Image Text to Speech Conversion

Reena Kinhekar[1] Sulbha Kamble[2] Anjali Suryawanshi[3] Mrs. Sayali. N. Mane[4]
[1,2,3]UG Student [4]Assistant Professor
[1,2,3,4]Department of Electronics and Telecommunication Engineering
[1,2,3,4]D.Y Patil College of Engineering Akurdi, Savitribai Phule Pune University, Pune-411007, India

Abstract— Computer vision requires Character Recognition. Optical Character Characters are recognized from images digitally in Recognition. In this paper an innovative, efficient and real-time cost beneficial technique that enables user to hear the contents of text images instead of reading through them has been introduced. OCR technique is used to changing it to voices and text to speech converter is device that scan and reads English alphabets. In raspberry Pi two methods are used that are the of Optical Character Recognition (OCR) and Text to Speech Synthesizer (TTS). This kind of system helps visually impaired people to interact with computers effectively through vocal interface. Text Extraction from color images is a challenging task in computer vision. In this paper, we describes the design, implementation and experimental results of the device. Two processing modules are consisted an image processing module and voice processing module.

Keywords: OCR, TTS, Image Text to Speech Conversion

I. INTRODUCTION

Optical character recognition has came into picture in late 19th decade. With the use of smart phones OCR can be used to extract text from image file captured. Commercial and open source OCR systems are available which can be used in integration with text to speech synthesizer to implement a device which can convert text on image into speech.

Language express one's thoughts by means of a set of signs, whether graphical gestural, acoustic, or even musical. Language express is distinctive nature of human beings, who are the only creatures to use such a structured system. Speech is one of its main components. Speech is used for communication between human being and others. People have studied it and often tried to build machines to handle it in acoustic way. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so the digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages.

We are taking images as an input. To extract text from the image we need to do image processing. We are implementing some concepts from curricular subject Digital Image Processing. We are doing Python programming for Raspberry Pi under Embedded Platform. We are implementing for English alphabets. Next step is TTS that is Text to speech conversion in this we have to convert recognized text from OCR into .wav file or simply in speech file.

II. BLOCK DIAGRAM

The hardware consists of the following parts:

1) Raspberry pi camera module
2) Raspberry pi 3 [model B] mounted with SD card
3) Speakers
4) Internet connection via Ethernet or Wi-Fi, laptop.

The Fig 3 gives the block diagram of system hardware design.

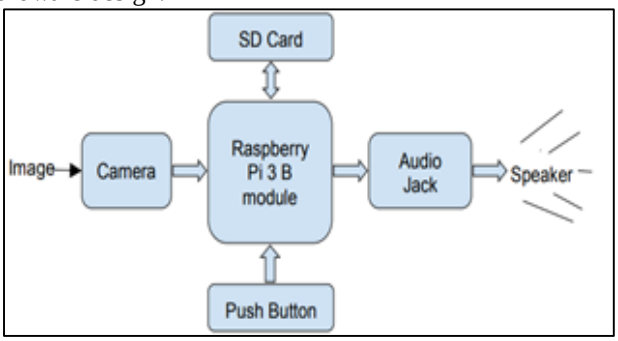

Fig. 1: Block Diagram

The Raspberry Pi 3 Model B is the third generation Raspberry Pi. This powerful credit-card sized single board computer can be used for many applications and supersedes the original Raspberry Pi Model B+ and Raspberry Pi 2 Model B. While maintaining the popular board format the Raspberry Pi 3 Model B has a more powerful processer, which is 10 times faster than the first generation Raspberry Pi processor. It also adds wireless LAN & Bluetooth connectivity thus making it the best solution for powerful connected designs.

The camera module is connected with the camera serial interface of the raspberry pi using the 15-pin ribbon cable. Enable the camera support in the configurations. This helps in capturing a 5 MP resolution image by a single command. The command is: sudo raspistill -o image.jpg The Fig. shows a raspberry pi 3 model B with a raspberry pi camera module connected via a 15-pin ribbon cable.

Text-to-speech device consists of two main modules which are image processing module and voice processing modules. Image processing module converts image into text using camera. Voice processing module processes data with specific physical characteristics so that the sound can be understood when the text is converted into sound.

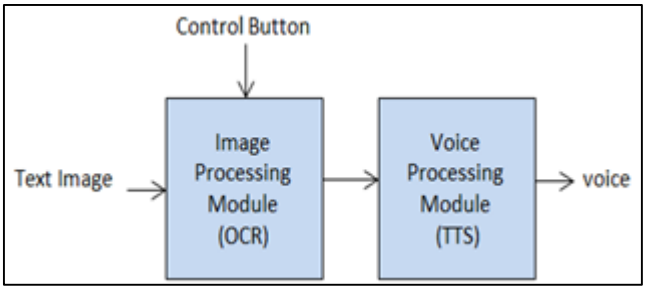

Fig. 2: Block Diagram of OCR & TTS

The block diagram of Text-To-Speech device is shown in Fig 2.

1st block-Image processing module- OCR converts .jpg to .txt form.

2nd block-Voice processing module- It converts .txt to speech.

OCR or Optical Character Recognition is a technique that recognize the character through the optical mechanism, this technology imitates the ability to senses of the human sight. Camera used as replacement for eye and through computer engine image processing is done. OCR engine has a Tesseract OCR type of engine which has matrix matching. The two main aspects of Tesseract engine is flexibility and extensibility of machines and the fact that many communities are active researchers to develop this OCR engine. Tesseract OCR supports 149 languages. In this method we are going to identifying English alphabets. Before sending the image to the OCR. It is converted to a binary image to increase the recognition accuracy. another open source tool for image manipulation is converting Image binary is done by using Imagemagick software. The output of OCR is the text, which is stored in a file (speech.txt). Machines have defects such a dim light effect and distortion at the edges. OCR engines has difficulties to get high accuracy text engines. in order to get the minimal defect. OCR needs some supporting and condition method.

The input image captured by the Logitech (C270) web camera has a size of 3 MPI (720 X 340 pixels). Tesseract OCR accuracy will decrease with the font size of 14pt. Based on the specifications of the Tesseract OCR engine, the minimum character size is 20 pixels which is helpful to read uppercase letters.

Via GPIO pin (23) that is connected to the button image is taken by the user, using interrupt function and the picture is taken by using raspistill program with sharpness mode to sharpen the image. The resulting image has a .jpg format with a resolution of 720 x 340 pixels. In this module text is converted to speech. The output of OCR is the text, which is stored in a file (speech.txt). Here, we use Festival software to convert the text to speech. Festival software is an open source text to speech (TTS) system, which is available in different types of languages. In this project, English TTS system is used for reading the text.

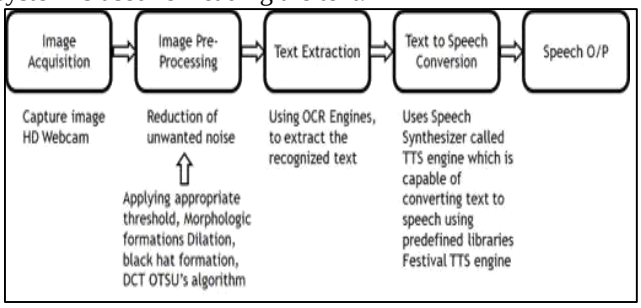

Fig. 3: Execution Flow

Image in the form of text is captured using camera and give to the next processing. Input image can be page of a book, web page etc. Images should contain English alphabets only. Before extracting text unwanted noise in the image is reduced. We can apply appropriate threshold, Morphologic transformations, Dilation, Black hat formation, DCT OSTU's algorithm. Text extraction is first important step in execution flow. Using OCR engines to extract recognized text. We are using Tesseract engine for this. Text to speech synthesizer is second important step in execution flow. Text to Speech engine is used for Speech synthesizer. It converts text to speech using predefined libraries in Festival TTS engine. Speech output i.e. .wav file is given to audio jack

III. RESULTS

The output observed for image text to speech conversion is Image is captured by the Camera. The captured image is converted into text format. This text is then converted into speech by the TTS and the output in the form of speech through Headphones or loudspeaker is heard.

Font size, Distance, Total letters& Correct letters this factors are responsible for efficiency .

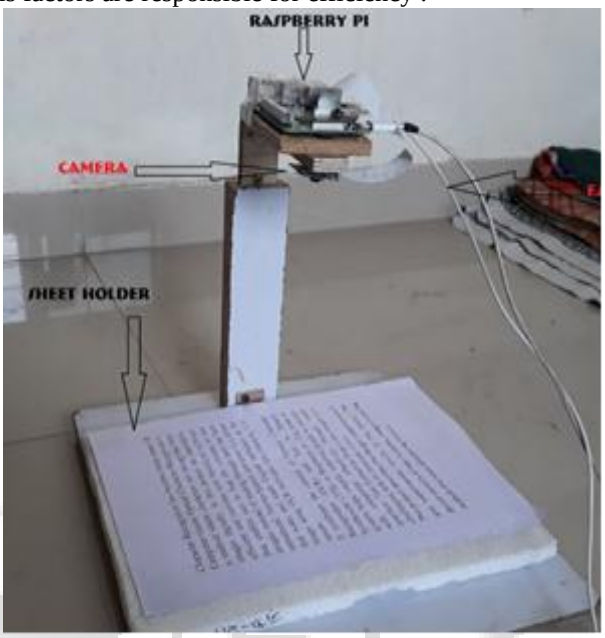

Fig. 4: Image text to speech result

Font Size	Distance	Total Letters	Correct letter	Efficiency
14	20cm	950	630	66%
	22cm	950	580	61%
14 Bold	20cm	950	760	78%
	22cm	950	490	51%
18	22cm	950	590	55%
	24cm	950	600	63%
18 Bold	24cm	950	850	89%
	26cm	950	650	68%
24	24cm	950	610	64%
	26cm	950	740	77%
24 Bold	26cm	950	690	72%
	28cm	950	810	85%

Table 1: Result for different font size

IV. CONCLUSION

Text-to-Speech model can change the text format from image into sound. This portable device can be used by visually impaired people to read boards, books, computer screens etc. This device can be used independently and does not require internet connection. Editing process of books or web pages becomes easier through this method.

REFERENCES

[1] J. Liang, et. al. "Geometric Rectification of Camera-captured Document Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 591-605, July 2006.
[2] K. Lakshmi, Mr. T. Chandra Sekhar Rao, "Design and Implentation Of Text To Speech Conversion Using Raspberry PI", International Journal of Innovative Technology and Research, Volume No. 4, Issue No. 6, October-November, 2016.
[3] Asha G. Hagargund, Sharsha Vanria Thota, Mitadru Bera, Eram Fatima Shaik, "Image to Speech Conversion for Visually Impaired", International Journal of Latest Research in Engineering and Technology, Volume 03, Issue 06, June 2017.
[4] "Digital Image Processing" 3rd Edition"-Gonzales and Woods
[5] Aaron James S, Sanjana S, Monisha M, "OCR based automatic book reader for the visually impaired using Raspberry PI", International Journal of Innovative Research in Computer and Communication Engineering, Volume 4, Issue 7, January 2016.
[6] T. Dutoit, "High quality text-to-speech synthesis: a comparison of four candidate algorithms," Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, vol.i, no., pp.I/565-I/568 vol.1, 19-22 Apr 1994.

