Head Mounted Device for Real World Text to Speech Conversion

Nikhil Varghese
Information Technology Department,
Sardar Patel Institute of Technology,
Mumbai, India

Gaurav Tripathi
Information Technology Department,
Sardar Patel Institute of Technology,
Mumbai, India

ABSTRACT
There is no low-cost aid for visually impaired people despite several advances in technology. This paper presents a mobile head-mounted device to detect and convert text in natural scenes to speech. The major components of the device are a Raspberry Pi, a high definition webcam, earphones and a portable power bank. The Raspberry Pi is connected to the webcam which captures the image. A text detection algorithm using Class Specific Extremal Regions (CSERs) is implemented to detect the text in complex natural scenes. The segmented image is passed to the Tesseract OCR engine for text detection. The identified text is converted to audio using the espeak Python module in the Raspberry Pi. Thus, a visually impaired person can use this device to hear all the text in his surroundings like the name of a shop, public notices, billboards, road directions, etc.

General Terms
Wearable technology, Computer Vision, Text recognition, Text-to-Speech synthesis, Natural Language Processing.

Keywords
Class-Specific Extremal Region, Head-mounted device, MSER(Maximally Stable Extremal Regions), Raspberry Pi, Tesseract OCR, Probabilistic Hough Lines Transformation

1. INTRODUCTION
According to World Health Organization (WHO) [1], 285 million people are estimated to be visually impaired worldwide: 39 million are blind and 246 million have low vision. In addition to this, about 90% of the world's visually impaired live in low-income settings. For long engineers have been innovating to improve upon existing computing devices. However, the visually impaired have not had commensurate improvement in the quality of their life. A number of hardware innovations in mobile have given rise to wearable technology which has resulted in commercial wearable computing products from companies like Apple, Google, Microsoft, etc. But, there has not been much progress in wearable technology which could potentially change the lives of millions of people, the visually impaired. Text detection along with speech synthesis plays a significant role in helping the visually impaired people to understand the scene in their surroundings. Kurzweil's reading machine [2] in 1975 is one of the earliest approaches of assistive technology which enabled book reading for blind people using a flat CCD scanner and a computer unit with optical character recognition (OCR) and text to speech synthesis (TTS) systems. iCARE portable reader [3], improved this layout using a camera and made document manipulation less cumbersome. A small camera mounted on a baseball cap was used for user navigation in an environment in [4]. A device comprising glasses with integrated camera and DSP-based processing unit for performing the recognition and speech synthesis was proposed by Chimel et al in [5]. However, this device was directed mainly towards document reading for the blind. At present there are a few other desktop and mobile solutions that are used widely. Google Translate [6] also offers a similar feature to obtain text-to-speech conversion by clicking pictures over a mobile phone. Google Translate offers translation service in over 90 languages. KNFB Reader [7] is a more recent example of an attempt to aid the visually impaired by making use of the camera on a smartphone. It offers text-to-speech conversion of texts in an image by taking a picture from the mobile phone.

Detecting and recognizing text also finds application for translation purposes, e.g. for tourists or robots. This lead to the Translation robot [8] which consisted of a camera mounted on a reading glass along with a head-mounted display used to play the output device for translated text. Carlos et al., designed a head mounted device for text detection in natural scenes [9]. The device was made up of components mounted on an aluminum framework which was embedded into a flat-cap. The metal framework comprised of a USB camera, a RC receiver and USB sound card connected to a USB hub that was in turn, is connected to a laptop. The problem of this design is the necessity of carrying around the laptop with the device. This paper presents a light, low cost, mobile head mounted device built using a Raspberry Pi and a USB camera powered using a portable battery power. The Raspberry Pi [10] is a low cost, credit-card sized computer that can be plugged into a TV and a keyboard and mice can be attached to it. It is a capable little computer that can is used in electronics projects, and for many of the things that one does on his/her desktop PC, like spreadsheets, word-processing, programming and games. This head mounted device enables the person to understand text in natural environments with the help of an audio output of the text. The device can be connected to the internet using WiFi for image processing to be performed on a remote server or a cloud service provider like HP Haven-on-demand or Amazon AWS. For the purpose of this paper, the image is processed on the Raspberry Pi.

2. HARDWARE DESIGN
The camera is placed at the centre and positioned at an angle so that the video feed resembles what the person's eyes would see. The Raspberry Pi 3(model B) is positioned at the right side and the battery source toward the left. These three components are held together using an elastic band with a velcro in order to fit the user in the best possible manner.

A low cost high definition USB webcam, Logitech HD Pro Webcam C920 [11] with autofocus capability is used to identify and take the best quality image of the text in natural environments. The device is powered through a power bank with 3200mAh power. The output audio is played on an earphone which is connected to the 3.5m jack in the Raspberry Pi.

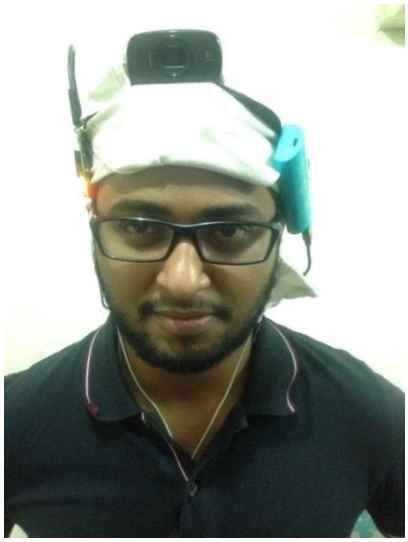

Figure 1: User wearing the device

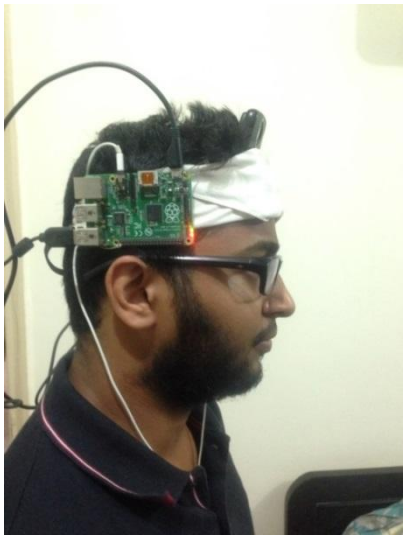

Figure 2: Powered up Raspberry-Pi

3. PROPOSED SYSTEM

A simplified schematic of the proposed system is shown in Figure 3. Initially, an image of the natural environment is captured when the user clicks a button. This is followed by text extraction. It consists of two stages, text detection and text identification. Text detection involves the localization of all candidate text regions in the image. Text identification involves using these localized regions and identifying a character through an OCR engine. Text detection of the possible text regions in the image input stream is done using a Class-specific Extremal Regions approach as described in [12]. Class-specific Extremal Regions is similar to the MSER, described in [13] and [14], and was used by Luis Gomez and Dimosthensis Karatzas in [15]. The skew angle for each candidate region is calculated using probabilistic Hough transform. The candidate regions are corrected by rotating them by the skew angle.Then, these regions are analyzed using the open source Tesseract OCR engine [16]. If the output has a confidence level higher than a certain level, the results of the Tesseract engine are passed on for spell checking followed by Text To Speech (TTS) [17] synthesis. Finally, the audio output is played back through the earphones to the ears of the user.

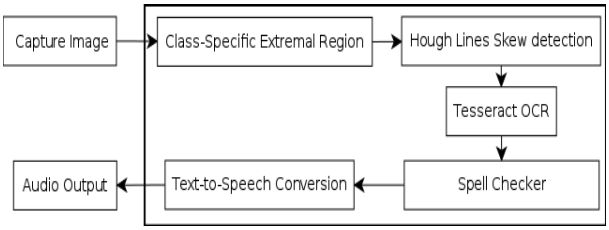

Figure 3: Flowchart of the application

3.1 The initial setup

The user straps the device on his head such that the webcam is positioned at the center of his forehead. The system boots when the power button is pressed by the user. This will execute all the scripts to prepare the device for required usage. The scripts include operations like checking whether all components are connected and are working properly and starting the camera capture module. Once the scripts are executed, an audio output saying that the device is working properly or not is played. When the user wants to click a picture, he will hit the capture button. This triggers the script to capture the image. The webcam captures the image using auto-focus and dynamically adjusts the brightness, contrast and saturation.

3.2 Text detection and identification

Class-specific Extremal Regions [18] are used to segment image for textual content. It is similar to Maximally Stable Extremal Regions (MSER). MSER was originally used as a method of blob detection in images, but it works well with text regions. Since performance close to real time text detection and identification is required, an algorithm with very low time complexity is implemented. It performs a sequential selection from a set of Extremal Regions(ER) [18]. Class Specific Extremal Region differs from the original ER in that the selection of ERs id done by a sequential classifier trained for character detection. This technique essentially drops the stability requirement of MSERs. Character recognition is performed using feedback loops which identify the most probable character segmentation by grouping ERs into words. This system presented in this paper makes use of the Class-specific Extremal Regions approach which is implemented in OpenCV 3.

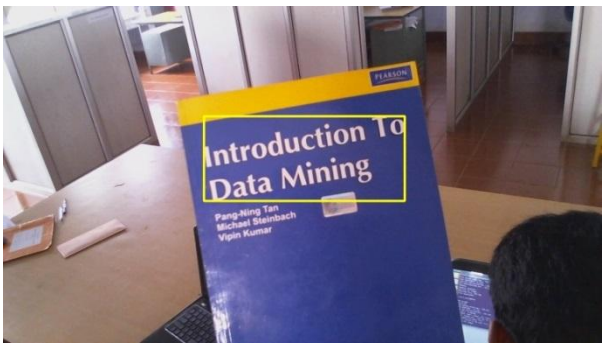

Figure 3: A candidate region is highlighted

Figure 4: Candidate region before probabilistic Hough lines transformation

The segmented regions typically contain text written in the same context, like a sign board on the road. The text in these regions is generally skewed. A sample candidate region is shown in Figure 3. Probabilistic Hough Lines Transformation is used to identify the skew angle for very candidate region. The mean angle of all the lines obtained from the transform is taken as the skew angle as shown in Figure 5.

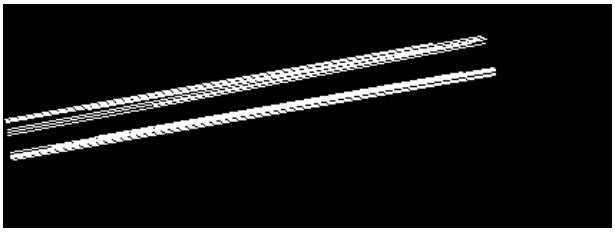

Figure 5: Candidate region after probabilistic Hough lines transformation

The image is rotated by this angle. Although this is a heavy and time consuming operation, it affects the result greatly. Once the regions are rotated, the image is passed to the Tesseract OCR engine. Tesseract has been trained with a huge dataset [19] to identify text from a varied set of environments.A sample of the input to Tessercat is shown in Figure 6. It gives a text output along with a confidence level which determines how likely it is to have identified the text correctly. If no text is identified despite having cadidate regions, the rotated image is further rotated by degrees of 180, 90 and 270 and passed to the Teserract engine until text has been identified. These rotations are relatively less resouce intensive.

Figure 6: Candidate region after rotation by skew angle

3.3 Text Validation

The result from the text detection and identification step is not always correct. Even a single misread character would obscure the speech. A simple solution is to add a spell checker to avoid silly mistakes. A trivial spell checker proposed by Peter Norvig [20] is modified to include numbers within text as a spell correction candidate. The final step is to order the extracted words. First, the regions are are ordered from left to right while gradually moving from top to bottom. Then the same algorithm is used to order the text inside each region. This ordered text is used for text to speech synthesis.

3.4 Text to speech synthesis

The text is used for speech synthesis using the espeak [21] module in available in Python. The audio output is played through the earphones. If there is no text found, a standard message is played to inform the user that no text could be found.

4. RESULTS

The HMD is tested with the MSRA-TD500 [22] dataset. It is challenging a dataset with most images containing text in two different languages, Chinese and English. It is an appropriate representation of the environment this device would be used it. However, it important to note that images which did not have any text in English have been skipped while evaluating the results mentioned below.

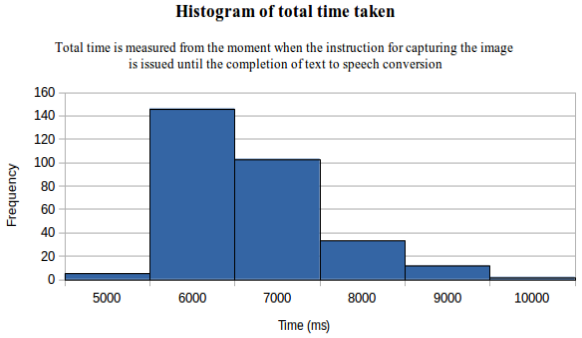

Figure 7: Histogram depicting the distribution of the total time taken for the end to end process for MSRA-TD500 dataset. The time is recorded the moment the user presses the button until the time when the text to speech synthesis is complete.

Table 1: The statistics related to the total time distribution from the MSRA-500 dataset.

Measures	Time(ms)
Mean	6226.6689550837
Standard Error	47.7384316537
Median	5994.2481517792
First Quartile	5715.1079177857
Third Quartile	6517.8627967835
Variance	685966.314883814
Standard Deviation	828.2308342991
Kurtosis	1.6327133196
Skewness	1.093054831
Range	5251.8389225006
Minimum	4113.0261421204
Maximum	9364.865064621

An overview of the time the visually impaired person would have to wait before he hears the text in the image when he pressed the button is shown in Figure 7. The maximum frequency is present in the 6-7 second bracket. The distribution has a low and broad peak indicated by the value of Kurtosis. Also, the distribution is moderately skewed to the right. A more detailed analysis is presented in Table 1. The median is approximately 6 seconds with a standard deviation of 828ms. Although the range is higher than 5 seconds, the difference between the first and third quartile is 802ms. Thus,

indicating a high probability of receiving the output in less than 6.5 seconds.

The results of the accuracy are described in Figure 8. Almost three-fourths of the images yielded a result in either a partial match or a complete match. The images where no match was found contained text in very challenging angles coupled with prominent text in Chinese. The results could be enhanced by adding support for multiple languages. However, this is out of scope for this paper.

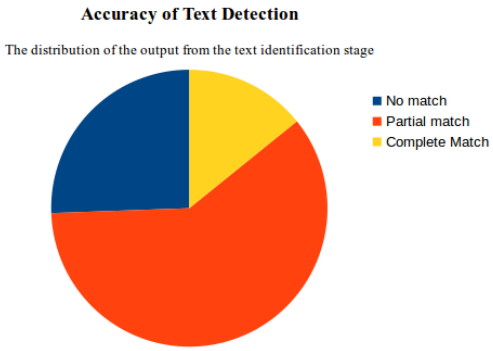

Figure 8: A pie chart depicting the distribution of the accuracy of results for MSRA-500 dataset. Complete match indicates that all the words in the image are identified. Partial match indicates that one or more words in the image are identified. No match indicates that none of the text in the image is identified.

More than half of the images returned a partial match. It was observed that prominent text was identified but the background text was not identified. This is not a big hindrance for the visually impaired person because generally the focus is on the prominent text.

5. FUTURE WORK

This paper presents a low cost aid to the visually impaired by helping them identify text at natural scenes. The Raspberry Pi can easily be connected to the internet using a WiFi module or by integrating using a SIM card. In areas with reliable internet connectivity, the device can offer much faster turn-around time by adding an option of processing on a remote server. This concept can be further extended to involve a GPS tracking and guiding mechanism to help the visually impaired to travel with a greater sense of confidence. With rapid increase in the speed and accuracy image processing and machine learning techniques, it is also possible to venture into deep semantics [23]. A description of the image which does not even contain textual reference can be huge step forward in improving the lives of the visually impaired people. Ever improving computer vision algorithms coupled with low cost add-ons like GPS and haptic feedback belts pave the way for further enhancements in the device while at the same time, being available to many under-privileged.

6. CONCLUSION

This paper presents a head-mounted device built on a simple Raspberry Pi computer with a camera attached to it powered by a portable power backup. The proposed system is thus a cost-effective model that performs robust text detection and its speech synthesis without a network connection. Although a significant amount of time taken is for the entire process to complete, the HMD is completely self-sustained. It is a cheap and can be used in urban and rural landscapes without any dependency. This increases the ease of use for the consumer dramatically. It is an extremely handy device for visually impaired people.

7. ACKNOWLEDGMENTS

Mr. Nikhil Varghese and Mr. Gaurav Tripathi have contributed equally to this paper. Special thanks Dr. Radha Shankarmani, Head of Department, Information Technology Department, Sardar Patel Institute of Technology and Dr. Prachi Gharpure, Principal, Sardar Patel Institute of Technology for backing the project from the ideation stage

8. REFERENCES

[1] (Aug. 2014). WHO | Visual impairment and blindness. [Online] Available: http://www.who.int/mediacentre/factsheets/fs282/en/

[2] R. Kurzweil, The age of spiritual machines: when computers exceed human intelligence. Viking Press, 1998

[3] T. Hedgpeth, J. A. Black, and S. Panchanathan, "A demonstration of the icare portable reader," in ACM SIGACCESS, 2006, pp. 279–280.

[4] H. Aoki, B. Schiele, and A. Pentland, "Realtime personal positioning system for a wearable computer," in ISWC, 1999, pp. 37–43.

[5] J. Chmiel, O. Stankiewicz, W. Switala, M. Tluczek, and J. Jelonek, "Read IT project report: A portable text reading system for the blind people," 2005

[6] About – Google Translate. [Online] Available: http://translate.google.co.in/about/intl/en_ALL/

[7] (2016). KNFB Reader. [Online] Available: http://www.knfbreader.com/

[8] X. Shi and Y. Xu, "A wearable translation robot," in ICRA, 2005.

[9] Carlos Merino-Gracia, Karel Lenc and Majid Mirmehdi, "A Headmounted Device for Recognizing Text in Natural Scenes", Visual Information Laboratory, University of Bristol, UK

[10] Help Videos - Raspberry Pi. [Online] Available: https://www.raspberrypi.org/help/what-is-a-raspberry-pi/

[11] (2016). Logitech C920 HD Pro Webcam for Windows, Mac, and Chrome OS. [Online] Available: https://secure.logitech.com/en-in/product/hd-pro-webcam-c920

[12] (Nov, 2014). Class-specific Extremal Regions for Scene Text Detection. [Online] Available: http://docs.opencv.org/3.0-beta/modules/text/doc/erfilter.html

[13] Chen, Huizhong, et al. "Robust Text Detection in Natural Images with Edge-Enhanced Maximally Stable Extremal Regions." Image Processing

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions." In BMVC, 2002 (ICIP), 2011 18th IEEE International Conference on. IEEE, 2011 Document Analysis and Recognition, 2013

[15] Gomez L. and Karatzas D., "Multi-script Text Extraction from Natural Scenes", 12th International Conference on Robust Text Detection in Natural Scene Images.

[16] GitHub Tessaract OCR. [Online] Available: https://github.com/tesseract-ocr/tesseract

[17] Thierry DutoitTTS research team, TCTS Lab:An Introduction to text-to-speech synthesis - TCTS Lab

[18] Neumann L., Matas J.: Real-Time Scene Text Localization and Recognition, CVPR 2012 (Providence, Rhode Island, USA)

[19] (2016).GitHub TessData. [Online] Available: https://github.com/tesseract-ocr/tessdata

[20] (Aug, 2016). Norvig, P. How to Write a Spelling Corrector. [Online] Available: http://norvig.com/spell-correct.html

[21] eSpeak text to speech. [Online] Available: http://espeak.sourceforge.net/

[22] (Oct, 2012). Yao, C. MSRA Text Detection 500 Database. [Online] Available: http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)

[23] Andrej Karpathy, Li Fei-Fei "Deep Visual-Semantic Alignments for Generating Image Descriptions", Department of Computer Science, Stanford University, 2014