

OCR-based Text Extraction from Images

Vijayalakshmi V.¹, Amruta Ashok Naik², Chinmayee M.S³

Assistant Professor, Information Science and Engineering, Atria Institute of Technology, Bangalore, India¹

Student, Information Science, and Engineering, Atria Institute of Technology, Bangalore, India²

Student, Information Science and Engineering, Atria Institute of Technology, Bangalore, India³

Abstract: “OCR Based Text Extraction From Images” is based on text recognition from image and text-to-speech conversion. It converts the text within an image into speech format and reads it out. Image has text characters which is the main source of information for content-based indexing. The goal of text recognition is to recognize the text from printed hardcopy documents to the desired format. However, these text characters are difficult to be detected and recognized due to their varying sizes and complex backgrounds. In the segmentation step, we model the distribution of grayscale values of pixels. Finally, they are processed by OCR. OCR is the technology that is the answer for extracting data from the images and any documents and convert into computer-readable forms which can be helpful for editing or searching. Images are converted to text files that will be further converted to audio files.

Keywords: OCR, TTS, Text detection, preprocessing, Gaussian blur, spell correction

I. INTRODUCTION

OCR, or Optical Character Recognition, is a process of recognizing text inside images and converting it into an electronic form. These images could be of handwritten text, printed text like documents, receipts, name cards, etc., or even a natural scene photograph.

OCR has two parts to it. The first part is text detection where the textual part within the image is determined. When the textual part is determined then it will be important for the next part where recognition to be done. Using these techniques together is how you can extract text from any image.

As the full form of OCR suggests optical character recognition, OCR technology faces the problem of detecting the all different types of characters present. Characters that are printed can be converted to digital text which can be readable by a machine. Think of any kind of serial number or code consisting of numbers and letters that you need digitized. By using OCR we can convert these images into a digital text file. The technology makes use of many different methods. In simple words image will be given as input and then it will be processed and text will be extracted by it, it should be recognized by the machine.

OCR cannot consider the properties of the object which is given as output. It just sees or scans the image we aim to convert into text characters. For example, when we scan a image machines will learn and recognize the characters, but not the sense of the word.

II. LITERATURE SURVEY

As discussed earlier text recognition from images is still active research in the field of pattern recognition. Many technologies were proposed by many researchers to rectify the problem in text recognition. In the forthcoming section, we present a detailed survey of approaches proposed to handle the issues related to text recognition.

In the paper, they provide a highly scalable architecture containing tabular data with and without borders into cells and reconstruct the tabular data while preserving the tabular format. Here they used the otsu algorithm (which is used to perform automatic image thresholding which determines the optimum threshold of the image)[1]. The main motive of this paper is to convert text to speech. This is also based on the region of interest using canny edge detection. Canny edge detection is a technique to extract useful structural information from different vision objects and reduce the amount of data to be processed[2].

The main motivation is to develop a highly efficient image-to-text converter as well as handwritten text as well. Here they used the python libraries such as Tesseract, OpenCV, and Tensorflow to convert images to text and they used Django, HTML, CSS, and JavaScript for web applications[3]. The main purpose of this paper is to convert images to text and to speech. After extracting text from images translate that into particular languages. This project will help visually impaired persons, and this project also acts as an OCR Translator so the language barrier won't be a problem[4].

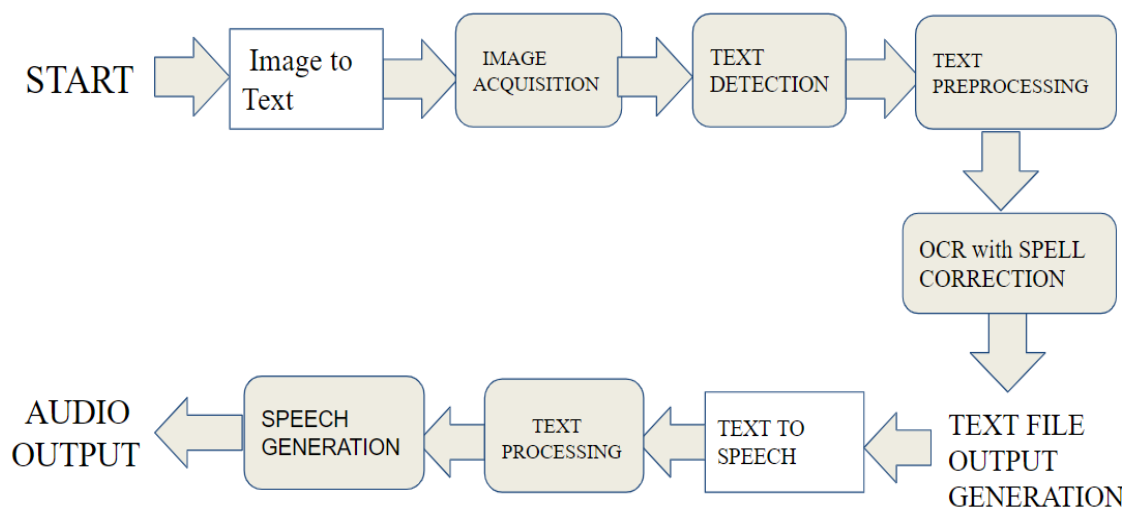
The main objective of this paper is to convert handwritten documents and provide efficiency because typed and printed are uncomplicated because of their well-defined shape and size but the handwriting of one person differs from another[5]. In this paper, they performed digitization of the textual information with the help of a scanner or camera. This digitized

information is processed with the help of LattePanda Alpha on board which is a tiny device that supports Linux and Windows OS[6].

The paper proposes to yield a program that is based on the web to convert any text or paragraphs into speech in the mp3 format. The people using this technology can listen to this audio file anywhere and any time they wish[7].The main objective of this paper is to convert images to speech using the LabVIEW framework.LabVIEW is a graphical programming dialect that utilizes symbols rather than lines of content to make applications[8].

In this paper, to extract text from images they have used the otsu’s algorithm for segmentation and the Hough transform method for skew detection[9].The objective of this paper is to recognize text from images for a better understanding of the reader by using a particular sequence of different processing modules. They have used a Document Image Analysis (DIA) to extract characters from images[10].

III. PROPOSED METHODOLOGY



A. Image to Text

1. Image Acquisition

The first step is to acquire images of paper documents with the help of optical scanners. So that an original image can be captured. The OCR system requires the input to be an image that contains the text that needs to be extracted. This image has to be acquired through some means. Some of the means of image acquisition are Scanned documents, Photographs, Digital images and art, Screenshots and etc.

2. Text Detection

Extract the ROI(Region of Interest) in the image instead of a complete image. This will discover all the focused text regions in the original image and the original image may contain multiple ROI. Here we apply the gaussian blur algorithm. This is an image unfocus technique that uses a Gaussian function (it also expresses the normal distribution in statistics) for calculating the modified pixel value for each pixel in the image. The formula for Gaussian function in one dimension is

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

In two dimensions, the formula is the same but used as a product of two Gaussian functions for each dimension as shown below.

$$G(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

where x is the distance from the origin on a horizontal axis, y is the distance from the origin on a vertical axis, and σ is the standard deviation(SD) of the Gaussian distribution. Gaussian Blur algorithm is used for noise reduction.

3. Text Preprocessing

The preprocessing is a fundamental stage proceeding to the feature extraction stage; it regulates the appropriateness of the outcomes for the successive stages. The OCR success rate is unpredictable in each stage.

OCR is a software that frequently pre-processes the images to improve the chances of successful recognition. The aim of text preprocessing is to improve actual image data. By this, unwanted distortions are suppressed and specific image features are enhanced. The preprocessing of images aims at selectively removing the redundancy present in captured images without affecting the details. The data which we collect or generate is mainly raw data, i.e. we can't use

this in applications directly. Therefore, we need to scan it first, carry out the required pre-processing, and then use it. It involves many methods such as resizing, noise removal, RGB to grayscale, Thresholding, etc.

4. OCR

The OCR is used to group the patterns and it contains several steps such as segmentation, feature extraction, recognition, etc.

Using an adaptive classifier, OCR can have many benefits. The adaptive classifier algorithm is based on a structural analysis method that analyses the character structure like curves and lines in character. The adaptive classifier uses isotropic baseline/x-height normalization. The baseline/x-height normalization makes it effortless to differentiate upper and lower case characters and improve immunity to noise specks. When compared to the static classifier, it will be easy to recognize the uppercase and lowercase characters in the adaptive classifier. In an adaptive classifier, it can identify and recognize any type of font style in the text. While in a static classifier, it cannot recognize any type of font in the text. In a static classifier, it can recognize subscripts and superscripts easier, but it needs an additional extra feature to identify the lowercase and uppercase alphabets. When compared to the static classifiers, adaptive classifiers are better suited to recognize the text in the image.

5. Post Processing(spell correction)

A spell checker basically checks the context of surrounding words, to only then present the possible error and its suggestions for correction. A spelling corrector is one step further and corrects the misspelled word automatically by choosing the most likely word from the dataset. SymSpell is based on the approach which takes a single word of a sentence and collects appropriate words from the dictionary at a time. Error detection and error correction are done to correct the spelling of the text extracted.

B. Text to Speech

1. TTS



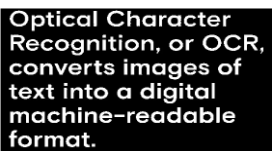
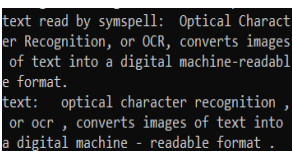
The main aim of a text-to-speech system is to convert an arbitrarily given text into a spoken output. The main modules of the text-to-speech system are

- Text processing: The Text processing technique will separate a text into words, subwords, and characters.
- Speech generation: The output obtained from the text normalization step is then converted into speech.

pyttsx3 is used to convert the text to speech. This module is a text-to-speech transformation library in Python. It works offline so it is amicable with both Python 2 and 3. An application invokes the pyttsx3.init() factory function to get an instance of a pyttsx3. It is a very convenient tool that converts the entered text into speech. The pyttsx3 module supports both male and female voices using "sapi5" for windows.

IV. EXPERIMENTS AND RESULTS

It was observed that the original image is given as input to the project and after that only the text region area was detected in the region of interest(ROI) step and the image is preprocessed using the adaptive thresholding technique which converts the image into a binary image. Now the binarized image is given as input to the OCR(optical character recognition) to detect the text from the binarized image. The text is detected from the image. The text is given as input to the sym-spell algorithm; it corrects the misspelled words and the corrected text file is given as input to the TTS(text-to-speech) module which turns the text file into audio.

 <p>Original Image</p>	 <p>Text Detection</p>
 <p>Text Preprocessing</p>	 <p>Text output</p>

V. CONCLUSION

This is a union of two composite systems, binding together for use better. Designing and executing a successful text extraction and the text-to-speech system is not just a single procedure but several concurrent ones. Together, they understand the basic strategies, systems, and procedures which are essential for an effective program. This system can be boosted by improving the characteristic of speech generation and the extraction of the text without any problem with the output generated. This system can be expanded to the video text extraction process in which the text appearing in the frames of video can be detected, extracted, and further converted to speech form.

REFERENCES

- [1]. Ashish Ranjan, Varun Behera, "OCR Using Computer Vision and Machine Learning" ResearchGate, 2021.
- [2]. Sadhana Suresh Chettiar, Bhuvana P, Harshitha P, Bhavana N M," Talkie Text: The Image Reader" IJERT,2021.
- [3]. Saurabh Dome, Asha P Sath," Optical Character Recognition using Tesseract and Classification", IEEE 2021.
- [4]. Nivetha, Kameshwari," Image to Text and Speech Converter" IRJET, 2020.
- [5]. K.Karthick, K.B.Ravindrakumar, R.Francis, S.IIankannan," Steps Involved in Text Recognition and Recent Research in OCR; A Study" IJRTE,2019.
- [6]. Dr. Samuel Manoharan, "Text Recognition, Information Extraction & Vocalization" JIIP 2019.
- [7]. Sanket Munot, Akshay Patil, Utkarsha Khandale, prof. Supriya S Ambarkar," Image To Speech Conversion Website ", IJSART 2019.
- [8]. J.J.Mullani, M Sanskar, Priyanka S Khade, Snehal H Sonalkar, Nikita L Patil," OCR Based Speech Synthesis System Labview", IEEE 2018.
- [9]. Neha Agarwal, Arashdeep Kaur," An Algorithmic Approach for Text Recognition from Printed Text Images ", IEEE 2018.
- [10]. Pratik Madhukar Manwatkar; Shashank H. Yadav,": Text Recognition from an Images", IEEE 2015.