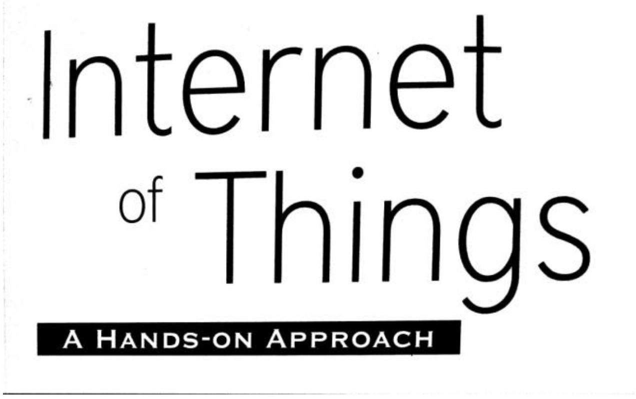
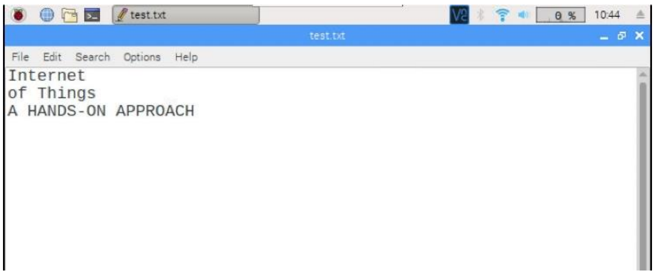


fig.9: Conversion to Gray scale

Tesseract OCR engine performs the text extraction and detection process by implementing a series of phases. The text recognition process uses Tesseract OCR engine which performs the pre-processing and post processing process of image. The extracted text from the image is saved in a.txt file which is further used for speech synthesis.



```
Internet
of Things
A HANDS-ON APPROACH
```

fig.10: Image to text conversion.

The .txt file is given as input to the speech synthesizer which is known as text to speech engine. The TTS text to speech engine converts the text to speech by implementing a series of steps like text analysis, normalization, Phonetic and prosodic analysis. A .flac extension file is generated which is an audio format file. All this process is done by eSpeak software installed in the Raspberry Pi which serves as a TTS.

The final step of the process will be speech translation where output sound is translated to desired language. The Google or Microsoft translator installed in Raspberry Pi does the task of speech conversion in desired language.

VI. Conclusion

Hence the purpose of helping the visually impaired will be successfully achieved by this system. This device helps the blind and visually disabled people by not only assisting them while reading but also in shopping by helping them to read the labels on products, indoor navigation at home and college where directions are labelled on boards in different locations. The overall cost of this system is around five thousand rupees which is a moderate cost and is more reliable and accurate. This approach saves a lot of money for them by reducing the cost of printing Braille books. Since the device is a portable version and uses a Bluetooth earphone it becomes very convenient for the blind people to carry it and use it. We can implement several other applications like image captioning which uses deep learning to improvise the system more but the process consumes lot of power and processing speed requirement is also high which restricts the usage of Raspberry Pi for this purpose.
However, Microsoft ARTIK Board can be used as core as improvisation to this system.

References

[1] Ezaki, Nobuo, et al. "Improved text-detection methods for a camera-based text reading system for visually impaired persons." Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE, 2005. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Ms.Rupali, D Dharmale, Dr. P.V. Ingole, "Text Detection and Recognition with Speech Output Visually Challenged Person", vol. 5, Issue 1, January 2016.

[3] Rajkumar N, Anand M.G, Barathiraja N, "Portable Camera Based Product Label Reading For Blind People.",IJETT, Vol. 10 Number 11 - Apr 2014.

[4] Boris Epshtein, Eyal Ofek, Yonatan Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform."IEEE, 2010,pp.2963-2970.

[5] N. Haering, P. L. Venetianer and A. Lipton, "The evolution of video surveillance: An overview," Machine Vision and Applications, vol. 19, no. 5–6, pp. 279–290, 2008.

[6] T. Winkler and B. Rinner. "Security and privacy protection in visual sensor networks: A survey," ACM Computing Surveys (CSUR), vol. 47, no. 1, article no. 2, 2014.

[7] Raspberry Pi 3 Model B,[Online].Available: https://www.raspberrypi.org

[8] Details of Webcam [Online]. Available: https://en.wikipedia.org/wiki/Webcam

[9] Anusha Bhargava, Karthik V. Nath, Pritish Sachdeva, Monil Samel (April 2015), "Reading Assistant for the Visually Impaired" International Journal of Current Engineering and Technology (IJCET), E-ISSN 2277 – 4106, P-ISSN 2347 – 5161, Vol.5, No.2.

[10] K Nirmala Kumari, Meghana Reddy J, "Image Text to Speech Conversion Using OCR Technique in Raspberry Pi" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering(IJAREEIE), ISSN (Print): 2320 – 3765, ISSN (Online): 2278 – 8875, Vol. 5, Issue 5, May 2016.

[11] Zoran Zivkovic, "Improved Adaptive Gaussian Mixture Model for Background Subtraction", IEEE,pp:28-31, 2004.