TEXT RECOGNITION IN IMAGES AND CONVERTING RECOGNIZED TEXT TO SPEECH

Ajay Pal
Dept. of computer science
and Information technology Engineering
RDEC, Ghaziabad
apal5774@gmail.com

Vasu Chaudhary
Dept. of computer science
and Information technology Engineering
RDEC, Ghaziabad
chaudharyvasu234@gmail.com

Rahul Singh
Dept. of computer science
and Information technology Engineering
RDEC, Ghaziabad
rahulsingh81two@gmail.com

Anmol Tyagi
Dept. of computer science
and Information technology Engineering
RDEC, Ghaziabad
anmoltyagi849@gmail.com

Abstract

Around 285 million people worldwide are visually impaired, including close to 39 million blind people. This has a significant impact on the lives of persons who are blind or visually impaired. Even though numerous attempts have been made to assist those who are blind in seeing objects through alternate senses like touch and sound, text-reading technology is still in its infancy. The system in use right now is either constrained in its application or expensive to maintain. Therefore, we require a system that can automatically recognize and read aloud text to a user base of visually impaired people that is both affordable and truly efficient. The main goal of this research is to develop a program that can identify text characters from turn any natural image into a voice signal. The programme need to carry out the identical action for any uploaded image and PDF file. The application should also have tools for pace modulation, voice choosing options, and storage capability for image to text output. The target audience for this programme can be expanded to include people with special needs who also have learning impairments, young children, and several other societal groups. The text is extracted from the image using optical character recognition (OCR), and the Windows API is utilised to turn the text into speech. The programming language for digital image processing is MATLAB.

Key Words: Digital image processing, optical character recognition, speech modulation, MSER Regions, stroke width algorithm, and image character recognition are some of the terms used in this document.

Cite this Article: Ajay Pal, Vasu Chaudhary, Rahul Singh, and Anmol Tyagi Conversion of Images to Speech Using Digital Image Processing.

1.Introduction

A popular area of computer technology is image-to-speech conversion. It establishes a crucial factor in how we engage with the system and interfaces on many platforms. It has long been a goal to replicate human abilities like reading through machines.

Machine reading, however, has developed from a pipe dream to a reality during the past 50 years. The most effective form of human communication is most likely speech. One of the most popular uses of technology in the fields of pattern recognition and artificial intelligence is optical character recognition.

The tool assists in converting textual information that is embedded in an image or scene into speech. This is not the only use it may be put to. It is beneficial to take text from PDF files and turn it into speech. All of the collected text can be stored as a text file in any location on the computer. While the text is being read aloud, it also offers the option to look up synonyms for words. Different paces may be comfortable for users to comprehend the language. As a result, a clause is added that allows for speech tempo modulation. Additionally, users can select from a variety of male and female speakers' voices as well as accents.

OCR, or optical character recognition, is a technique we use to extract text from photographs. After that, a text-to-speech (TTS) module turns the text into audio. We can see that this procedure was split into two modules. The first is picture recognition, and the second is speech conversion for that image:

optical character recognition Optic character recognition is referred to as OCR. Through this procedure, the application will be able to recognise a character automatically using an optical method. OCR is the conversion of captured photographs of printed or typewritten text into digitally changeable information.

Speech synthesis: Without directly using a human voice, speech synthesis creates speech that is more human-like than robotic. A voice synthesiser is, more broadly speaking, a type of technology that creates fake speech through the creation of symbols and signals. It has been possible to modify the speech's cadence. Additionally, the application includes a variety of voices and accents.

The above can be depicted by the following illustration:

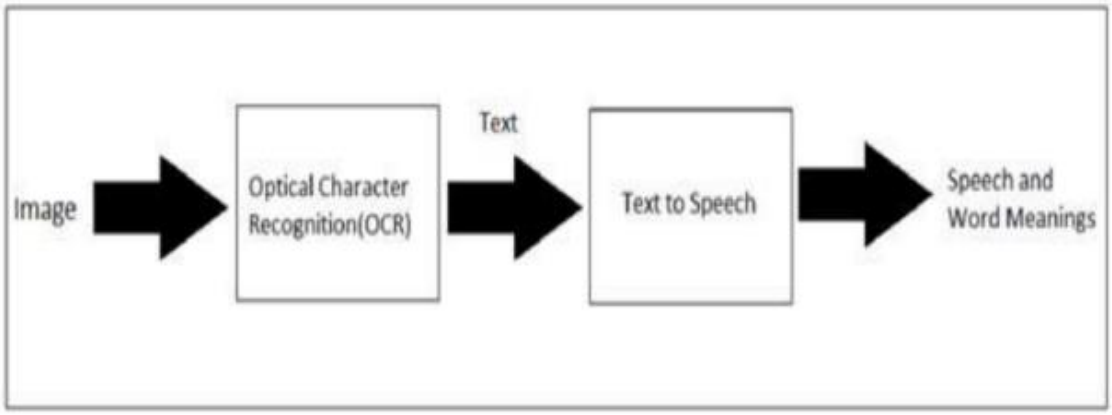

Figure 1 An overview of the system

2.FrameWork And Architecture

2.1. High Level Design

The high-level operation of the system is depicted in the diagram below.

The graphic demonstrates that the system is made up of two primary modules:

The system's two primary components are depicted in the figure as :

(a).image-to-text conversion and PDF generation.

(b).Conversion from text to speech

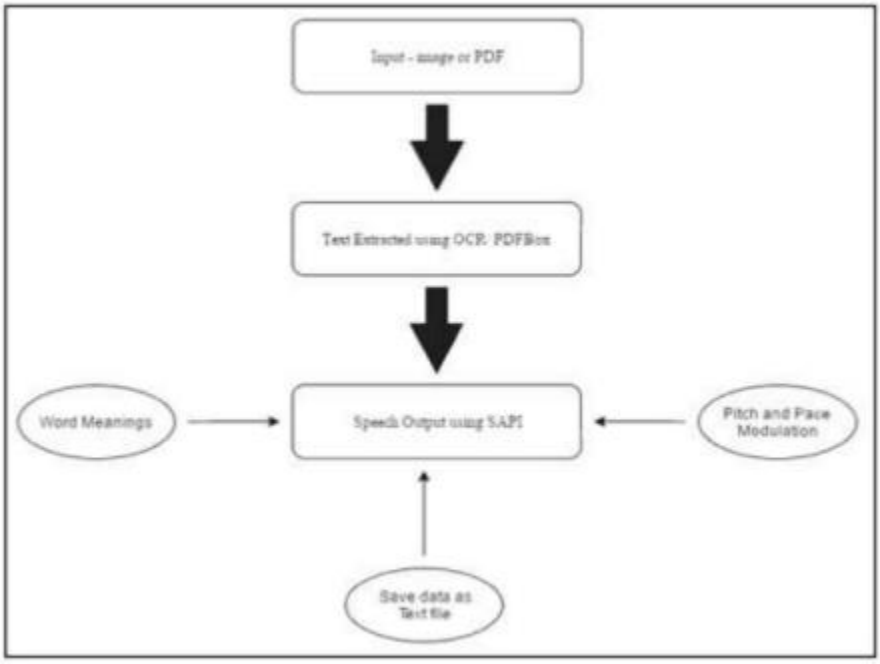

Figure 2 High Level Diagram of the System

The conversion of a picture into speech is handled by these two modules. The text to speech module also includes a number of other features, such as word definitions and synonyms, pitch and tempo adjustments for speech output, and the ability to store extracted data as text files or in the.txt format.

2.2. Image/PDF to Text Conversion

The diagram shows the various functionalities of this module.

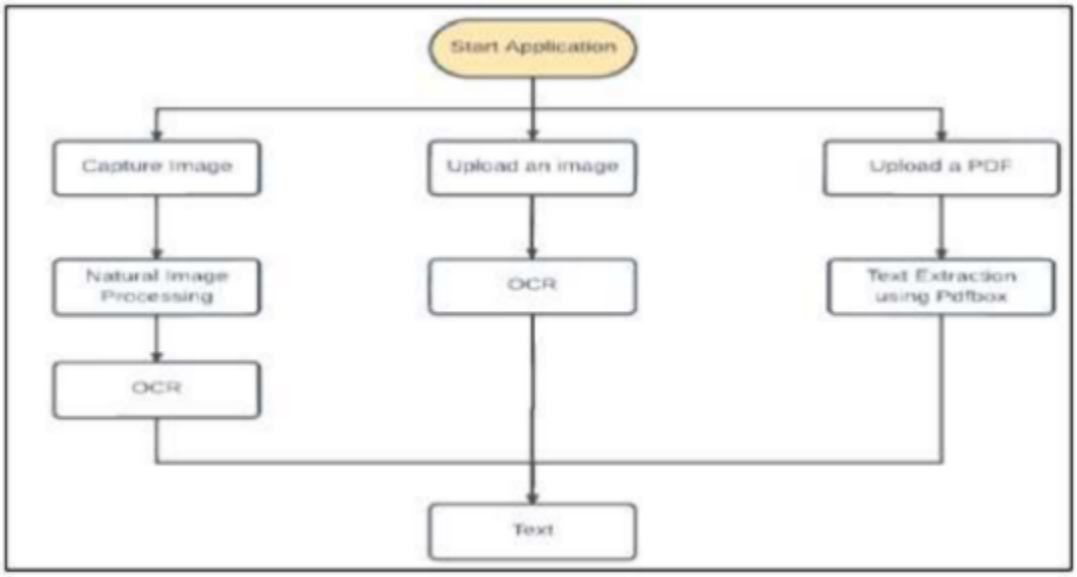

Figure 3 Detailed Diagram of Image/PDF to Text Conversion

The Image to Text Conversion module is further subdivided into three sub-modules that are in charge of Image to Text conversion.

- Capture an image
- Upload an image
- Upload a PDF file

The technique of extracting text from various sources differs. As seen in the illustration, word extraction from a taken image necessitates natural image processing. It begins by finding the text-containing portions of the picture. MSER (Maximally Stable Extremal Regions) and Stroke-Width algorithms are

utilized to perform this as well as locate characters. The recognized letters are then combined into words and phrases using Optical Character Recognition (OCR). Finally, a text file with the extracted text is created.
The graphic below depicts this process:

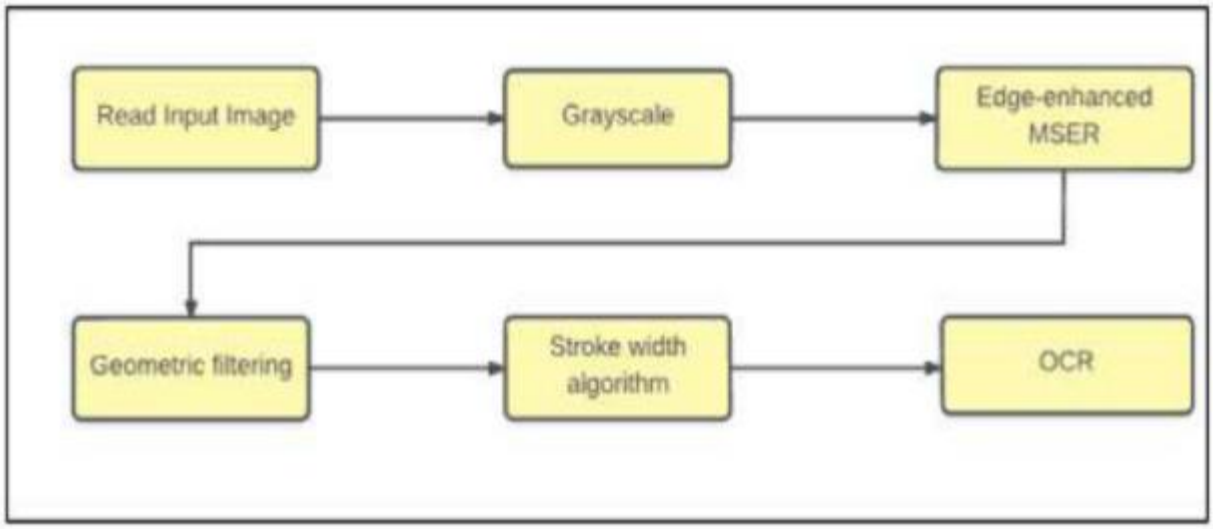

Figure 4 Natural Image Processing for Captured Image

To improve efficiency, all noise in the image is clipped before to processing. The picture is then OCR' d. The extracted text is subsequently written to a file in the working directory for further processing. The PDFBox.jar file is used to extract text from PDFs. It is integrated with the program to allow for smooth extraction from PDF files.
It is important to remember that the algorithms used to analyze natural images are designed in such a way that the quality of the image does not interfere with the precision with which the text is retrieved. The OCR works well enough for photos with non-straight text orientations and low resolution. In an image or a PDF, text can be in any font style and size.
The algorithms have been designed to correctly recognize text in any style or size. When it is converted to text format, however, these details are lost because the algorithms are only equipped to read the text and not its font size and font style. The entire module has been designed to minimize latency and maximize efficiency.

2.3. Text to Speech Conversion

Text retrieved from any type of input is saved in the working directory as a text file. Microsoft supplies a Speech API that is utilized to supply different voices, and an algorithm is created to translate this text into speech.

Image to Speech Conversion Using Digital Image Processing

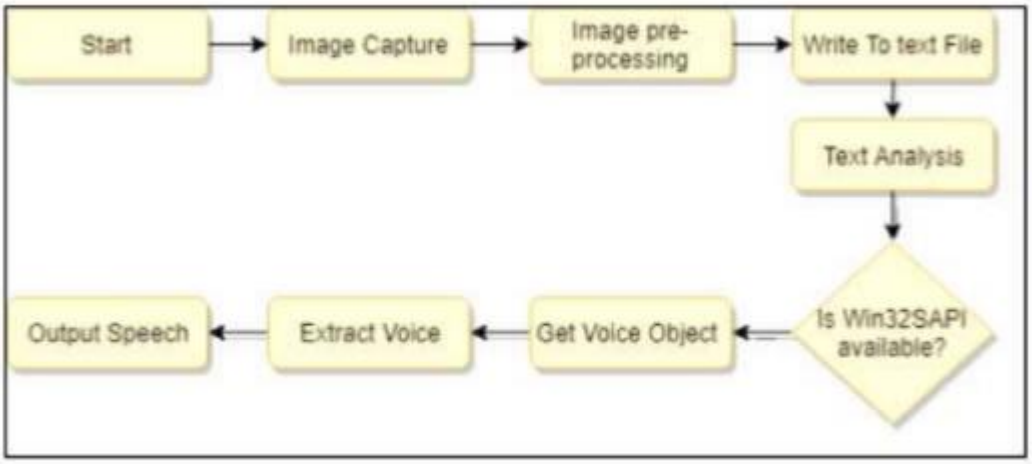

Figure 5 Text to Speech Conversion Flowchart

The first image is taken, or, in the case of a PDF, it is uploaded. Using the prior module, text from the input is extracted and transformed to text format. The text is then analyzed to divide into sentences. These sentences must appear in the correct order alongside the voice output. When a full stop ("."), a question mark ("?"), or an exclamation mark ("!") is encountered, the text is broken into sentences. If Windows 32 SAPI is not available, the default configuration for the speech module is utilized. This module also allows users to alter the tempo and timing of their speech. A real-time play and pause feature is also available.

This module includes it. This module also includes highly useful features such as word definitions, which can be used offline because it is supported by the Microsoft Word API. The extracted text can be stored in any format for the user's future reference. Finally, this module allows the user to store the document in a variety of formats and easily access it for future reference.

2.4. MSER Regions

For blob detection in photos, the program uses the Maximally Stable Extreme Regions algorithm. MSER changes with the picture threshold; given a threshold value, pixels below that threshold value are "white," while those over or equal to that value are "black." Because of the uniform color and great contrast of the text, the MSER feature detector performs well for locating text sections. The first step in constructing MSER is to execute a basic luminance threshold of the picture by sweeping the intensity threshold from black to white. After that, the extracted related components or Extreme Regions are executed. Following that, a threshold is determined when an external region is maximally stable. Finally, the descriptions of the areas as MSER features are acquired.

Although the MSER algorithm detects the majority of the text, it also detects several other stable regions in the image that are not text. The stroke width algorithm can assist in resolving this.

2.5. Stroke Width Algorithm

This algorithm has been implemented to retrieve the most likely stroke containing the required pixel. The algorithm receives an RGB image and returns an image of the same size, where the regions of suspected text are marked.

The first step is the stroke width transform which is an operator which determines the width of the most likely stroke containing the pixel for each and every pixel. The output produces by the SWT is an image of the same size as of the input image where each element contains the width of the stroke associated with that pixel. We have now obtained a map of the most likely stroke-widths for each pixel in the original image.

The next step is to group all these pixels into letter candidate which is done by selecting two neighboring having similar stroke width, and then applying several rules to distinguish the letter candidates.

For this reason, we have modified the classical Connected Component algorithm by altering the association rule from a binary mask to a predicate which compares the SWT values of the pixels. To increase the efficiency and reliability we strive forward to group the letters. Since single letters are not expected to appear in images, closely positioned letter candidates are gathered together into regions of text.

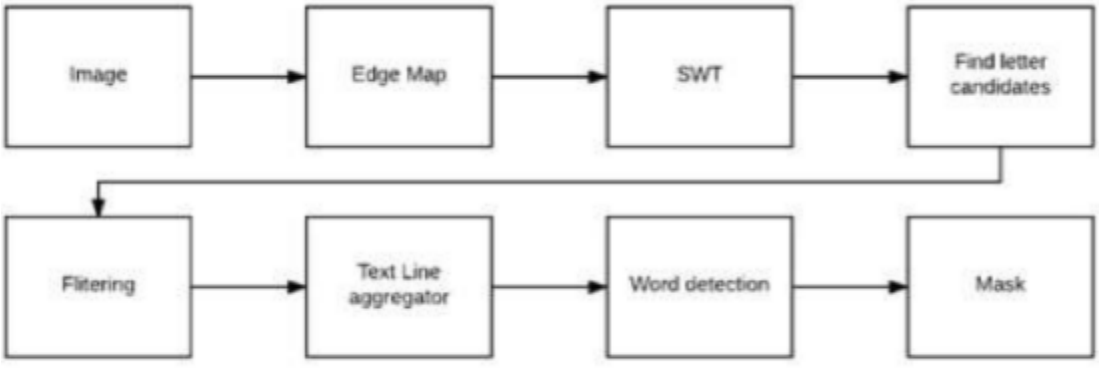

Figure 6 Stroke Width Algorithm Flowchart

2.6. Optical Character Recognition

The method of translating text from a digital picture into editable text is known as optical character recognition. It allows a machine to recognize the characters using optical methods.

The OCR result should preferably be the same as the input in formatting. The procedure begins with the picture file being pre-processed, followed by the acquisition of critical information about the textual text. MSER and the Stroke Width Algorithm aid in this process.

3.Algorithms use

(A).Tokenization

It's the process of breaking down the text into sentences and phrases. The work entails breaking down a text into smaller chunks (known as tokens) while discarding some characters, such as punctuation.

Consider the following example:

Text input: Potter walked to school yesterday.

Potter went to school yesterday, according to the text output.

The major disadvantage of this strategy is that it works better with some languages and worse with others. This is particularly true when it comes to tonal languages like Mandarin or Vietnamese.

Depending on the pronunciation, the Mandarin term ma can signify "a horse," "hemp," "a scold," or "a mother." The NLP algorithms are in grave danger.

(B). Bag of Words

This paradigm represents a text as a bag (multiset) of words, neglecting syntax and even word order while keeping multiplicity. In essence, the bag of words paradigm generates a matrix of incidence. These word frequencies or instances are then employed as features in the training of a classifier.

Unfortunately, there are some drawbacks to this paradigm. The worst is the lack of semantic meaning and context, as well as the fact that such terms are not appropriately weighted (for example, in this model, the word "universe" weighs less than the word "they").

(C). Text Summarization

As the name implies, NLP approaches can assist in the summarization of big volumes of text. Text summarization is commonly utilized in situations such as news headlines and research studies.

Text summarization can be done in two ways: extraction and abstraction. By deleting bits from the text, extraction methods create a rundown. Abstraction tactics produce summaries by constructing new text that conveys the essence of the original content.

Different NLP algorithms can be used for text summarization, such as Lex Rank, Text Rank, and Latent Semantic Analysis. To use Lex Rank as an example, this algorithm ranks sentences based on their similarity. Because more sentences are identical, and those sentences are identical to other sentences, a sentence is rated higher.

(D).Lemmatization and Stemming

Two of the strategies that assist us to develop a Natural Language Processing of the tasks are lemmatization and stemming. It works nicely with a variety of other morphological variations of a word.

These strategies allow you to limit a single word's variability to a single root. We can, for example, reduce "singer," "singing," "sang," and "sang" to a singular version of the word "sing." We can quickly reduce the data space required and construct more powerful and robust NLP algorithms by doing this to all the terms in a document or text.

Thus, lemmatization and stemming are pre-processing techniques, meaning that we can employ one of the two NLP algorithms based on our needs before moving forward with the NLP project to free up data space and prepare the database.

Both lemmatization and stemming are extremely diverse procedures that can be done in a variety of ways, but the end effect is the same for both: a reduced search area for the problem we're dealing with.

(E).Keyword Extraction

Keywords Extraction is one of the most important tasks in Natural Language Processing, and it is responsible for determining various methods for extracting a significant number of words and phrases from a collection of texts. All of this is done to summarise and assist in the relevant and well-organized organization, storage, search, and retrieval of content.

There are numerous keyword extraction algorithms available, each of which employs a unique set of fundamental and theoretical methods to this type of problem.

There are various types of NLP algorithms, some of which extract only words and others which extract both words and phrases. There are also NLP algorithms that extract keywords based on the complete content of the texts, as well as algorithms that extract keywords based on the entire content of the texts.

The following are some of the most prominent keyword extraction algorithms:

Text Rank: This algorithm operates on the same idea as Page Rank. Google uses this method to rank the importance of various websites on the internet.

Term Frequency – Inverse Document Frequency (TF-IDF): The full version of TF-IDF is Term Frequency – Inverse Document Frequency, which tries to better define the importance of a term in a document. Also, take into account the relationships between texts from the same corpus.

RAKE: RAKE stands for Rapid Automatic Keywords Extraction and is a type of NLP method. This can extract keywords and key phrases from a single document's content without taking into account other documents in the same collection.

(F).Knowledge Graphs

Knowledge graphs are a collection of three items: a subject, a predicate, and an entity that explain a method of storing information using triples.

The subject of approaches for extracting knowledge-getting ordered information from unstructured documents includes awareness graphs.

Knowledge graphs have recently become more popular, particularly when they are used by multiple firms (such as the Google Information Graph) for various goods and services.

Building a knowledge graph requires a variety of NLP techniques (perhaps every technique covered in this article), and employing more of these approaches will likely result in a more thorough and effective knowledge graph.

(G).Words Cloud

sometimes known as a tag cloud, is a data visualization approach. Words from a text are displayed in a table, with the most significant terms printed in larger letters and less important words depicted in smaller sizes or not visible at all.

Before applying other NLP algorithms to our dataset, we can utilize word clouds to describe our findings.

(H). Named Entity Recognition

Another significant technique for analyzing natural language space is named entity recognition. It's in charge of classifying and categorizing persons in unstructured text into a set of predetermined groups. This includes individuals, groups, dates, amounts of money, and so on.

There are two sub-steps to named entity recognition;

Named Entity Identification (the identification of prospective NER algorithm candidates) and

Named Entity Classification are two of these phases (assignment of candidates to one of the predefined categories)

(I).Sentiment Analysis

Sentiment analysis is the most often used NLP technique. Emotion analysis is especially useful in circumstances where consumers offer their ideas and suggestions, such as consumer polls, ratings, and debates on social media.

In emotion analysis, a three-point scale (positive/negative/neutral) is the simplest to create. In more complex cases, the output can be a statistical score that can be divided into as many categories as needed.

Both supervised and unsupervised algorithms can be used for sentiment analysis. The most frequent controlled model for interpreting sentiments is Naive Bayes.

A sentiment-labeled training corpus is required, from which a model can be trained and then utilized to define the sentiment. Naive Bayes isn't the only machine learning method that can be used; it can also employ random forest or gradient boosting.

4.RESULT AND DISCUSSION

MATLAB is used to implement the method. The method has been tested on a variety of scanned and printed document pictures. Various images/documents were considered, with minimal overlapping and fewer broken and connected characters. As predicted, the program delivered a less efficient text extraction with small details for handwritten notes with moderately blurred text.

The application is capable of identifying text form natural images as shown below:

Figure 7 Captured Image of a street Sign

Figure 8 Applications correctly identifies the text

The application is capable of identifying hand written text as shown below:

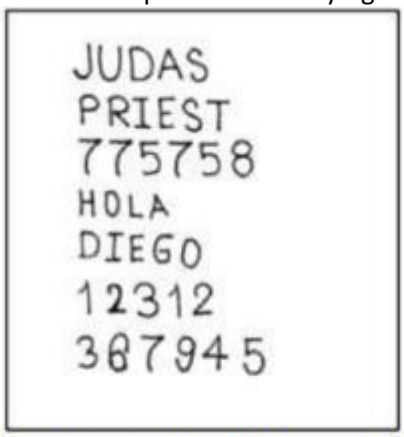

Figure 9 Captured Image of a handwritten text

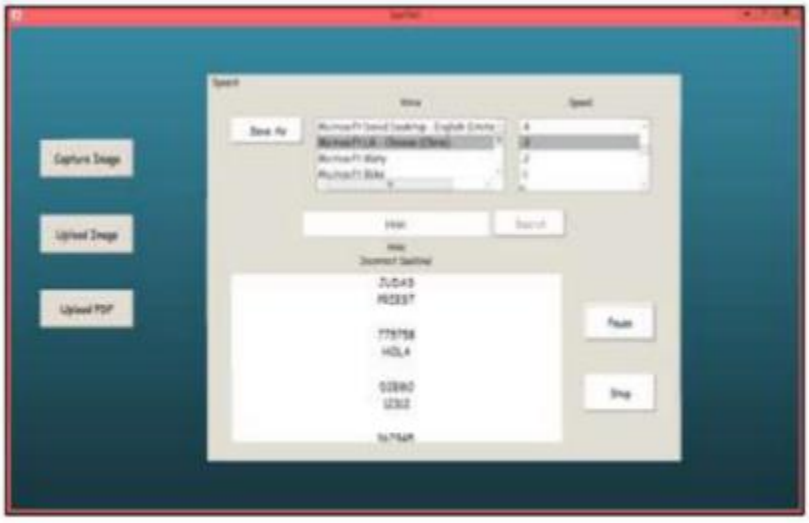

Figure 10 Application detects the correct text

The application is capable of identifying text form uploaded PDF as shown:

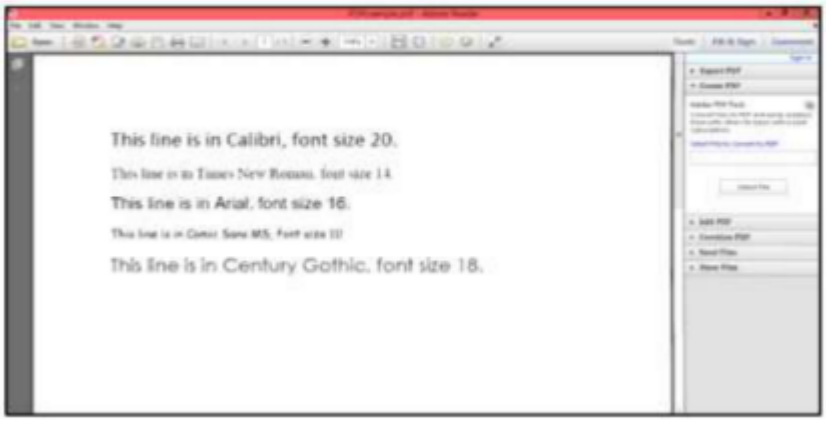

Figure 11 Uploaded a PDF file containing text in different fonts and font sizes

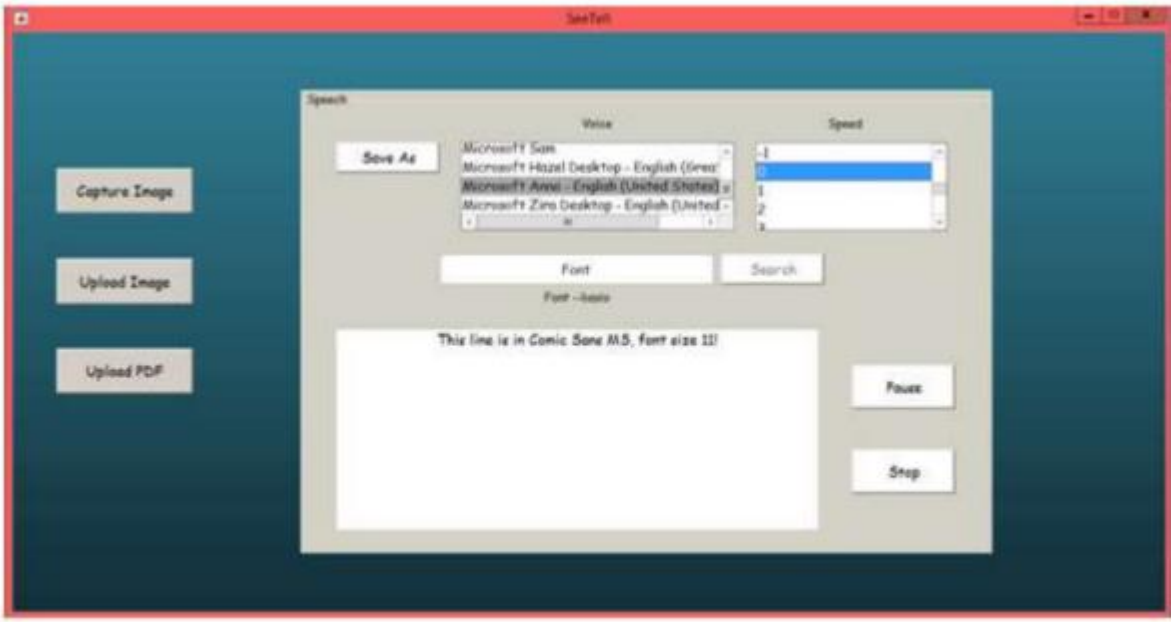

Figure 12 Application detects the correct text from PDF file

The intermediate text is presented sentence by phrase alongside the voice in this example.
Other features are available. The pitch is set to Microsoft Anna - United States, and the tempo is set to 0, which is the typical pace. The definition of the word "font" is looked up and shown. The system is tested using the methods outlined above. Testing is essential to offer information about the quality of the developed system. It determines if the system is fit for usage.

Table 1 Integration Test Cases and Results

Test case	**Description**	**Expected Result**	**Result**
1.	Capture Blur or Low-Resolution Image from Camera	Error: “Text cannot be extracted, please retry”	Pass
2.	Give Incorrect Path to Upload Image	Error: “Image does not exist, please retry”	Pass
3.	Give Incorrect Path to Upload PDF File	Error: “File does not exist, please retry”	Pass
4.	Search for a non-existent word meaning	Pause/Stop recitation Error: “Incorrect spelling, please recheck”	Pass
5.	Start application, Capture Image from Webcam, Recite Text	Display text sentence wise and recite the same Provide options of Pause/Play, Stop,Save Text File, Search Word Meanings, Change Speech Voice and	Pass

		Speed	
6.	Start application, Upload Image from PC, Recite Text	Display text sentence wise and recite the same Provide options of Pause/Play, Stop,Save Text File, Search Word Meanings, Change Speech Voice and Speed	Pass
7.	Start application, Upload PDF File from PC, Recite Text	Display text sentence wise and recite the same Provide options of Pause/Play, Stop, Save Text File, Search Word Meanings, Change Speech Voice and Speed	Pass

5.CONCLUSION

Text-to-speech conversion is a rapidly evolving feature of computer technology that has become an essential factor in deciding how humans engage with systems and interfaces across a wide range of platforms. This study will suggest a technique for image-to-speech conversion combining optical character recognition and voice synthesis. The designed programme is simple to use, relatively cost effective, portable, and applicable in real time. It can read text from any natural image, any document image, and PDF documents, as well as generating synthesised voice using a computer's speakers. Along with speed control, the developed software includes features such as word meaning assistance and voice modulation. This allows users to multitask and save time.

By listening to background materials while performing other chores. This approach can also be utilised in sections. For example, if we simply want to convert text, we may do so; moreover, only text to speech can be conducted individually. Expensive hardware, support software, a current operating system version, or even an internet connection are not necessary. A webcam to capture images is an optional requirement. This programme allows those with visual impairments or full blindness to read papers and books. This application is also suitable for those with learning difficulties. It is comfortable and simple to use for individuals who do not understand how computers function but still utilise the dictionary application.

Utilise this application. It is pleasant and simple to use for individuals who do not understand how computers function but still use this programme. The dictionary feature included intended to assist interpret the content. Users no longer need to look for synonyms and word meanings in several locations because the programme has them. Tests were carried out to verify the conversion, and positive findings were obtained.

5.1. Future Scope

___There is room to expand the current application's capabilities. Support for languages other than English may be included in one of the extensions. The algorithms used to pre-process a natural image function well in this system. However, there is still room for improvement. in addition to OCR methods. Support for additional image input formats can also be added.

Text recognition algorithms may be created to recognise text in low resolution and blurred photos. This will allow users to upload damaged historic manuscripts and scrolls and extract text from them.

REFERENCES

[1] Archana A. Shinde, D.G.Chougule “Text Pre-processing and Text Segmentation for OCR” IJCSET , January 2012.

[2] Benjamin Z. Yao, Xiong Yang, Liang Lin, MunWai Lee and Song-Chun Zhu, “I2T: Image Parsing to Text Description”.

[3] Bernard Gosselin Faculté Polytechnique de Mons, Laboratoire de Th´eorie des Circuits et Traitement du Signal, “From Picture to Speech: An Innovative Application for Embedded Environment”.

[4] Huizhong Chen1, Sam S. Tsai1, Georg Schroth, David M. Chen, Radek Grzeszczuk and Bernd Girod, “Robust text detection in natural images with edge-enhanced maximally stable extremal regions”, International Conference on Image Processing • September 2011

[5] Jisha Gopinath, Aravind S, Pooja Chandran, Saranya S S, “Text to Speech Conversion System using OCR”, International Journal of Emerging Technology and Advanced Engineering, January 2015.

[6] Itunuoluwa sewon, Jelili Oyelade, Olufunke Oladipupo, “Design and Implementation of Text To Speech Conversion for Visually Impaired People”, International Journal of Applied Information Systems (IJAIS, 2014).

[7] Yao Li and Huchuan Lu, “Scene Text Detection via Stroke Width”, 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.