I. J. Engineering and Manufacturing, **2023, 4, 44-49**
Published Online on August 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2023.04.05

An Efficient Approach for Text-to-Speech Conversion Using Machine Learning and Image Processing Technique

Swaroopa Shastri
Assistant Professor, Department of CSE (MCA), Visvesvaraya Technological University, Centre for PG Studies, Kalaburagi, India
E-mail:Swaroopam.vtu@gmail.com
ORCID iD: https://orcid.org/0000-0002-8897-5878

Shashank Vishwakarma*
Student, Department of CSE (MCA), Visvesvaraya Technological University, Centre for PG Studies, Kalaburagi, India
E-mail: shashankmv17@gmail.com
ORCID iD: https://orcid.org/0000-0001-9792-4960

Received: 11 September, 2022; Revised: 16 October, 2022; Accepted: 15 November, 2022; Published: 08 August, 2023

Abstract: This study explores the conversion of English to Hindi, first to text, and subsequently to speech. The first part of the implementation is the text recognition from images, in which two approaches are used for text character recognition: a maximally stable extensible region (MSER) and grayscale conversion the second part of the paper deals with the geometric filtering in combination with stroke width transform (SWT). Subsequently, letter/alphabets are grouped to detect text sequences, which are then fragmented into words. Finally, a 96 percent accurate spell check is performed using naive Bayes and decision tree algorithms, followed by the use of optical character recognition (OCR) to digitize. The word Give our text-to-speech synthesizer (TTS) the recognized text to convert it to Hindi language using the text-to-speech model. Based on aspects such speech speed, sound quality, pronunciation, and clarity.

Index Terms: Image processing MSER, OCR, Geometrical properties, SWT, TTS Synthesizer.

1. Introduction

There is a picture anywhere around us, and we see the picture and examine our daily life's textual content. Like bus numbers, inn names, newspapers, etc. however the query is how visually impaired or blind humans can understand this textual content. Virtually they want a few helps to examine the textual content. This paper says a look at of powerful prototype device to assist visually impaired humans' self -governing. At gift, there is plentiful TTS and text extraction can be used in conjunction to enable computer-assisted reading for those with visual impairment and reading difficulties. In this work, a novel connected component analysis-based text identification framework is provided. For extracting CCs, which are regarded as letter candidates, MSER algorithms are used. Based on their geometric characteristics and variation in stroke width, CCs that are likely to be characters are chosen.

The selected objects are subsequently sorted into words and text sequences that have been detected. The words are identified using optical character recognition, and then the recovered text is turned into the appropriate voice using a text-to-speech synthesiser. Assistive technology gadgets together with transferring chairs, vibrating watches, and talker tools are not unusual to place examples. Images with transcripts act as valuable communique mediums for conveying data to blind humans. Reading is one of the maximums not unusual place in the future societies, however, the low visually peoples are examining the textual content is a maximum tough one in today. Speech is one of the oldest and maximum herbal ways of fact change among humans. Over the years, Attempts had been made to expand vocally interactive computer systems to realize voice/speech synthesis. Such an interface could yield extraordinary benefits. In this example unique, the tech of a laptop can synthesize textual content and deliver a speech.

Text-To-Speech Synthesis is a Technology that offers a method of changing written textual content from a descriptive shape to a spoken language this is effortlessly comprehensible via way of means of the quit consumer (Basically in the English Language). It runs on the PyCharm platform, and turned into Object Oriented Analysis and

Development Methodology; whilst Expert System turned into integrated for the inner operations of the program. This layout might be geared closer to supplying a one-manner communique interface wherein the laptop communicates with the consumer via way of means of studying out a textual report for the cause of short assimilation and studying development. As part of Xerox PARC's work in speech and text image processing, Xerox PARC has expanded and redefined the role of recognition technology in document-oriented applications in two different ways. One is the development of systems that provide functionality similar to that of text processors but operate directly on audio and scanned image data. Using speech as an interface to systems that work primarily with text-based material has been the focus of spoken language processing to document creation and information retrieval. Speech and text-picture popularity to retrieve arbitrary, user-specified data from files with sign content.

The focus is once more on partial record fashions which might be described in only sufficient elements to fulfill mission requirements. Depending upon the application, layout conversion can also be added might not be the desired goal. For example, in retrieving applicable quantities of a lengthy audio record through keyword spotting, an easy time index is sufficient. On the alternative hand, extracting numerical data from tabular snapshots facilitate online calculations and requires at the least partial transcription into symbolic form. The textual content may be executed via way of means of Optical man or woman popularity (OCR). The output of this device may be manipulated via way of means of a computer. For image popularity and information access with inputs from information data and published files along with commercial enterprise cards, resumes, passports, income invoices, and financial institution statements are extensively used. A textual content report is generated as an output via way of means of spotting the characters in a photo or scanned report with the assistance of a program. Extracting the text from the Audio Files Is a Demanding Application in Many Fields. Speech synthesis is the production of human voice or speech by a machine. It is mostly used to convert written information into spoken information for convenience.

By extracting the text from the audio files, sentiment analysis can be performed. Speech Recognition (Speech to Text Conversion) is the process of extracting or converting audio to text. Dragon Dictation, Ever Note for Android, and Voice Assistant are popular speech-to-text conversion applications (Boris). The Conversion of Speech to Text by Using Computers Is Known as Automatic Speech Recognition (Asr) Or Compute Speech Recognition or Speech to Text (Speech Recognition).

2. Literature Review

In [1], the Authors of This Study Are Niblack, W Et Al. The Query by Image Content (Qbic) Project Investigates Ways to Query Big Online Image Collections Based on The Content of The Pictures Themselves. Content Examples Include the Color, Texture, And Form of Individual Visual Objects and Areas. Medicine, Photography, Retail, And Business Are Just A Few Of the Many Applications for These Images. There Are Several Uses for Them, Including in The Medical and Photography Sectors, And Retail. An Essential Question Is How to Extract and Calculate Relevant Properties from Photos and Objects., How to Retrieve Information Based on Similarity Rather Than Exact Match, How to Conduct Searches Using Images As Examples Or User-Drawn Images, And How To Design User Interfaces That Support Query Refinement And Navigation. As Of Right Now, We've Constructed an Rs/6000 Prototype System In X/Motif and C With A Test Database Of Over 1,000 Photographs And 1,000 Objects Sourced From Commercially Available Photo sip Art. Throughout This Project, We Describe And Demonstrate The Major Algorithms We Utilize For Color Texture, Shape, And Sketch Queries; We Also Address Plans.

In [2], the Authors Discuss the Text Region from The Complex Background and Give A High-Quality Input To The OCR. The Text, Which Is the Outcomes of The OCR Is Given to The Tts Engine Which Provides the Speech Output.

In [3], the Authors Talk About the Devanagari Script Evolved from The Brahmi Script, Which Is Regarded as The First Writing System in Ancient India, According to Some. Deva (God) And Nagari (City) Are Sanskrit Terms That Scholars Say Correctly Combine to Become 'Scythe Script of Gods,' Or, More Simply, The "City of The Gods." It Is Used to Recognize Hindi Terms from Bilingual or Multilingual Texts using a Support Vector Machine. Structural and statistical aspects of the differentiated words are used to segment, which leads to the detection of GHIC segmented characters. To determine how far apart two photos are, the Harsdorf distance is used. It is more accurate to recognize perfect pictures than to recognize images that are unionized.

In [4], the authors discuss the OCR and TTS synthesizers were actualized to extricate the content data from images and convert it into sound.

A.K. et al [5] extract just text areas from images. Useful for a variety of applications, including database indexing and electronic document conversion. It is impossible to execute classic text localization algorithms on mobile phones due to the intrinsic complexity of natural settings, particularly in multi-context situations.

In [6], the authors discuss an extract just text areas from images. Useful for a variety of applications, including database indexing and electronic document conversion. It is impossible to execute classic text localization algorithms on mobile phones due to the intrinsic complexity of natural settings, particularly in multi-context situations.

In [7], this approach can efficiently distinguish the object of attention from the background or other objects in the camera view. OCR is used to make word identification on the surrounded text fields and transform it into speech output for blind users.

In [8], the authors talk about Wolf, C. It's not difficult to derecognize photographs using the approaches that already exist. Texture estimation and edge detection are used by most of them, and then these characteristics are gathered together. All except one of the approaches enforces geometrical limitations. Only a morphological post-processing step performs this function. To fix a poor detection in a post-processing phase is very difficult, if not impossible. As part of the detection process, we offer a text model that incorporates geometrical restrictions directly into the text "probability" picture. After that, we determine the geometric characteristics of the surrounding text area for each pixel.

In [9], there's a novel approach to word recognition in video frames that relies on the BANN and CED operators. Using a novel color image edge operator (CED), the picture is first segmented into basic candidate text blocks. To further classify video frames into text and non-text blocks, a neural network is then used. Bootstrap is used in the neural network's training to increase its performance. This strategy is beneficial to the outcomes of experiments.

The authors of [10], this approach can efficiently distinguish the object of attention from the background or other objects in the camera view. OCR is used to make word identification on the surrounded text fields and transform it into speech output for blind users.

The creators of this paper found that many papers zeroed in on one of the two cycles Text to Discourse. They likewise found that the dialects chipped away at were for the most part unfamiliar and Indian Provincial dialects. Just a single paper utilized brain organizations, one more utilized guileless Bayes machines, and the rest utilized manual strategies for classification, the primary person of double planning techniques.

3. Methodology

MSER, OCR is used to convert text to Images and text to speech. Additionally, it can automatically detect speech and recognizes the content as text. This work has the potential to assist the masses who may lack the necessary education to write fluently and correctly. Converts the audio speech to text format. Furthermore, the proposed system's performance is assessed after it has been integrated with the proposed English spell corrector mechanism, resulting in a significant improvement over the previous method. The same contents convert to Hindi. Using Naive Bayes and Decision Tree algorithms for accuracy and spell check purposes. Dataset Creation and Preprocessing.

3.1 Data Set

A dataset in machine learning is, quite simply, a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes. This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do. For this, after collecting the data, it's important to preprocess it by cleaning and completing it, as well as annotate the data by adding meaningful tags readable by a computer**.**

The text extraction from images the data set consists of 239 images for testing to extract images from text conversion of text to speech and spell check the purpose.

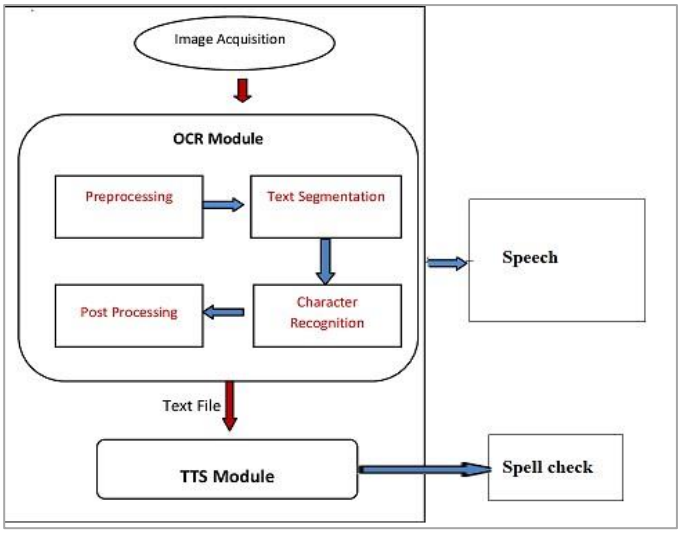

Fig. 1. Flowchart of the proposed methodology

3.2 Image Processing:

To create a more appealing image or extract useful information from an image, image processing involves turning it into a virtual factor and applying positive attributes. It's a kind of significant time when the input is an image, combined with a video body or image, and the output is an image or function connected to that image. In most cases,

the AWS picture processing tool treats photos as if they were identical symbols, even when using the predefined techniques.

1. Preprocessing

The grayscale picture is addressed by utilizing 8 pieces esteem. The pixel worth of a grayscale picture goes from 0 to 255. The change of a variety picture into a grayscale picture is finished by changing over the RGB values (24 digit) into grayscale values (8 bit). One technique for changing RGB over completely to grayscale is to take the normal of the commitment from every pixel (R+G+B)/3.

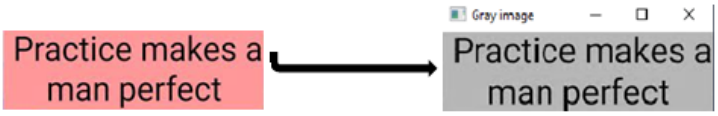

Fig. 2. Image before and after grayscale conversion.

A. MSER

Maximally stable extremal regions (MSERs) are used to identify blobs in photographs. This method became proposed by friends et al. To find a connection between various aspects of a photograph taken from a variety of perspectives. The wide-baseline matching is made possible by this method of extracting a large number of relevant image elements, and it has led to improvements in stereo matching and item popularity algorithms. The MSER item assesses the version of the place vicinity length among depth thresholds.

B. Stroke width Transformation

The stroke width transform (SWT) is a well-known operation for the undertaking of detecting texts from herbal photos due to the fact the characters intrinsically have an elongated form of almost uniform width. The side pairing method turned into lately evolved through Epshteyn et al. and is popularly used because of its simplicity and effectiveness in the single-sided image. For double-sided images, various methods were employed to differentiate between Recto and Verso dots for the final Braille character segmentation process.

C. Geometrical properties

A stable frame or particle may have geometrical qualities that may be determined by its geometry. They are very important because they allow the size and shape of an abnormally formed particle to be easily measured.

D. OCR

OCR stands for "Optical Character Recognition." It is a generation that acknowledges textual content inside a virtual photograph. It is generally used to understand textual content in scanned files and images.

4. Machine Learning Algorithms

4.1 Decision Tree

In the extended family tree of supervised learning algorithms, the decision tree method is a close cousin. Like other supervised learning methods, the decision tree may be used to address regression and categorization problems. When a Decision Tree is used, the purpose is to develop a learning model for predicting the magnitude or cost of the target variable based on past data gathered through examining simple choice rules (schooling data). In Decision Trees, the root of the tree is used to forecast a category label for a document. When comparing a document's characteristic to its base characteristic, we arrive at an evaluation. A comparison of costs leads to us taking a similar path and moving on.

4.2 Na ïve Bayes

The Bayes theorem provides the foundation for the naive Bayes algorithm, which is a collection of principles for supervised learning based on the Bayes theorem. Textual material that includes a high-dimensional education dataset is a unique use case. As one of the simplest and only Classification algorithms, the Naive Bayes Classifier may be used to rapidly construct a device that can learn from models and predict future events. An unsupervised learning collection of rules, the Naive Bayes set of rules is based on the Bayes theorem and used to solve type issues.

5. Results and Discussions

The model is developed using python programming language. PyCharm IDE is used as a platform for building the model. Integrated Development Environment (IDE) provide a large number of fundamental instruments for Python designers, firmly coordinated to establish a helpful climate for useful Python, web, and information science

improvement. Standard alone application provides the option of loading the text images and after When the text images are dropped or clicked on browse file button then it will file explore and it converts gray scale and extract text from images.then extract from text. This is the entire process which needs to be followed for text to speech and showing text from English to hindi. The overall accuracy of the model which we have achieved is 100%, which is considerably a good accuracy and na ïve bayes and decision tree algorithm applied for spell checking purpose.

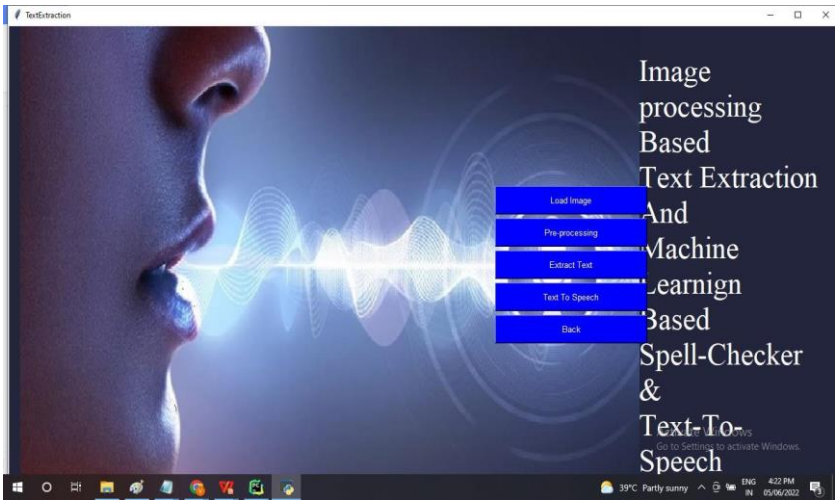

Fig. 3. Home Page

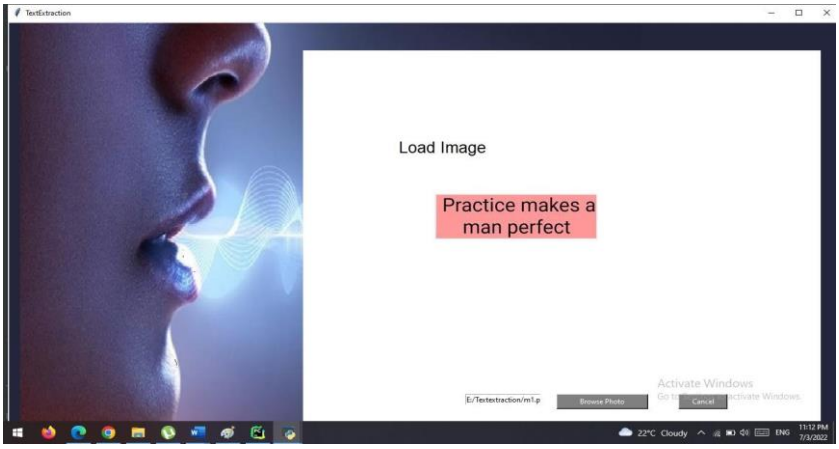

Fig. 4. Loading Images

```
C:\Users\dell\AppData\Local\Programs\Python\Python37\pyt
Practice makes a
man perfect


0
Accuracy:100%
original text: Practice makes a
man perfect


अभ्यास मनुष्य को परिपूर्ण बनाता है
corrected text: Practice makes a
man perfect



Process finished with exit code 0
```

Fig. 5. Result

6. Conclusion and Future Enhancement

Multiple strategies were used in this research to obtain crucial data. Additionally, it takes a while to extract textual information from the visual image, which irritates the user. In this paper, we present a method for extracting text from pictures that recovers text information more correctly than earlier methods. We apply our Maximally Stable Extensible Region method to extract text from photographs (MSER). When compared to their historical counterparts, the MSER regions have a notably rich depth. The value of 8 bits is used to signify grey scale conversion. Its changing-color image transforms into grey sun sunglasses with no discernible color. If there is text in a photo, document, or pdf file, that text can be extracted. 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 future use.

References

[1] Niblack, W. 1993. The QBIC Project: Querying Images by Content Using Color, Texture, and Shape. In Proc. Storage and Retrieval for Image and Video Databases, SPIE Bellingham, Wash,173-187

[2] Asha G. Hagargund, Shasha Vanaria Thota, Mitadru Bera, Eram Fatima Shaik (2017) "Image to Speech Conversion for Visually Impaired", International Journal of Latest Research in Engineering and Technology, ISSN: 2454- 5031, Issue 06, Vol. 03, No. 0, pp. 09-15.

[3] A. V. Bapat and L. K. Nagalkar, "Phonetic Speech Analysis for Speech to Text Conversion," 2008 IEEE Region 10 and the Third International Conference on Industrial and Information Systems, 2008, pp. 1-4, DOI: 10.1109/ICIINFS.2008.4798390.

[4] Kiran Rakshana R, Chitra C(2019) "A Smart Navguide System for Visually Impaired", International Journal of Innovative Technology and Exploring Engineering, ISSN:2278- 3075, Issue 6S3, Vol. 8, No. 0, pp. 0.

[5] Jain, A.K., and Yu, B. 1998. Automatic Text Lo cation in Images and Video Frames, Pattern Recognition Society. Vol. 31(12), 2055-2076.

[6] Wolf, C., and Jo lion, J.M. 2004. Model-Based Text Detection in Images and Videos: A Learning Approach. Technical Report LIRIS RR.

[7] Vaibhav V. Govekar, Meenakshi A (2018) "A Smart Reader for Blind People", International Journal of Science Technology & Engineering, ISSN: 2349-784X, Issue 1, Vol.5, pp. 0.

[8] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation," 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2020, pp. 439-444, DOI: 10.1109/CISPBMEI51763.2020.9263564.

[9] Hao, Y., Yi, Z., Zeng-Guang H., and Min, T. 2003. Automatic Text Detection in Video Frames Based on Bootstrap Artificial Neural Network and CED. Journal of Winter School of Computer Graphics (WSCG), Vol. 11.

[10] Misran, C., and Swain, P.K. 2011. An Automated HSV-Based Text Tracking System from Complex Color Video. LNCS, Vol 6536, 255-26

Authors' Profiles

Ms. Swaroopa Shastri working as an Assistant Professor in the Department of Computer Science and Engineering(MCA) at Visvesvaraya Technological University, Centre for PG Studies, Kalaburagi. She is in the field of Computer Appications at VTU's, CPGS, Kalaburagi, Karnataka, India. She is in teaching profession for more than 10 years. She has presented/published 20 papers in National and International Journals and Conference. Her main area of interest includes Machine Learning, Image Processing and Artificial Intelligence.

Shashank Vishwakarma is a student studying Master of Computer Application at Visvesvaraya Technological University, Center for PG Studies, Kalaburagi, India. Research areas are Machine Learning, python IDE(pycharm).

How to cite this paper: Swaroopa Shastri, Shashank Vishwakarma, "An Efficient Approach for Text-to-Speech Conversion Using Machine Learning and Image Processing Technique", International Journal of Engineering and Manufacturing (IJEM), Vol.13, No.4, pp. 44-49, 2023. DOI:10.5815/ijem.2023.04.05