IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105285 |

Image to Text to Speech Using Machine Learning

Siddhesh Pisal, Chaitanya Patil, Shlok Jain, Akash Kumar, Ms. Ashwini Gavade.

UG Student, Department of Information Technology, Sinhgad Institute of Technology and Science, Narhe, Pune, India

Professor, Department of Information Technology, Sinhgad Institute of Technology and Science, Narhe, India

ABSTRACT: This project consolidates the idea of Image Text to Speech Synthesizer (TTS) and Optical Character Recognition (OCR). This sort of framework assists visually impaired people by connecting with computers successfully through vocal interface. Text-toSpeech conversion is a strategy that scans and reads 38+ languages and numbers that are in the image utilizing OCR method and transforming it to voices. This project implements two modules, voice processing module and image processing module. There are many techniques, for example, Edged Based method, connected component Method, texture Based Method, Mathematical Morphology Method is been utilized previously, yet they have some restrictions when estimated by exactness, f-score and re- view. Image Text is the content data installed or written in Image of various structure. Image text can be found in magazines, captured images, newspapers, banners and so on These image texts are exception- ally accessible these days and they are vital in addressing, describing and moving data which help people in communication, accessibility, making of new sorts of jobs, cost viability, efficiency, tackling issues, globalization and so on.

KEYWORDS: Text to Speech, Text Extraction, OCR, GTTS.

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 textto-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.

II.BACKGROUND

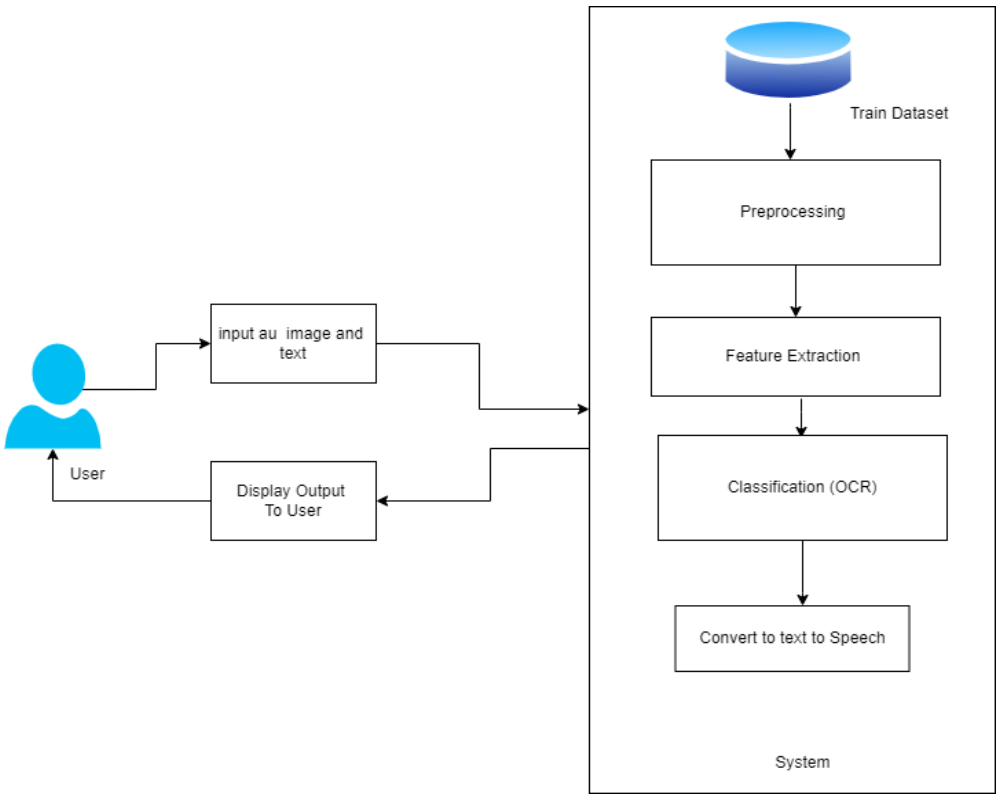
Text detection and recognition is a conventional problem that has been researched and constantly improved according to the increasing challenges in the images and videos on the web. Several methods of text extraction from images and videos have been suggested in the past few years. The methods can be 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, frequency transform and many other. Although texture based method seems to be a better classification method of text/non text regions, Also the complex background and bad quality of the images can hamper the performance of texture based approach .

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105285 |

III.SYSTEM ARCHITECTURE



IV.METHODOLOGY

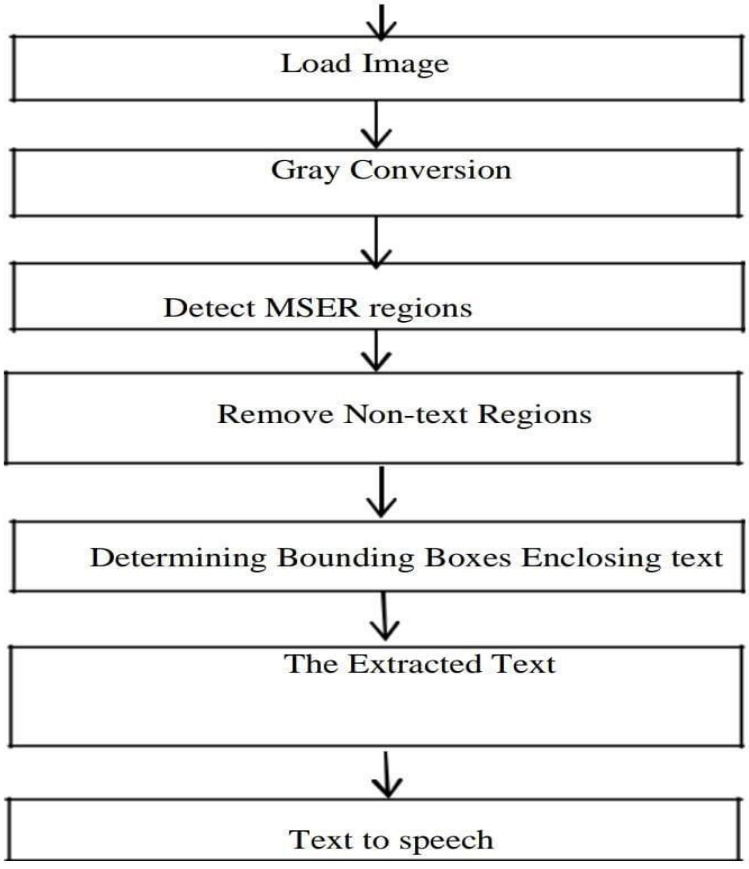
Optical Character Recognition (OCR) is the process of detecting and reading text in images through computer vision. Detection of text from document images enables Natural Language Processing algorithms to decipher the text and make sense of what the document conveys. Furthermore, the text can be easily translated into multiple languages, making it easily interpretable to anyone. OCR, however, is not limited to the detection of text from document images only. Novel OCR algorithms make use of Computer Vision and NLP to recognize text from supermarket product names, traffic signs, and even from billboards, making them an effective translator and interpreter. OCR used in the wild is often termed as scene text recognition, while the term "OCR" is generally reserved for document images only. Optical character recognition is a technology that converts typed or handwritten text and printed images containing text into machine-readable digital data format. OCR algorithms help turn large amounts of paper documents into digital files, facilitating text storage, processing, and searching. 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 areselected. 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-

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105285 |

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:
• Edge: Edge is a group of points having strong gradient magnitude in an image.
• Corner (or Point of Interest): Corner is a group of points having a high level of curvature in the gradient in an image.
• Region: A region is a contiguous set of adjacent pixels.
• 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.
• 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.
• 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. • Maximally Stable Extremal Region (MSER): An extremal region is termed as maximally stable when its variation w.r.t. a given threshold isminimal. The various components of the proposed system are as follows:-
• 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• 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.
• 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.
• Geometric properties The following geometric properties are taken into consideration:



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105285 |

– Bounding Box: Bounding boxes are rectangular 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.
– 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
– 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.
– 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.
• 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
• 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.
• Text to speech :- A text-to-speech (TTS) systemconverts the normal human readable language text toSpeech
boxes created around the region of interest. It contains all the pixel values within the enclosing boundary.
– Eccentricity: The eccentricity is the ratio of the

IV.CONCLUSION

Text to speech conversion is a fast-growing aspect of computer technology and has become an important criterion in determining the way we interact with the system and interfaces across a variety of platforms. This Images intends to propose an approach for image to speech conversion using optical character recognition and speech synthesis. The application developed is simple to use, very cost effective, portable and applicable in the real time. Using it we can read text from any type of images, and also it can generate synthesized speech through a mp3 file format. The developed software has incorporated features like word meaning assistance and voice modulation along with speed control. This can enable user to multitask and save time by listening to background materials while doing some other tasks. This system can be used in parts as well. For instance, if we want just the text conversion it is also possible moreover only text to speech can also be performed separately. Expensive hardware component, support software, updated Operating System version or even internet connection is not required..

REFERENCES

[1]AlexandreTrilla and Frances Al´ıas. (2013), “Sentence Based Sentiment Analysis for Expressive Text-to-Speech”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, Issue. 2. pp. 223-233.
[2]Al´ıas F. Sevillano X. Socoro J. C Gonzalvo X. (2008), Towards high quality ´ nextgeneration text-to-speech synthesis, IEEE Trans. Audio, Speech, Language Process, Vol. 16, No. 7. pp. 1340-135
[3] Balakrishnan G. Sainarayanan G. Nagarajan R. and Yaacob S. (2007) Wearable realtime stereo vision for the visual
[4] Chucai Yi. YingLiTian.AriesArditi. (2014), Portable Camera-based Assistive Text and Product Label Reading from Hand-held Objects for Blind Persons,IEEE/ASME Transactions on Mechatronics, Vol. 3, No. 2, pp. 1-10
[5]Deepa Jose V. and Sharan R. (2014), A Novel Model for Speech to Text Conversion, International Refereed Journal of Engineering and Science (IRJES) Vol.3, Issue.1, pp. 39- 41
[6]Goldreich D. and Kanics I. M. (2003), Tactile Acuity is Enhanced in Blindness, International Journal of Research And Science, Vol. 23, No. 8,pp. 3439–3445.
[7]JoaoGuerreiro and Daniel Gonc¸alves (2014), Text-to Speech: Evaluating the Perception of Concurrent Speech by Blind People, International journal of computer technology, Vol. 6, No. 8, pp. 1-8.
[8] J. Liang D. and DoermannH. (2005), Camera-based analysis of text and documents: a survey, International Journal on Document Analysis and Recognition, Vol.7, No-6, pp.
[9] Chucai Yi. YingLiTian.AriesArditi. (2014), Portable Camera-based Assistive Text and Product Label Reading from Hand-held Objects for Blind Persons,IEEE/ASME Transactions on Mechatronics, Vol. 3,No. 2, pp. 1-10

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details