International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Real Time Text Detection & Recognition System to Assist Visually Impaired

Chaitanya R. Kulkarni[1], Ashwini B. Barbadekar[2]

Abstract: ***Number of blind or visually impaired people in the world are constantly increasing due to reasons like diabetes, accidents, ageing etc. Along with navigation, they also needs assistance in reading various boards, labels etc. in their daily life. A reading assistance system based on text detection and recognition from natural scenes is proposed in this work. Camera based system detects text in natural images using Maximally Stable Extremal Regions algorithm. Detected text is recognized using Optical Character Recognition and converted into speech using Text-to-Speech synthesizer for blind user to listen through headphones. Proposed method is evaluated on ICDAR 2011 dataset.***

Keywords: Text detection, Recognition, MSER, TTS

1. Introduction

Approximately 39 million blind people are there worldwide whereas 285 million people are visually impaired according to World Health Organization. 19 million children are suffering with blindness or impaired vision. This number is increasing day-by-day due to diabetes, accidents, aging population and other reasons. Predictions suggest that there will be 8 million blind or visually impaired people in U.S. by 2050 which is twice the current count. Various studies suggest that count of blind or visually impaired people will increase at least until 2021 since main cause of this issue is age related.

Blind people need to use their other senses to overcome their loss of eyesight. Most of the times they rely on other people for information which should be received by eyes. Along with problems they face during navigation, inability to read is severe loss to blind people. Right from reading label of food packets to reading traffic boards blind or visually impaired people need assistance from another human being. Increasing amount of material is made available in Braille as well as in forms of audio books but reader must rely on choice made by others. Thus there needs to be a device to allow blind people to read text datafrom natural scenes in its original form. Reading Assistance device will help blind or visually impairedpeople to understand their surroundings in better way. It will allow them to read signboards, public notices, etc. without help from anyone. It will also assist themin identifying currency notes, reading labels etc. It will boost up their confidence to leave an independent life.

In this work, a reading assistance system is proposed as an application of text detection and recognition from natural scene images. Maximally Stable Extremal Regions (MSER) algorithm is used for detecting text regions from frames of video captured using camera mounted on user. Non-text components are removed using geometry-based filtering. Text regions obtained are applied to Optical Character Recognition (OCR) engine for recognition. Recognized text is given to user in form of audio through headphones by using Text-to-Speech synthesizer.

The paper is organized as follows. Section 2 presents previous work in this field. System design is presented in section 3 whereas text extraction and recognition procedure proposed in this work is explained in section 4. Section 5 focuses on analysis of experiments and results. Section 6 concludes the paper.

2. Previous Work

Research in the field of text detection and extraction from natural scenes is increasing day-by-day. But work done in the field of reading assistance applications for blind or visually impaired people has been limited. Ezaki et al. developed camera based reading assistance system for blind people[1] using various approaches for text detection like connected components extraction, mathematical morphology based approach, etc. Shoulder mounted camera detects text characters of small size using global and local binarization on three color channels, top-hat processing using erosion of thin text characters as well as using connected component extraction. Detected text character are zoomed in for recognition and recognized data is read out to blind person using voice synthesizer. In [2] a PDA based embedded reading device has been developed for blind people. Images captured with PDA camera by user undergo text recognition using Gabor filter based texture recognition technique. This is followed by denoising and binarization of recognized text area and character segmentation along with feedbacks from user after getting output to allow designer to improve proposed system.

Various approaches like connected component extraction, features based classification, stroke width transform, adaptive thresholding, etc. have been used to solve problem of text detection and recognition from natural scene images. Adaptive thresholding is used in [3] to segment scene images into regions, and text regions are extracted using difference between gray-level forms of adjacent regions. It is followed by character segmentation using similarities between standard properties of characters from dictionary and properties of character candidates. Features based classifiers have been used for text extraction in scene images in [4] and [5]. Intensity histogram is used for classification at bottom level in [4] to remove non-character regions which is followed by SVM classifier at next level using texture information obtained from Haar wavelet. Unsupervised learning approach is used in [5] in which system is trained using k-means clustering to detect text and to extract characters by using small sized image patches.

Stroke Width Transform (SWT) is applied to image, pixel-by-pixel in [6] to obtain text character candidates by applying text specific constraints to SWT information of image. These character candidates are then aggregated into lines of text by using characteristics of text lines. Huang et al. have proposed Stroke Feature Transform (SFT) in [7]which is extension of SWT with color cue of text pixels to separate inter-components and connectintra-components in text extraction process. Stroke colormap isused along with stroke width map for robust grouping of text pixels, whereas features derived from obtained stroke pixels are used for Text Covariance Descriptors (TCD) which is used to extracts text components as well as text lines. Connected Components (CC) Analysis is used in [8] and [9] for text extraction from natural scenes. In [8], Histogramof Oriented Gradients (HOG) and Multi-Scale Local Binary Pattern are used to obtain text features which is followed by cascade AdaBoost classifier to classify text and non-text. After classification, text lines are generated using grouping method and text is extracted precisely using CC analysis based on Markov Random Fields (MRF). In [9] text in Devnagari and Bangla Language is extracted using CC analysis using headlines of text which is peculiarity of both scripts. Image is initially binarized using Otsu's Binarization which is followed by text extraction using CC analysis.

It is observed that majority of previously proposed approaches focus on text detection and extraction rather than its applications. Thus main focus of these methods is accurate detection and recognition of text without consideration of computation time. However proposed method is for real-time application of reading assistance for visually impaired, where time is main constraint. Unlike other methods, inclusion of any training based classifier or other time-complex algorithms are avoided by using very fast MSER algorithm in proposed method which results into real-time performance of system.

3. System Design

Figure 1 represents general design of proposed system. Proposed system consist of a digital camera mounted on user that captures video of surrounding, from which frames are extracted. Frames are processed to detect text regions using Maximally Stable Extremal Region (MSER) algorithm that finds connected components in image. Non-text components are removed by applying constraints like Aspect ratio, Solidity, etc. Character regions obtained are grouped together to form words or group of words. Images of text words are applied to Optical Character Recognition (OCR) module to obtain text. This text is applied to text-to-speech synthesis stageto convert it into audio. Audio output is provided to user through headphones.

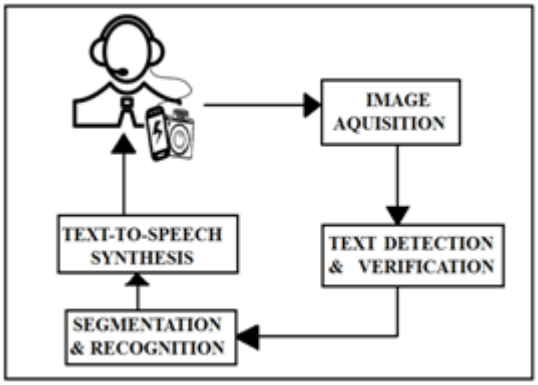

Figure 1: System Design

In proposed system, Raspberry Pi 3 model B processor is used for implementation of proposed text detection and recognition algorithm. It is 1.2 GHz quad-core processor with 1GB RAM. Raspberry Pi NoIR camera V2 module (8 MP) is interfaced to processor board which captures video with 1080p at the rate of 30fps. This video is converted into frames on which further processing is done using Raspberry Pi processor. Speech synthesized from recognized text is made audible to user through headphones.

4. Text Extraction and Recognition

In proposed system, MSER is used to detect text region from image (video frames). Non-text regions obtained from MSER algorithm is removed by applying geometry based filtering. Text characters obtained through this process are merged together to form words or group of words based on their positioning. This process gives text regions from image which are further applied to 'Tesseract' open-source OCR engine for optical character recognition which converts image of text into actual text. Recognized text is fed to 'eSpeakNG', an open-source speech synthesizer, which gives audio output.This audio output is fed to user through headphones which guides user about its surroundings.

4.1. Text Extraction using MSER

MSER algorithm is fast and robust algorithm for text detection. It basically finds out connected components in the image that are stable over a sequence of increasing threshold. It is independent of transforms of image intensities, scaling of image as well as illumination, which helps in better text detection in various conditions. Despite of favorable features of MSER, it is sensitive to image blur which makes it difficult to use for detection of very small sized character. Chen et al. have proposed a method in [10] to combine MSER with Canny Edge Detection to detect small sized character. However for Blind assistance application, major important text information is available in large sized text form which allows use of MSER algorithm in its simplest form in proposed work.

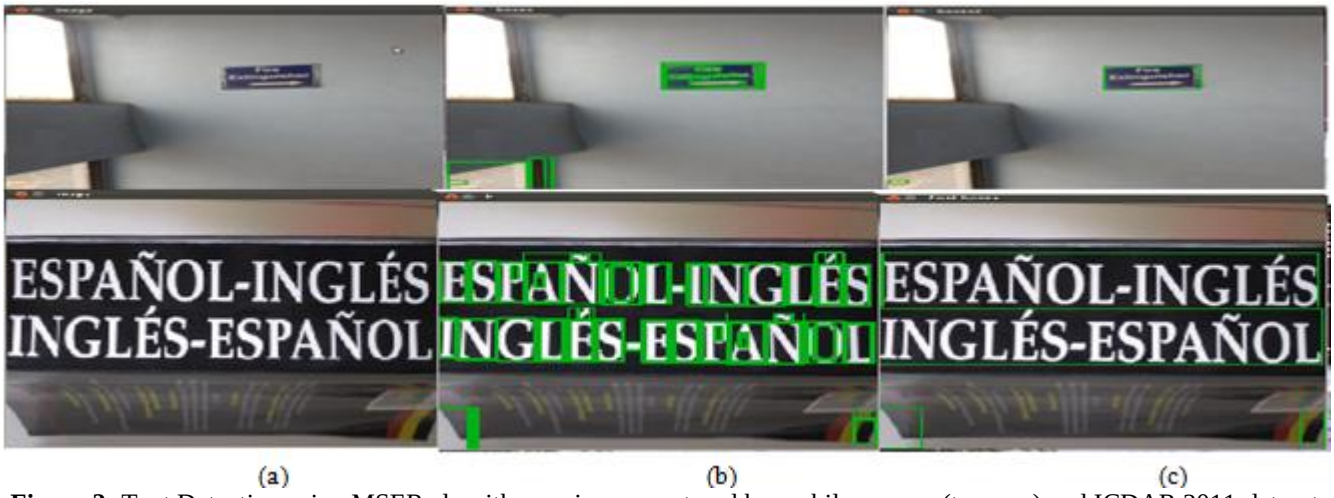

Figure 2: Text Detection using MSER algorithm on image captured by mobile camera (top row) and ICDAR 2011 dataset image (bottom row) .(a) Original Image (b) Detected MSER Regions (c) Text Regions obtained after applying proposed algorithm

4.2. Geometry Based Filtering

MSER gives few non-text components along with text candidates which needs to be filtered out before further processing. [11] performs local consistency analysis to find out consistency between neighbor regions and studies their projections to remove non-text candidates. Various geometry based filtering methods can also be used to remove non-text components. Geometry based filtering methods are used in this work. In proposed method, aspect ratio, extent and solidity are used to filter out non-text components. These are defined as follows.

$$\text{Aspect ratio} = \frac{\text{Width}}{\text{Height}}$$

$$\text{Extent} = \frac{\text{Component Area}}{\text{Area of Bounding Box}}$$

$$\text{Solidity} = \frac{\text{Component Area}}{\text{Area of Convex Hull binding the Component}}$$

These parameters help to remove non-text components from detected MSER regions. Figure 2 represents implementation of MSER algorithm on images containing text. Original image in figure 2.a) undergoes MSER detection as shown in figure 2.b). After applying Geometry based filtering and merging bounding boxes based on their positioning, text region is obtained as shown in figure 2.c).

4.3. Text recognition and Text-to-Speech Synthesis

Text regions obtained in previous stage is applied to 'Tesseract' open source OCR module for character recognition. Tesseract is Google funded OCR engine with accurate recognition results with support of 100 languages. Image needs to be scaled and preprocessed for this OCR to give better results. This OCR engine is tested on various manually captured images as well as on ICDAR 2011 Text localization dataset and good accuracy was observed.

After OCR engine performs text recognition, this text is fed to Text-to-Speech (TTS) synthesizer to obtain audio output. 'eSpeakNG' is a speech synthesizer used for this purpose, which is open source synthesizer based on formant synthesis method. It supports more than 80 languages in very small memory size. Audio output thus generated through TTS is fed to user through headphones as final output.

5. Experiments and Results

The proposed method has been evaluated on ICDAR 2011 dataset for text detection and recognition. It gives good detection and recognition results although there are few non-text regions getting detected along with text regions. However during recognition phase, these non-text regions are ignored by proposed OCR giving correct outputs for text regions only. Figure 3 represents implementation of proposed text detection algorithm on ICDAR 2011 dataset.

Parameters like Precision, Recall and F-measure are calculated using method described in [12]. Precision defines ratio of correct estimates to the total number of estimates however Recall is ratio of number of correct estimates to the number of targets. F-measure combines Precision and Recall into single quality measure by assigning equal weights to them. Important advantage of proposed system is its fast detection and recognition rate compared to previous approaches. Proposed system performs both detection and recognition in average 0.9 sec per image on ICDAR 2011 dataset. Computation time is calculated by computing time difference between start and end instances of program. Python module 'time' was used for this purpose.

Computation time = time (end instance) – time (start instance)

Figure 3: Implementation of proposed text detection method on ICDAR 2011 dataset

Table 1: Performance analysis of proposed method and existing methods using ICDAR 2011 Dataset

Method	Precision	Recall	F-measure	Avg. time (s)
Proposed Method	0.707	0.832	0.757	0.9
Risnumawan et. al	0.83	0.71	0.77	13.9
Neumann and Matas (2012)	0.73	0.65	0.69	1.8
Neumann and Matas (2013)	0.854	0.675	0.754	0.6
Shi et. al (2013)	0.833	0.631	0.718	1.5

Table 1 compares various parameters like Precision, Recall, F-measure and computation time of proposed approach with previous approaches. It can be seen that proposed system provides better recall rate compared to others which indicates that proposed system detects more text from ground truth compared to others. However due to presence of small amount of non-text components, precision rate is little affected. However this problem is handled by OCR in recognition stage by avoiding recognition of non-text regions which satisfies necessity of proposed application. Since proposed application is reading assistance for visually impaired, time is more crucial parameter than others and proposed system performs better than previous ones.

Table 2 compares time required for text detection using proposed algorithm for two different datasets. Our dataset consists of scene images captured with mobile camera as well as raspberry-pi camera. Average text area is less than 25 % of total image area in our dataset, whereas ICDAR dataset has more than 30 % image area as text area. This results into faster computation for our dataset compared to ICDAR 2011 dataset as shown in Table 2.

Table 2: Time analysis of proposed algorithm with 2 different datasets

Datasets	Avg. time (s)
ICDAR 2011 Dataset	0.9
Our Dataset	0.4

6. Conclusion and Future Scope

Reading assistance system for blind or visually impaired people is developed as an application of Text detection and Recognition from natural scenes imagery in the proposed work. Use of MSER algorithm has allowed fast detection of text regions which helps to achieve near real-time text reading from images. In future, system can be modified for document analysis, text reading assistance for unmanned vehicle, label recognition system, and many more text recognition applications.

References

[1] Nobuo Ezaki1, Kimiyasu Kiyota2, Bui Truong Minh3, Marius Bulacu4 and Lambert Schomaker.Improved Text-Detection Methods for a Camera-based Text Reading System for Blind Persons.*Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)*

[2] Jean-Pierre PETERS, Céline THILLOU, Silvio FERREIRA. Embedded Reading Device for Blind People: a User-Centered Design.*Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop (AIPR'04)*

[3] Jun Ohya, Akio Shio, and Shigeru Akamatsu. Recognizing Characters in Scene Images.*IEEE transactions on pattern analysis and machine intelligence, vol. 16, no. 2, february 1954*

[4] Takuma Yamaguchi and Minoru Maruyama.Character Extraction from Natural Scene Images by Hierarchical Classifiers.*Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04) IEEE*

[5] Adam Coates, Blake Carpenter, Carl Case, SanjeevSatheesh, Bipin Suresh, Tao Wang, David J. Wu, Andrew Y. Ng. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning.*2011 International Conference on Document Analysis and Recognition*

[6] Boris EpshteinEyalOfek Yonatan Wexler.Detecting Text in Natural Scenes with Stroke Width Transform.*IEEE 2010*

[7] Weilin Huang, Zhe Lin, Jianchao Yang, and Jue Wang.Text Localization in Natural Images using Stroke Feature Transform and Text Covariance Descriptors.*ICCV 2013, IEEE*

[8] Yi-Feng Pan, XinwenHou, Cheng-Lin Liu. A Robust System to Detect and Localize Texts in Natural Scene Images.*The Eighth IAPR Workshop on Document Analysis Systems, IEEE 2008*

[9] U. Bhattacharya, S. K. Paruiand S. Mondal. Devanagari and Bangla Text Extraction from Natural Scene Images.*10th International Conference on Document Analysis and Recognition, IEEE 2009*

[10] Huizhong Chen, Sam S. Tsai, Georg Schroth, David M. Chen, RadekGrzeszczuk and Bernd Girod. Robust text detection in natural images withedge-enhanced maximally stable extremal regions. *18th IEEE International Conference on Image Processing, 2011*

[11] Le Kang, Yi Li, and David Doermann.Orientation Robust Text Line Detection in Natural Images.*CVPR 2014, IEEE*

[12] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young. ICDAR 2003 Robust Reading Competitions.

[13] Shi, C., Wang, C., Xiao, B., Zhang, Y., &Gao, S. (2013). Scene text detection usinggraph model built upon maximally stable extremal regions. *Pattern Recognition Letters, 107–116.*

[14] Neumann, L., &Matas, J. (2012). Real-time scene text localization and recognition.*In Proceedings of the CVPR (pp. 3538–3545).*

[15] Neumann, L., &Matas, J. (2013). On combiningmultiple segmentation in scene textrecognition. *In ICDAR 2013 (pp. 523–527).*

[16] AnharRisnumawan, Palaiahankote Shivakumara, Chee Seng Chan, Chew Lim TanA robust arbitrary text detection system for natural scene images. *Elsevier, 2014*