SeeTell

An Image to Speech Converter

Shantanu Das
Information Science and Engineering
BMS College of Engineering
Bangalore, India

Rozelle Jain
Information Science and Engineering
BMS College of Engineering
Bangalore, India

***Abstract*— There are close to 39 million blind people and around 285 million visually impaired people globally. There is a huge impact on the lives of visually disabled people due to such disability. Although there have been several attempts made for helping visually disabled to see objects via some other alternating sense such as sound and touch. The development of text reading device is still at a dormant stage. The system currently in existence either has a limited scope or requires a heavy investment. Therefore we need a cost effective and truly efficient system that will be able to automatically identify and recite text aloud to visually challenged user base. The main purpose of this project is to recognize the text character from any natural image and convert it into speech signal. The same can be done for any PDF documents and uploaded image. Along with this the application also provides features like dictionary, pace modulation and voice selection options and saving capabilities.**

Keywords*— *Image to speech converter, SeeTell, Speech converter

I. Introduction

Image to speech conversion is a trending aspect of computer technology. It determines an important criterion in which we interact with the system and interfaces across a variety of platforms. Machine replication of human functions, like reading, is an ancient dream. However, over the last five decades, machine reading has grown from a dream to reality. Speech is probably the most efficient medium for communication between humans. Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence.

The advantages of digital imaging are increasing demand of the document processing community where cameras are used to image printed documents, or natural scenes containing textual data. The challenges of complex content and layout, noisy data and variations in font and style presentation still persist. Work in the field covers many different areas including pre-processing, physical and logical layout analysis, optical and intelligent character recognition (OCR/ICR), graphics analysis, form processing, signature verification and writer identification, and has been applied in numerous domains, including office automation, forensics, and digital libraries.

For those who are just learning the language, it might be challenging to understand road signs and signals or reading addresses and letters etc. There should be a way to help them read and understand the same. One may simply click a picture of the material and it could be read out to them. Not only that, difficult word meanings should also be available at all times. Such an application will clearly prove to be very helpful.

A. Literature Survey

Archana A. Shinde, et.al [1] - This paper proposes a segmentation method for printed text image. The processed document is segmented into lines and words. The proposed method can be applied without any further modifications for segmentation of characters in any part of the text document. This method can find number of lines, number of words from the input text document. It can also determine number of words in a specific line.

Benjamin Z. Yao, et.al [2] - This paper proposes a framework that provides a complete solution for parsing image and video content extracting video event, and providing semantic and text annotation. It generates text descriptions of image and video content based on image understanding. One major contribution is the AoG visual knowledge representation. The AoG is a graphical representation for learning categorical image representations and symbolic representations simultaneously from a large-scale image. The image and video contents are expressed in both OWL and text format, this technology can be easily integrated with a full text search engine, as well as SPARQL queries, to provide accurate content-based retrieval. Users can retrieve images and video clips via keyword searching and semantic-based querying. The I2T framework discussed in this paper is an exploratory prototype with many challenging problems yet to solve.

Bernard Gosselin, et.al [3] - Developed a system which is able to automatically identify and recognize text zones in images taken from a camera. It works well for a wide variety of document images without any prior information about the document layout, character size, type, and orientation. A new threshold algorithm has been proposed and discrimination between different kinds of documents enables to apply the new method on corresponding documents, such as extremely degraded ones. The proposed algorithm aims at handling various situations despite of different hardware constraints, typical of mobile environment. The paper presents the overall description of the system.

Boris Epshtein, et.al [4] – Presents an image operator that helps to find the value of stroke width for each image pixel, and depicts its use for text detection in natural images. It defines the notion of a stroke and derives an efficient algorithm to compute it, producing a new image feature. Once obtained, it provides a feature that has proven to be reliable and flexible for text detection. Compared to the most recent available tests, the algorithm reached first place and was about 15 times faster than the speed reported there. The feature was dominant enough to be used by itself, without the need for actual character recognition step as used in some previous works.

Huizhong Chen, et.al [5] – This paper proposes a simple text detection algorithm, which uses edge-enhanced Maximally Stable Extremal Regions to identify the candidate text regions. These candidates are then filtered using geometrical methods and stroke width algorithm to exclude non-text objects. Lettersare paired to identify text lines, which are subsequently separated into words. The system is evaluated using the ICDAR competition dataset and our mobile document database.

Itunuoluwa Isewon, et.al [6] – Identifies the different steps and operations involved in text to speech synthesis. It also develops a simple and attractive graphical user interface which lets the users enter the text in the text field provided in the application. This application works only for English as of now.

Jisha Gopinath, et.al [7] **–**Discusses an approach for image to speech conversion using optical character recognition and text to speech technology. The application developed was simple to use, very cost effective and applicable in the real time. By this approach it is possible to read text from a document, PDF or any electronic book and it can also generate synthesized speech through a computer's speakers.

Kaladharan N, et.al [8] **–** Proposes a simple method for text to speech conversion. Text inputs like the alphabets, sentences, words and numbers are fed into the system. After that text to speech conversion is achieved. This is audible and perfect.

This paper is demonstrating to convert the international language English text into speech sign. The exchange of text to speech is made by the speech synthesizer. Speech synthesis is the imitation technique of human speech. Text handling and speech generation are two main mechanisms of text to speech system.

K. Kalaivani, et.al [9] – The paper talks about conversion of ASCII formatted text file form raw images using Optical Character Recognition, and further the converted text can be made as speech through Text to Speech Synthesizer. Through this system the editing process of books or web pages is made easier. The audio file can be saved for future use also. This kind of system enables visually impaired people to interact with computers effectively through vocal interface. It also enables users to listen the audio created from the contents of web pages, e-books, documents and almost all electronically available data thereby enhancing the human-computer interactions.

M. Nagamani, et.al [10] – This paper primarily focuses on conversion of Telegu numerical digits in the image into speech for numerals. This conversion system for Telugu language is developed for only numeric digits. This system takes an input image with 32 pixel height and n pixel width (n varies).This application is more helpful in banking, toys and many other applications like checking marks, railways, aid to the physically challenged persons, language education and fundamental and applied research etc.

Partha Sartha Giri [11] – Describes the challenges in locating text region in natural (non -document) image with complex background which is a very challenging yet an important problem. It also proposes an accurate text region extraction algorithm based on two methods with grey-information. The proposed methods work smoothly on text region in simple images.

Sneha Sharma, et.al [12] – This paper compares two basic approach to text extraction in natural (non-document) images: edge-based and connected-component based. Both the algorithms are implemented and evaluated using a set of images of natural scenes that vary along the dimensions of lighting, scale and orientation. Various features like accuracy and precision for each approach are analysed to determine the success rate and limitations of each approach.

Sukhpreet Singh, et.al [13] - Presents a literature review on English OCR techniques. Various already existing techniques are studied to find the best technique, but it is observed that the techniques which provide better results are slow in nature while fast techniques mostly provide inefficient results.

ShreejitAchari, et.al [14] – This paper talks about image to text conversion using OCR, also it makes use of Tesseract OCR engine provided by Google for extracting text from image. It also uses separate module which generates an audio file of the text extracted.

Yao Li, et.al[15] - In this paper, a CC-based methodology for text detection in natural images has been proposed. MSERs are first utilized to detect potential text regions. The significant difference in this method is that it applies skeleton to extract stroke width. Moreover, our robust CC grouping method can not only group characters into separated words, but also eliminate false positives at the same time.

II. DESIGN

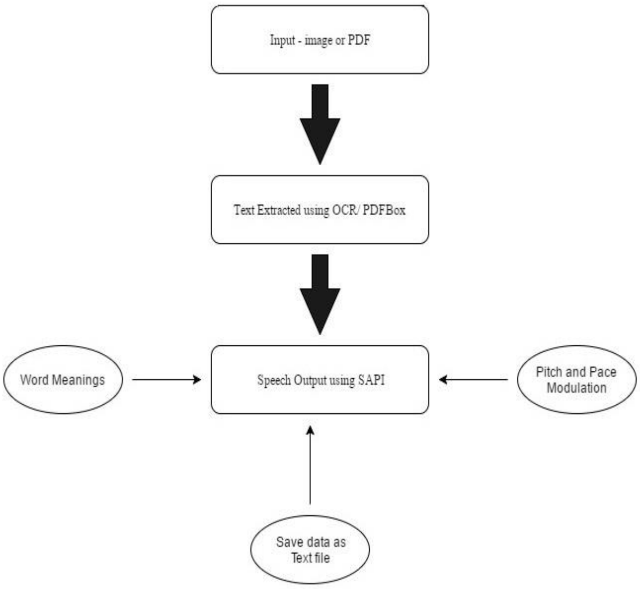


FIGURE 1 HIGH LEVEL DIAGRAM OF THE SYSTEM

The diagram shows that the system consists of two main modules which are

- Image to text conversion or text extraction from image and PDF
- Text to speech conversion

These two modules are responsible in the conversion of image to speech. The text to speech module also offers various other functionalities which are:

- Word meanings and synonyms
- Pitch and pace modulation of speech output
- Save extracted data as text file or in .txt format

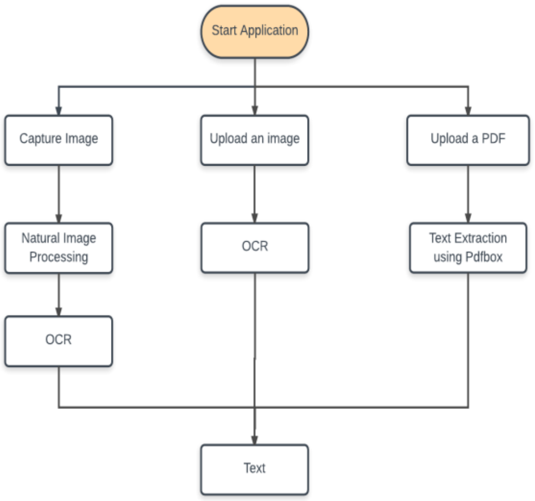


FIGURE 2 DETAILED DIAGRAM OF IMAGE/PDF TO TEXT CONVERSION

On starting the application the users are presented with three options namely

- Capture an image
- Upload an image
- Upload a PDF file

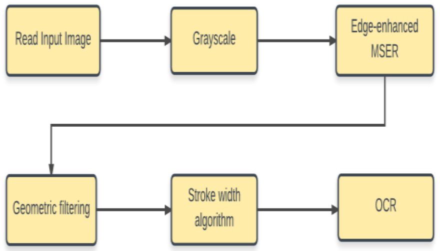


FIGURE 3 NATURAL IMAGE PROCESSING FOR CAPTURED IMAGE

The entire process of image/PDF conversion to text may take variable time. This process requires most of the code execution and is deemed to take most amount of processing time. However, the average time taken by the process has to be minimized to be close to negligible. This is because if this process has high latency, the consequent processes depending on it will be delayed exhibiting overall latency. This will reduce the efficiency of the application.

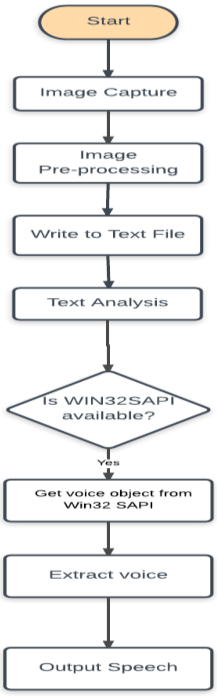


Figure 4 Text to Speech Conversion Flowchart

First image is captured or PDF is uploaded, as the case maybe. Using the previous module text is extracted from the input and converted to text format. Then text analysis is done to separate into sentences. These sentences are to be displayed in proper order alongside the speech output. Text is split into sentences whenever a full stop ('.') or a question mark ('?') or an exclamation mark ('!') is encountered.

The next step is verifying if Windows 32 SAPI is available in the system or not. It provides various voices like Windows Sam, Windows Mike, Windows Hazel, etc. It is useful in providing various pitches of voices to the application. If it is available, the voice objects are extracted and enumerated in GUI to be selected by users. Else only the default voices are enumerated.

This module also gives other functionalities to users. It allows users to set the pace and timing of speech. By default it is set to the slowest pace which is -10. Speed 0 is normal rate of speech. Fastest pace of speech provided is 10. Users are allowed to change the pace whenever requires.

The next facility is to interrupt the output. Users may pause or play as required. The recitation continues from the point the reading was paused. There is an option to stop the recitation altogether. This can also be availed to at any point of time during the recitation. If now the play button is pressed, it continues recitation from the beginning.

III. Implementation

MSER is a method for blob detection in the images. This technique was first proposed by Matas et al. to find any similarities between the image elements from two images having different viewpoints. This method of extracting a comprehensive number of corresponding image elements contributes to the broad-baseline matching, and it has led to improved matching and recognition algorithms. MSER depends on the threshold of the image, if we feed them with some threshold value the pixels below that threshold value are 'white' and all those above or equal are black

The MSER feature detector works well for finding text regions because of the consistent colour and high contrast of text leads to stable intensity profiles. MSER can be implemented as follows:

- The first step is to sweep threshold of intensity from black to white performing a simple luminance threshold of the image.
- Once that is done extraction of the connected components or the Extremal Regions is performed.
- After that a threshold is found when an extremal region is maximally stable.

Finally the regions descriptors as features of MSER are obtained.

As it can be observed in the above image, there are many non-text regions detected alongside the text. Although the MSER algorithm detects most of the text, it also detects several other stable regions in the image that are not text. For removing non-text region a rule based approach can be used.

The detection of text from a natural image is a complicated task . The performance of optical character recognition (OCR) algorithms can be highly increased by identifying the regions containing text in the image. Text detection in natural scenes is a challenging scenario, and there are numerous approaches for solving this problem. However, most text detection schemes restrict the user to specific languages, scale and direction of the text.

The stroke Width algorithm is a local image operator which computes per pixel width of the most likely stroke containing the pixel. The algorithm receives an RGB image and returns an image of the same size, where the regions of suspected text are marked. It comprises of three main steps: The first step is the stroke width transform, then the letter candidates are grouped based on their stroke width, and finally, grouping of letter candidates into regions of text.

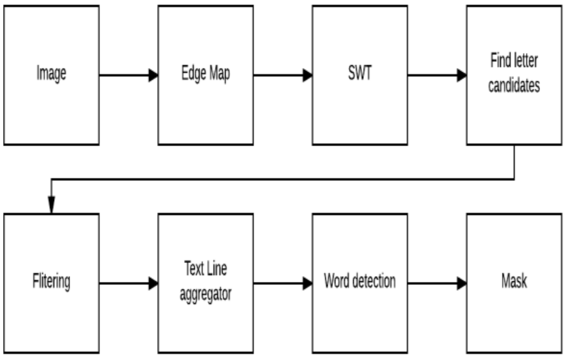


Figure 5 Flowchart of the algorithm

Optical Character Recognition is the process of converting the text, present in digital image, into editable text. It enables a machine to identify the characters through optical mechanisms.

The output of the OCR should ideally be same as input in formatting. The process involves pre-processing of the image file and then acquiring of important information about the written text.

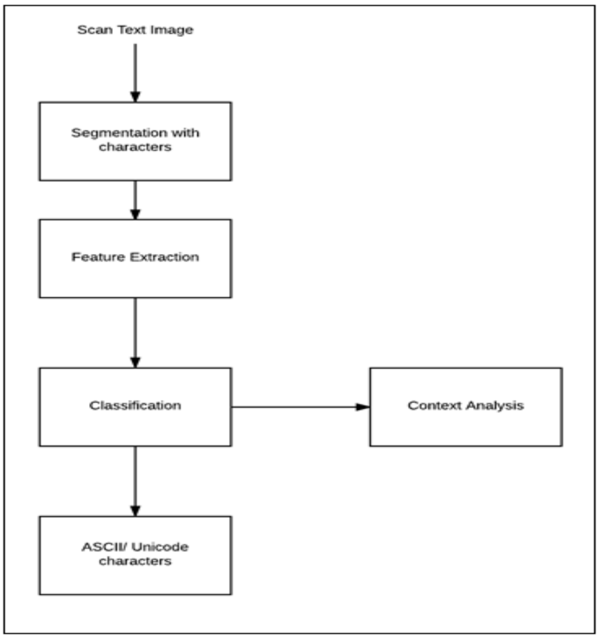
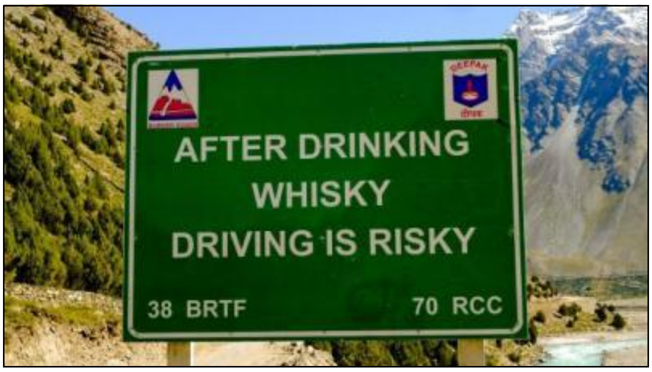


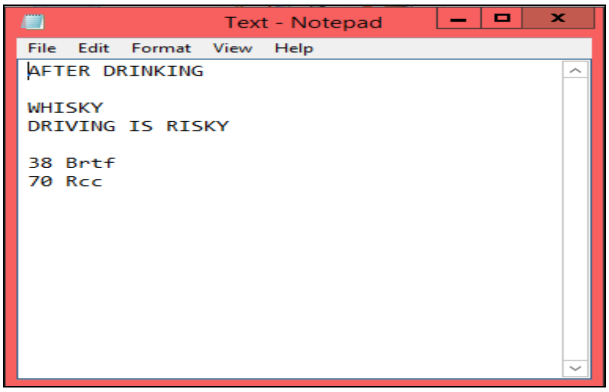
Figure 6 Stages in OCR Design

IV. Results

The algorithm is implemented in MATLAB. The algorithm is tested with various printed and handwritten document, images and PDF. We have considered only good quality of printed documents where there are no overlapping, connected, or broken characters.

The application consists of two phases namely image to text conversion and text to speech conversion. Input is an image or PDF. Intermediate stage consists of a text file. Finally the output is in form of speech. Supporting visuals are present in the application along with various other facilities. Some of the screenshots of the process is given





```
AFTER DRINKING

WHISKY
DRIVING IS RISKY

38 Brtf
70 Rcc
```

Accuracy rates can be measured in several ways, and how they are measured can greatly affect the reported accuracy rate. For example, if word context (basically a lexicon of words) is not used to correct software finding non-existent words, a character error rate of 1% (99% accuracy) may result in error rate of 5% (95% accuracy) or worse if the measurement is based on whether each whole word was recognized with no incorrect letters.

References

[1] Archana A. Shinde, D.G.Chougule “Text Pre-processing and Text Segmentation
for OCR” IJCSET , January 2012.

[2] Benjamin Z. Yao, Xiong Yang, Liang Lin, MunWai Lee and Song-Chun Zhu, “I2T: Image Parsing to Text Description”.

[3] Bernard GosselinFacult´ePolytechnique de Mons, Laboratoire de Th´eorie des Circuits et Traitement du Signal, “From Picture to Speech: an Innovative Application for Embedded Environment”.

[4] Boris Epshtein, EyalOfek, Yonatan Wexler, “Detecting Text in Natural Scenes with
Stroke Width Transform”, Microsoft Corporation.

[5] Huizhong Chen1, Sam S. Tsai1, Georg Schroth2, David M. Chen1, Radek Grzeszczuk3 and Bernd Girod1, “Robust text detection in natural images withedge-enhanced maximally stable extremal regions”, International Conference on Image Processing · September 2011

[6] ItunuoluwaIsewon, JeliliOyelade, OlufunkeOladipupo, “Design and Implementation of Text To Speech Conversion for Visually Impaired People”, International Journal of Applied Information Systems (IJAIS, 2014.

[7] JishaGopinath, Aravind S, PoojaChandran, Saranya S S, “Text to Speech Conversion System using OCR”, International Journal of Emerging Technology and Advanced Engineering, January 2015.

[8] Kaladharan N, “An English Text to Speech Conversion System”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 10, October-2015

[9] K.Kalaivani, R.Praveena,V.Anjalipriya,R.Srimeena, “Real time implementation of image recognition and text to speech

conversion”, International Journal of Advanced Engineering Research and Technology (IJAERT), September 2014.

[10] M. Nagamani, S.Manoj Kumar, S.UdayBhaskar,” Image to Speech Conversion System for Telugu Language”, International Journal of Engineering Science and Innovative Technology (IJESIT) , November 2013

[11] ParthaSarthaGiri, “Text Information Extraction And Analysis From Images Using Digital Image Processing Techniques”.

[12] Sneha Sharma, “Extraction of Text Regions in Natural Images”.

[13] Sukhpreet Singh, “Optical Character Recognition Techniques: A survey”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), June 2013

[14] ShreejitAchari, NishkarshaKotian, SuyogMarde& Prof SeemaRedekar, “Blind Speech Using OCR”, Imperial Journal of Interdisciplinary Research (IJIR) , 2016.

[15] Yao Li and Huchuan Lu, “Scene Text Detection via Stroke Width”, 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.