Text Extraction From Image and Text to Speech Conversion

[1st] Prof. Teena Varma
Department Of Computer Engineering.
Xavier Institute Of Engineering Mumbai University
Mumbai,India

[2nd] Stephen S Madari
Department Of Computer Engineering.
Xavier Institute Of Engineering Mumbai University
Mumbai,India

[3rd] Lenita L Montheiro
Department Of Computer Engineering.
Xavier Institute Of Engineering Mumbai University
Mumbai,India

[4th] Rachna S Pooojary
Department Of Computer Engineering.
Xavier Institute Of Engineering Mumbai University
Mumbai,India

***Abstract*— The recent technological advancements in the field of Image Processing and Natural Language Processing are focusing on developing smart systems to improve the quality of life. In this work, an effective approach is suggested for text recognition and extraction from images and text to speech conversion. The incoming image is firstly enhanced by employing gray scale conversion. Afterwards, the text regions of the enhanced image are detected by employing the Maximally Stable External Regions (MSER) feature detector. The next step is to apply geometric filtering in combination with stroke width transform (SWT) to efficiently collect and filter text regions in an image. The non-text MSERs are removed by using geometric properties and stroke width transform. Subsequently, individual letter/alphabets are grouped to detect text sequences, which are then fragmented into words. Finally, Optical Character Recognition (OCR) is employed to digitize the words. In the final step we feed the detected text into our text-to-speech synthesizer (TTS) for text to speech conversion. The proposed algorithm is tested on images from documents to natural scenes. Promising results have been reported which prove the accuracy and robustness of the proposed framework and encourage its practical implementation in real world applications.**

Keywords*— *Image Processing, Text Recognition and Extraction, MSER (Maximally Stable Extremal Regions), OCR (Optical Character Recognition), SWT (Stroke Width Transform). TTS (text-to-speech synthesizer).

I. INTRODUCTION

Languages are the oldest way of communication between human beings whether they are in spoken or written forms. In the recent era, visual text in natural or manmade scenes might carry very important and useful information. Therefore, the scientists have started to digitize these images, extract and interpret the data by using specific techniques, and then perform text-to-speech synthesis (TTS). It is done in order to read the information aloud for the benefit and ease of the user. Text extraction and TTS can be utilized together to help people with reading disabilities and visual impairment to listen to written information by a computer system. In this work, a novel text detection framework is proposed which is based on connected component analysis and MSER algorithms are employed for extraction of CCs, which are taken as letter candidates. CCs that are likely to be characters are selected on the basis of their geometric properties and stroke width variation. Afterwards, the selected objects are grouped into detected text sequences, which are then fragmented into isolated words. Optical character recognition is employed to recognize and extract the words and finally the extracted text is converted to appropriate speech using text-to-speech synthesizer. The proposed algorithm is tested on images representing different scenes ranging from documents to natural scenes. Promising results have been reported which prove the accuracy and robustness of the proposed algorithm and encourage its practical implementation in real world scenarios.

Rest of the paper is structured as follows: Section II covers the background of the research problem addressed in this paper and related methods. The proposed algorithm is presented in Section III followed by experimental analysis in Section IV. Results of the experimental analysis are discussed in section V. Conclusion and future prospects of this research are summarized in Section VI. Acknowledgement is in section VII and section VIII contains the References

II. BACKGROUND

Text detection and recognition is a conventional problem that has been researched and constantly improved according to the increasing challenges in the images and videos on the web. Several methods of text extraction from images and videos have been suggested in the past few years. the methods can be broadly classified into two main types those are region-based approach and texture-based approach.

In region-based approach the properties of the textual region which distinguish them from the background are taken into consideration such as color, intensity, edges ,etc. Region-based method can further be subdivided into various categories such as edge-based, color-based, stroke-based and many other. Region-based method is a bottom-up approach where we detect small candidate regions and then group them into text regions. Whereas texture-based approaches follow top-down approach since they are based on the textual properties of the text such as statistical features, Gabor filters, frequency transform and many other. Although texture based method seems to be a better classification method of text/non text regions but the are

computationally demanding and sensitive to text alignment. Also the complex background and bad quality of the images can hamper the performance of texture based approach.

III. PROPOSED METHODOLOGY

In this work, we propose a robust MSER method to extract the text from images. The MSER regions are areas that have a relatively distinct intensity compared to their background contrast. They are retrieved through a process of attempting numerous thresholds. The regions that preserve constant shapes over a wide range of thresholds are selected. Segmenting the text from a scene via MSER intensively helps in further processing of image for detecting text regions. Once the MSER regions are detected those region are further processed using geometric properties, connected components and stroke width variation. Once the text regions are detected, the other non-text regions are removed. MSER is compatible with text due to the constant color and high contrast with the background, which together give us stable intensity profiles. However it is highly likely that a number of non-text regions that are stable are also selected. Geometric properties such as eccentricity, bounding box ,solidity, euler number are also taken into consideration for detection of text regions. Connected components within a region are also considered for detecting the region of interest. To remove the non-text regions the stroke-width is considered. Text characters tend to have little variation when it comes to stroke widths of the lines and curves, whereas non text areas display a high stroke width variance. So the regions that have high stroke width variation are removed as they more likely to be non-text regions. The detected text regions undergo OCR (Optical Character Recognition) for digitizing the text regions and to detect and extract the text from image. Finally the detected text is converted to speech using text-to-speech synthesizer. In our work we make use of the Microsoft text-to-speech system available for Windows.

Some important terminologies associated with this research problem are defined below:

a. Edge: Edge is a group of points having strong gradient magnitude in an image.

b. Corner (or Point of Interest): Corner is a group of points having a high level of curvature in the gradient in an image.

c. Region: A region is a contiguous set of adjacent pixels.

d. Blob (or Region of Interest): Blob is the area in which some properties (color, brightness, etc.) are invariant or slightly variant in an image, i.e. points in a blob are similar.

e. Boundary: Boundary of a region is the group of pixels neighboring at least one pixel of that region but not a part of that region.

f. Extremal Region: If all the pixels in a region have values greater than (or smaller than) that of the boundary, the region is called extremal region.

g. Maximally Stable Extremal Region (MSER): An extremal region is termed as maximally stable when its variation w.r.t. a given threshold is minimal.

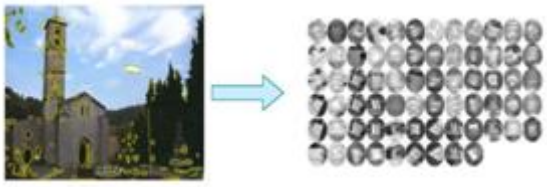


Fig. 1. Example of MSER regions

Flow chart of the proposed system

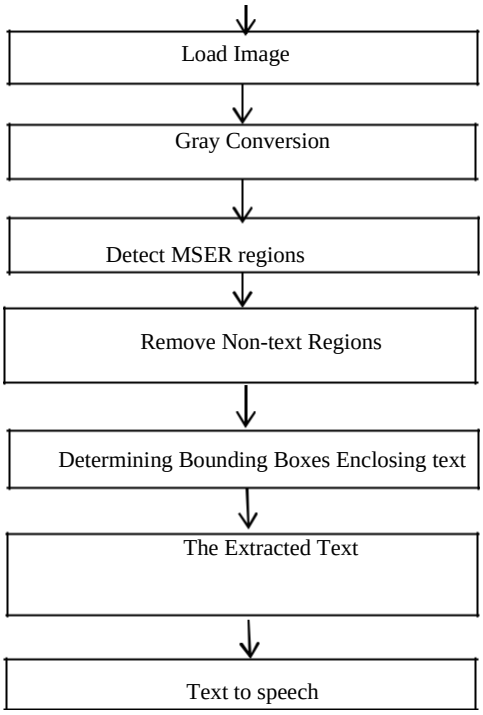


Figure 2. Architecture of proposed method

The various components of the proposed system are as follows

A. Gray scale conversion

The grayscale image is represented by using 8 bits value. The pixel value of a grayscale image ranges from 0 to 255. The conversion of a color image into a grayscale image is done by converting the RGB values (24 bit) into grayscale values (8 bit). One method of converting RGB to grayscale is to take the average of the contribution from each pixel (R+G+B)/3.

B. MSER Regions:

Maximally stable extremal regions are used as a method of blob detection in images. MSER regions are connected areas characterized by almost uniform intensity throughout a range of thresholds. The selected regions are those that maintain unchanged over a large set of thresholds.

C. Connected components:

Connected components of an image are the regions which have continuous pixels within that region. The pixels in the connected components are connected to each other through either 4-pixel, or 8-pixel connectivity.

D. Geometric properties

The following geometric properties are taken into consideration:

a. *Bounding Box*: Bounding boxes are rectangular boxes created around the region of interest. It contains all the pixel values within the enclosing boundary.

b. *Eccentricity*: The eccentricity is the ratio of the distance between its major axis length and the foci. The value should be between 0 and 1. An ellipse is said to be circle if its eccentricity value is 0 whereas if the eccentricity value is 1 then the ellipse is a line segment.

c. *Solidity*: Solidity also known as convexity of an image is the area of the image divided by area of its convex hull. It is the proportion of the pixels in the convex hull that are present in the region to the actual number of pixels in the image

d. *Extent* : Extent of an image is defined as the ratio of the pixels in the image to the number of pixels in the total bounding box in that image.

e. *Euler* : Euler number is defined as the total number of pixels in the image minus the number of holes in that region. Holes in a region indicates there are no pixels in the region. We can use either 4 or 8-connectivity.

E. Stroke width transform

A stroke in an image is a continuous band of a nearly constant width. As the name suggests stroke width variation calculates the width of the most likely stroke containing the pixel for each pixel in that stroke

F. OCR

OCR stands for Optical Character Recognition. As the name suggests OCR is used to detect the normal human readable language which may be present in the form of textual matter present in image or any documents or pdf files and convert it into editable formats.

G. Text to speech

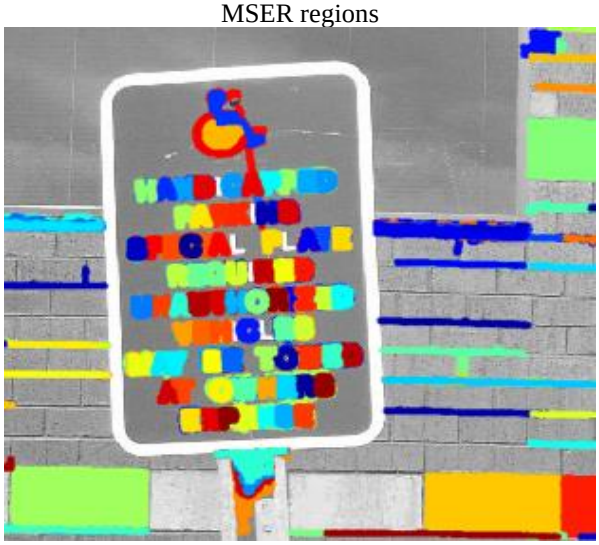
A text-to-speech (TTS) system converts the normal human readable language text into speech.

IV. EXPERIMENTAL ANALYSIS

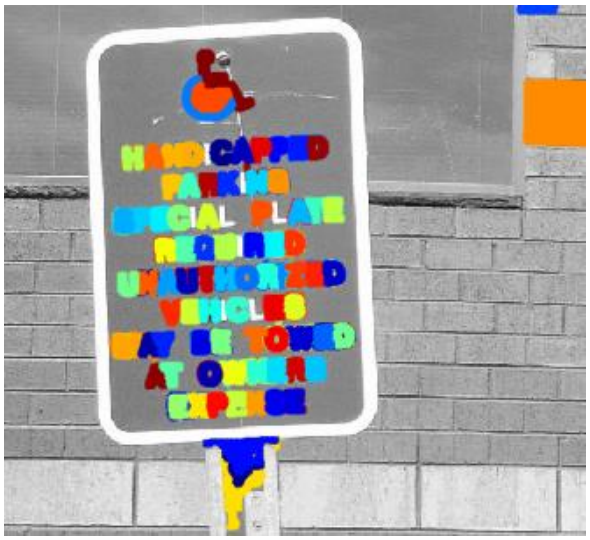
Original Image

Gray scale conversion

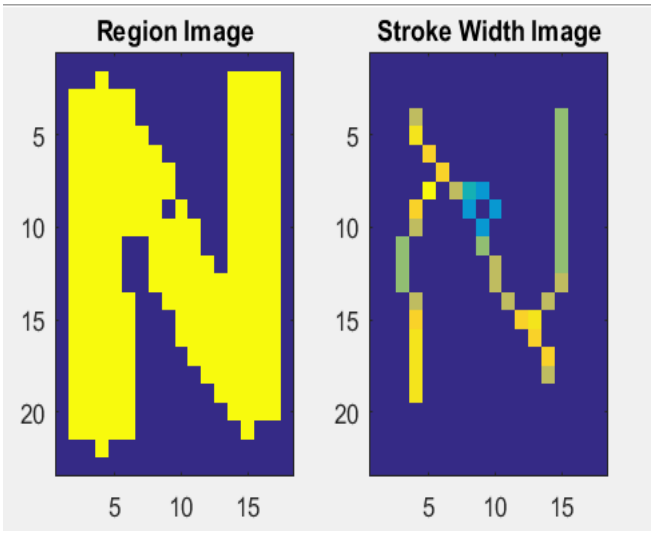
MSER regions



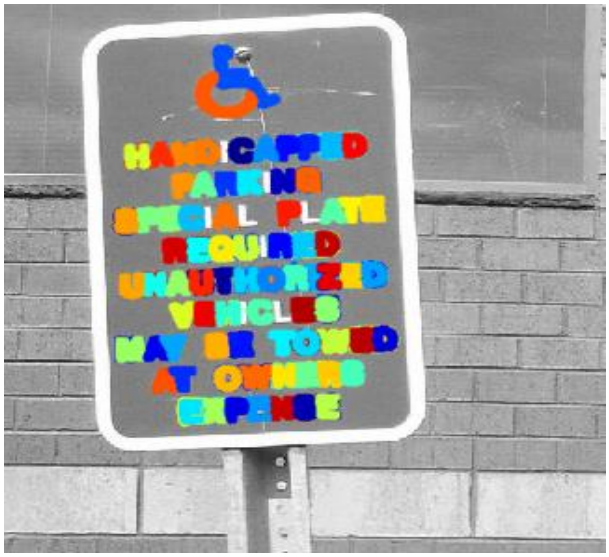
Removing non-text regions based on Geometric properties



Stroke width transform



Removing non-text regions based on Stroke width variation



Bounding boxes

Detected text

Extracted text

```
ans =

HANDICAPPED
PARKING
SPECIAL PLATE
REQUIRED
UNAUTHORIZED
VEHICLES
MAY BE TOWED
AT OWNERS
EXPENSE
```

V. RESULTS AND DISCUSSIONS

The proposed method successfully detects the text regions in most of the images and is quite accurate in extracting the text from the detected regions. Based on the experimental analysis that we performed we found out that the proposed method can accurately detect the text regions from images which have different text sizes, styles and color. Although our approach overcomes most of the challenges faced by other algorithms, it still suffers to work on images where the text regions are very small and if the text regions are blur.
Below is the word-confidences of the words that we retrieve after performing the optical character recognition on the image which is tested in the experimental analysis section of this paper.

TABLE I. ANALYSIS OF WORD CONFIDENCE

WORD	WORD CONFIDENCE
HANDICAPPED	0.6431808
PARKING	0.7250697
SPECIAL	0.8774834
PLATE	0.9105222
REQUIRED	0.7836785
UNAUTHORIZED	0.8626139
VEHICLES	0.8679733
MAY	0.8386983
BE	0.8924372
TOWED	0.8627644
AT	0.9176741
OWNERS	0.8632467
EXPENSE	0.8248840

Word confidence is a metric indicating the confidence of the recognition result. Confidence values ranges between 0 to 1 and should be interpreted as probabilities. As we can see from the above table that the words having fewer number have characters have a better word confidence than the words which comprises of more number of characters. The average word-confidence comes out to be 0.8361.

VI. CONCLUSION AND FUTURE WORK

Nowadays, there is increasing demand of text information extraction from image. So, many extracting techniques for retrieving relevant information have been developed. Moreover, extracting text from the color image takes time that leads to user dissatisfaction. In this paper we have proposed a method to extract the text from image which extracts text more accurately. Using our method it is possible to extract information within short time. Although, our connected component based approach for text extraction from color image method has several features than existing method but it becomes less effective when the text is too small and if the text region is not clearly visible or the color of the text is not visible clearly. In future, this work can be extended to detect the text from video or real time analysis and can be automatically documented in Word Pad or any other editable format for further use.

VII. ACKNOWLEDGEMENT

First of all we are grateful to Almighty God to give us the opportunity and capability to complete this project. Then we remember our parents who are always thinking of our success. While accomplishing this work we faced so many problems and to solve the problem we got help from various sources. We would like to express our sincere regard to our project guide Prof. Teena Varma for her co-operation, suggestion, guidance, supervision and continuous encouragement.

VIII. REFERENCES

[1]. Ranjit Ghosal, Ayan Banerjee ," An Improved Scene Text And Document Image Binarization Scheme",Recent Advances In Information Technology (RAIT) 2018.
[2]. Muhmmad Jaleed Khan, Naina Said, Aqsa Khan, Naila Rehman, Khurram Khurshid Automated Latin Text .
[3]. Satish Kumar, Sunil Kumar , Dr. S, Gopinath " Text Extraction From Images", International Journal Of Advanced Research In Computer Engineering & Technology, June 2012.
[4]. Nitin Sharma And Nidhi , "Text Extraction And Recognition From The Normal Image Using MSER Feature Extraction And Text Segmentation Methods." Indian Journal Of Science And Technology May 2017.
[5]. Amani Jamal, Noora Alhindi, Raghdah Nahhas, Somayh Al-Amoudi "Image Assistant Tools For Extracting, Detecting, Searching Images And Texts".2019.
[6]. Saeed Mian Qaisar†, Raviha Khan, Noofa Hammad "Scene To Text Conversion And pronounciation For Visually Impaired People" 2019.
[7]. B.Gatos, I.Pratikakis, K.Kepene And S.J. Perantonis, "Text Detection In Indoor/Outdoor cene Images" 2005.
[8]. Kamrul Hasan Talukderr, Tania Mallick "Connected Component Based Approach For Text Extraction From Color Image", International Conference On Computer And Information Technology (ICCIT) 2014.
[9]. Jaswant P ,Anusuya S, Anil Kumar M, Dhikhi T ," Enhanced Mser For Text Extraction", International Journal OfComputational Intelligence And Informatics ,March 2016.

.