Text to Speech Conversion Using Google Vision Api

Karun Somasunder M[1], Amal J.S[2], Gopal Gopakumar[3], Suraj V Thomas[4], Keerthi Krishnan[5]

[1]*Systems Engineer, Infosys, Department of Computer Science & Engineering, Anna University, Chennai 600 025, India*

[2,3]*B.Tech Student, Department of Computer Science & Engineering, Anna University, Chennai 600 025, India*

[4] *Information Security Analyst, Department of Computer Science & Engineering, Anna University, Chennai 600 025, India*

[5] *Research Scholar, Department of Computer Science & Engineering, Anna University, Chennai 600 025, India*

***Abstract* - With recent advancement within the technology, we shall implement an assistive device that's capable of capturing a picture from a camera and extracting the text from the captured image and further to convert the text to speech as voice-based output to assist the people. The captured image is analyzed using Google Cloud Vision API Optical Character recognition (OCR). So as to extract text, we use image preprocessing methods to obviate any noise or blur within the captured image so that the accuracy is often increased. Further, we include software-based text to speech to convert the text to speech as voice output. The Google Cloud Speech API integrates with Google Cloud Storage for data storage.**

***Index Terms* – OCR, Google Cloud Vision API, Text to Speech Conversion, gTTS, Flask.**

I.INTRODUCTION

OCR is an electronically based conversion of various documents such as images into machine encoded format like Unicode UTF-8, UTF-16 text formats. OCR provides alphanumeric recognition text extraction from the images or documents. Now it has been one of the major researches oriented with various applications including healthcare, vehicle plate recognition etc. With advancement in the recent technology of computer vision Artificial intelligence and pattern recognition provides more accurate results in extracting text features from the image.

In recent years, OCR has been widely used to help the visually impaired person to assist them in reading the documents. Especially as a book reader device to assist the visually impaired people to read the text by automatically extracting text and converting the text to speech format. In general, Optical Character Recognition (OCR) comprises translation using various algorithms such as Pattern recognition to extract the text from any documents. Nowadays, with the help of computers most of the OCR are electronically extracting the text from the documents that are in various image and readable formats such as JPG, PDF and HTML files. Therefore, OCR bridges the gap between the main gap between the person and therefore the machine interface. Even though there are various Pattern recognition algorithms used to extract the text from the documents, sometimes the text is not extracted exactly as expected due to low quality of the documents and degraded quality of the documents especially with the old documents. OCR mainly focuses on character recognition with various document types. If the printed or handwritten documents are of degraded quality, then appropriated image enhancement needs to be done to extract the exact text, such extraction is accomplished using computer vision OpenCV and Python Image Processing Libraries.

Different OCR are available in recent years such as Server OCR example Google OCR, Desktop OCR example Tesseract, Python OCR and Web OCR such as Google Cloud OCR, Amazon OCR and Microsoft OCR and the expected text extraction may vary from OCR type to another due the effective Pattern Recognition algorithms used in such OCR. In this paper, we would be focusing on preprocessing and Google Cloud OCR so as to extract text from a dynamic environment like low light.

This paper is organized in the following order; related works are discussed in Section II as well as on Optical Character Recognition. Section III describes the proposed system architecture and Section IV describes the methodology. Further the results are discussed in Section V followed by conclusion and future work.

II.RELATED WORKS

Google Cloud Vision API was used for image analysis [1]. Their project detects individual objects and faces within images, also finds and reads printed words contained within images. Paper evaluates the robustness of Google Cloud Vision API to input noise. Especially, a set of images are taken and noise is added to them then the API is unable to detect the right text or object as if the noise is removed then the output is analogous thereto of the primary image. Cloud vision API can enjoy noise filtering

A prototype was proposed that helps people to hear the text content of the image in their native language [2]. The text is extracted from the image and then text is converted to translate speech of the user's native language. Camera captures the image and then the OCR engine converts the image to text. Then text is converted into speech using the gTTS [2]. This file is then converted into the desired language by using a python script.

A system was proposed that reads text on a captured Image [4]. It is performed as text is extracted from scanned images using Tesseract Optical Character Recognition (OCR) and then converting the text to speech. First captured image is converted to grayscale and then filtered using Gaussian filter to reduce noise adaptive Gaussian thresholding is used then it is converted to binary image and cropped and loaded to tesseract OCR for text recognition and output of tesseract is text file which is input for e-speak, which produce audio [7].

The current system in existence is that words of any language can be typed manually and translated to any language as required. It is not possible to convert the image of an entire text book from one language to another. Some mobile applications which tried to do the above showed major errors in converting. The existing system [3], where the traditional OCR is used, fails to detect text from the images which are blur or are having low resolution, low contrast, high noise, and distortions. The noise in the images distort the final result. Thus, creating an understanding problem for the users. Some systems where words are pronounced wrongly. Most of the accurate text to speech Applications are coming in paid format on market.

A. Optical Character Recognition

Optical character recognition (OCR) is a system that converts input text into machine-encoded format. Today, OCR is helping not only in digitizing the handwritten medieval manuscripts, but also helps in converting the typewritten documents into digital form. OCR involves the following stages that may convert a document into ASCII code and Unicode mapped to each ASCII code. OCR engine includes the following preprocessing methods which helps to enhance the image and to extract the text. Appropriate methods may be used depending on the factors such as paper quality, resolution of the image, age of the document, and the layout of text etc.

- Binarization,
- Noise removal,
- Thinning,
- Skew Detection & Correction,
- Line, Word and Character Segmentation,
- Feature Extraction and Selection,
- Classification.

Binarization, Thinning, Skew detection & correction are applied to the scanned image to extract the line and each character or word separation. The preprocessing in document analysis helps to remove the noise due to various factors such as low-quality image or lighting background illumination.

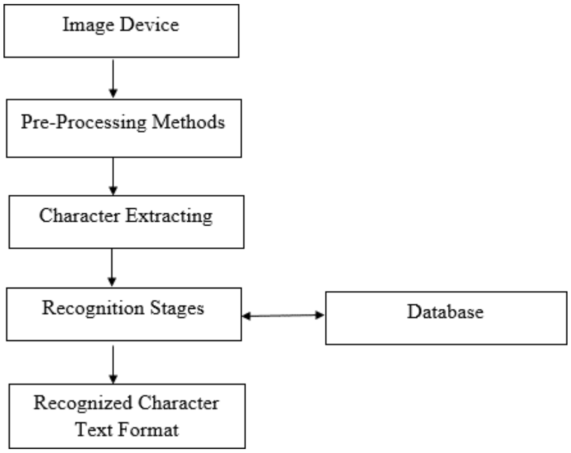


Fig. 1. General OCR system

The accuracy of OCR is evaluated based on the recognition rate and rejection rate and error rate is determined based on the ratio between the Classified and Misclassified data rate [8].

III.PROPOSED SYSTEM

In this section we describe our TEXT TO SPEECH application. The application uses inbuilt smartphone camera to capture the images. The captured image or an existing image is then sent for Text extraction and Conversion processing as a http request to the server. The Extraction and Conversion process are written in a python code which is stored inside a server. Further using Google Cloud Vision, we intend to extract text from any type of documents such as printed or handwritten document. In order to accomplish this, we intend to use various software tools such as Google Cloud Vision API, Google Text to Speech (gTTS), Python Image Libraries, Flutter (App Development). Further, Image enhancement preprocessing technique are applied to detect the edges of each character from the captured image. After extracting the edges, we apply Google Cloud OCR engine to extract text from the image using Google Cloud Vision API. Flutter and Python modules are joined using Flask framework.

The major component of the proposed system is Google Cloud Vision API. It was released in the year 2015. It enables developers to analyze the content of images. It uses powerful machine learning tools to extract the required data from images. It can perform different functions like label detection, face detection, Logo detection, Optical Character Recognition (OCR). Finally, the UTF-8 Unicode text is fed to a Text to Speech engine such as gTTS which converts it into speech. The converted speech is then sent back to the application from the server where the user can access it. For Future scope, we can use Goslate to translate from one language to another, thereby supporting Multi Language translation. The mobile application is connected to the server using python Flask framework Flask is a micro web framework written in Python. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself.

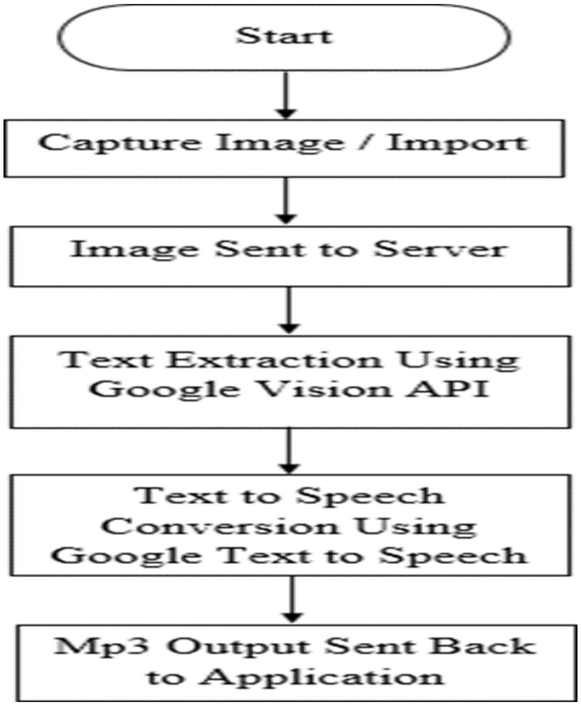

Start
Capture Image / Import
Image Sent to Server
Text Extraction Using Google Vision API
Text to Speech Conversion Using Google Text to Speech
Mp3 Output Sent Back to Application

Fig. 2. Flowchart

IV.METHODOLOGY

The current system in use is that words of any language are often typed manually and translated to any language as needed. With the use of existing systems, it is impossible to convert the image containing text in a different language to speech in a common standard like English. Most of these mobile applications are paid which can’t be accessed by all. Some mobile applications in which we tried to do the above showed major errors while converting. The android application developed is user friendly. The application uses inbuilt smartphone camera to capture the images. The captured image or an existing image is then sent to the server for Text extraction and conversion processing as a http request. Further using Google Cloud Vision API, we extract text from the image. During image extraction processing the image is resized to 720 x 480 pixels since greater the dimension, greater is the time taken for OCR, after resizing the image is converted into grayscale. All. the Google Cloud then performs OCR (Optical Character Recognition) on the processed (resized and grayscale) image sent to it. The Google Cloud can recognize different languages. The text is obtained and then converted into speech using the gTTS engine. The converted speech is then sent back to the application from the server where the user can access it. The output is in the form of a mp3 file and can be played using an mp3 player.

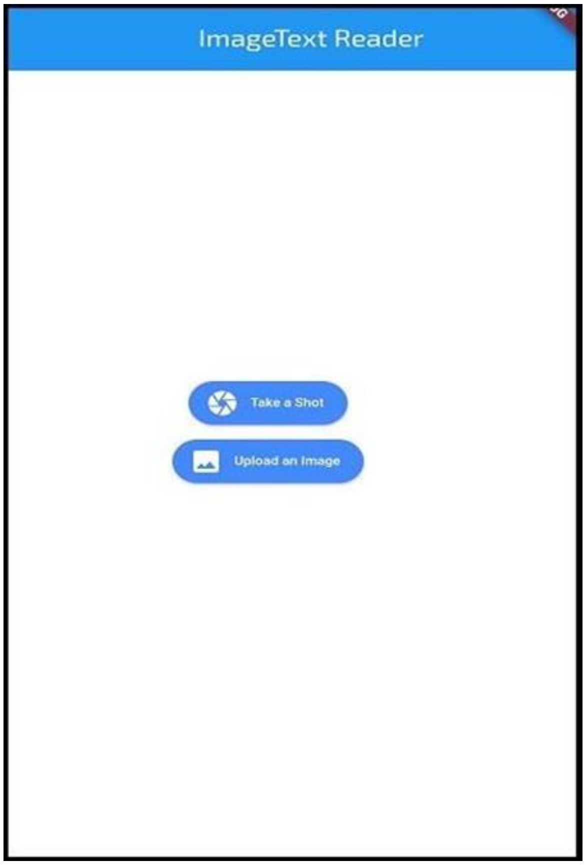


Fig. 3. Application UI

V.RESULTS AND DISCUSSION

Text is extracted from the image and converted to audio Figure 6. It recognizes different fonts. Skewed text images are also identified and converted into speech. The model recognizes the text which is readable by human eyes. In order to study the accuracy of the application, we tested it with an image inputs Figure 4 and the corresponding output is shown in the Figure 5.

Existing Text to Speech system fails to detect handwritten texts, texts with cursive fonts and produces inaccurate results for images containing noise or blurriness. Our system Text to Speech convertor using Google vision API is capable of producing more accurate and better results for the same. We observed that using our system we could get the desired output even if the image was written with a different font or had noise or distortions.

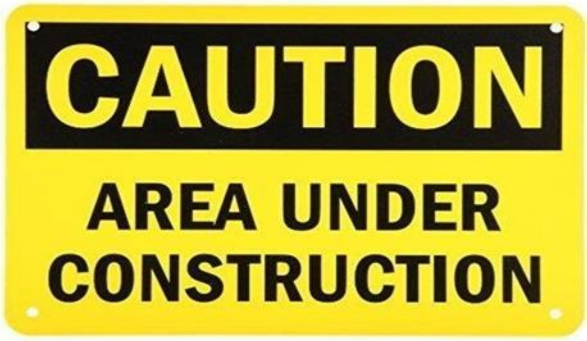
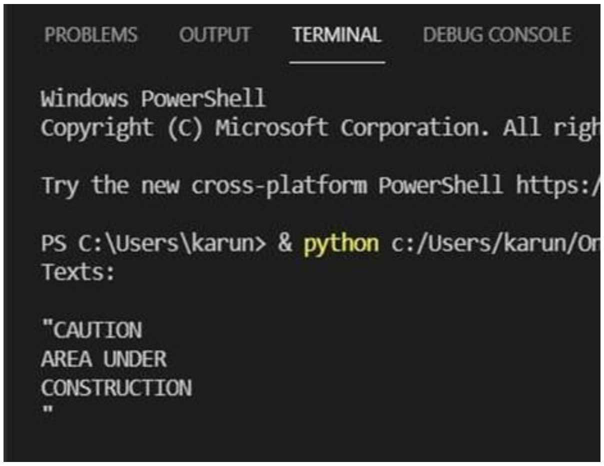


Fig. 4. Test Input 1



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All righ

Try the new cross-platform PowerShell https:/

PS C:\Users\karun> & python c:/Users/karun/Or
Texts:

"CAUTION
AREA UNDER
CONSTRUCTION
"
```

Fig. 5. Output

Fig. 6. Mp3 Speech Output

VI.CONCLUSION AND FUTURE WORK

Text to speech synthesis is a rapidly growing aspect of technology and is increasingly playing a more important role in the way we interact with the system and interfaces across a range of platforms. We identified the different operations and processes involved in text to speech synthesis and developed a very simple and attractive graphical user interface which allows the user to feed an image as an input and to get the corresponding speech as output. Our system interfaces with a text to speech engine developed for English (US).

For future work, a module which helps the users to convert the speech to their regional language and get the corresponding audio file as output could be added. Conversion of text image with multi-lingual script can be implemented and also cursive characters could be identified and converted to speech.

REFERENCE

[1] Hossein Hosseini, Baicen Xiao and Radha Poovendran “Google’s Cloud Vision API Is Not Robust to Noise” 16th IEEE International Conference on Machine Learning and Applications December 18-21, 2017.

[2] Rithika.H, B. Nithya santhoshi “Image Text to Speech Conversion in The Desired Language by Translating with Raspberry Pi” International Conference on Computational Intelligence and Computing Research 2016.

[3] Mr. Rajesh M., Ms. Bindhu K. Rajan Ajay Roy, Almaria Thomas K, Ancy Thomas, Bincy Tharakan T, Dinesh C “Text recognition and face detection aid for visually impaired person using

[4] raspberry pi” International Conference on circuits Power and Computing Technologies [ICCPCT] July 2017.

[5] Text Reader for Visually Impaired Using Google Cloud Vision API - IEEE Journal.

[6] Davide Mulfari, Antonio Celesti, Maria Fazio, Massimo Villari and Antonio Puliafito “Using Google Cloud Vision in Assistive Technology Scenarios” IEEE Workshop on ICT solutions for eHealth 2016.

[7] Yasuhisa Fujii, "Optical Character Recognition Research at Google", IEEE 7th Global Conference on Consumer Electronics (GCCE), December 2018.

[8] Text localization and extraction in images using mathematical morphology and OCR Techniques; 2013.

[9] Mithe R, Indalkar S, Divekar N. Optical character recognition. International Journal of Recent Technology and Engineering. 2013 Mar; 2(1).

[10] Archana A, Shinde D. Text pre-processing and text segmentation for OCR. International Journal of Computer Science Engineering and Technology. 2012:810–12.

[11] Ani R, Effy Maria, J Jameema Joyce, Sakkaravarthy V, Dr.M.A. Raja, “Smart Specs: Voice Assisted Text Reading system for Visually Impaired Persons Using TTS Method”, IEEE International Conference on Innovations in Green Energy and Healthcare Technologies (ICIGEHT’ 17), 2017.