INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

OPTICAL CHARACTER RECOGNITION SYSTEM

[1.] MR. B.NARSINGAM, [2.] A.SANDHYA,[3.] K.VARSHINI, [4.] D.MANISHA

[1.] Assistant Professor Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Rangareddy (TS).India.

Email-:[1.] tkecsebnarsingam@gmail.com

[2,3,4.]B.Tech StudentstDepartment of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Rangareddy (TS).India.

Email-:[2.] arepellysandhya@gmail.com, [3.] reddyvarsha336@gmail.com,

[4.] dm257448@gmail.com .

Abstract- This project speaks about the objective of developing a system that can help disabled people in their day-to-day activities. Many challenges were faced by visually impaired people. In most instances, they require constant support in all situations, especially in their daily activities. Some of the major challenges include difficulty in moving from one place to another without the assistance of others. In addition to that, they face difficulty in recognizing people, detecting obstacles, etc. To overcome this condition, we put forward "optical character recognition". This system guides the visually impaired person. A camera will capture the live footages. Optical Character Recognition (OCR) is used to extract the text from the images. The data read is converted to speech using text to the speech synthesizer. The system is implemented using various available technologies that helps visually impaired. The paper discusses the design of system and the challenges involved in designing the device. .

KEYWORDS: Optical Character Recognition (OCR), GTTS, Camera, Visually Impaired, Human Emotion, Pyttsx3.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

1. INTRODUCTION

OCR is defined as the process of electronic or mechanical conversion of documents or text embedded in images to computer understandable ASCII characters. Actually, text embedded in images or documents is converted into machine code non variation. It also minimizes human effort and increases data reliability. With OCR, you'll be able to convert any physical paper document, or an image into an accessible electronic version with text. For example, if you scan any image or document with a scanner, the output will most likely be a digital file (Which isn't machine-encoded) with ".jpg" or ".pdf" format, but on loading the identical with an OCR tool (Tesseract engine, for example). The tool recognizes the text and saves the output as an editable & computer-understandable text file. OCR involves two important steps, viz- text/entity detection, in which the text embedded in an images or document is localized. Followed by text recognition, text extraction from image. OCR can be used for both handwriting recognition and printed text recognition. Handwriting character recognition is a complex task because of variety of writing styles and letter strokes of different users. This SLR is purely based upon printed character recognition system. Any discussion related handwriting OCR is outside the scope of this research paper. There are various types of OCR engines used in existing research studies and text extraction, such as: Desktop OCR, Server OCR, Web OCR, OmniPage Ultimate, Readiris, etc. But only a few of them are free and open source. Their accuracy varies from 70% to 98%, depending upon document images & noise in them. In this research study, we've utilized Tesseract OCR engine.

2. LITERATURE SURVEY

Image Processing

Image processing is analysis and manipulation of a digitized image, so as to enhance its quality with the help of mathematical operations by using any kind of signal processing where the input is a picture or an image or a video frame. The output of image processing will be either a

picture or set of characters or parameters associated with the given input image. This is a set of computational techniques for analyzing, enhancing, compressing and reconstructing image.

Text Recognition Software

Text recognition software is an automatic tool that recognizes text within image files, so that you do not have to retype it yourself. This is especially important when working with scanned documents and image files that contain text that might not be recognized by the computer. Text recognition software, known as OCR, resolves this problem by making the documents text-searchable.

Text in images can exhibit many variations with respect to the following properties:

1. Geometry:

Size: Although the text size can vary a lot, assumptions can be made depending on the application domain.

Alignment: The characters in the caption text appear in clusters and usually lie horizontally, although sometimes they can appear as non-planar texts as a result of special effects. This does not apply to Text, which can have various perspective distortions. Text can be aligned in any direction and can have geometric distortions.

Inter-character distance: Characters in a text line have a uniform distance between them.

2. Colour:

The characters in a text line tend to have the same or similar colours. This property makes it possible to use a connected component-based approach for text detection. Most of the research reported till date has concentrated on finding 'text strings of a single colour (monochrome)'. However, video images and other complex colour documents can contain 'text strings with more than two colours (polychrome)' for effective visualization, i.e., different colours within one word.

3. Motion:

The same characters usually exist in consecutive frames in a video with or without movement. This property is used in text tracking and enhancement. Caption text usually moves in a uniform way: horizontally or vertically. Scene text can have arbitrary motion due to camera or object movement.

4. Edge:

Most caption and scene text are designed to be easily read, thereby resulting in strong edges at the boundaries of text and

background.

5. Compression:

Many digital images are recorded, transferred, and processed in a compressed format. Thus, a faster TIE (Text Information Extraction) system can be achieved if one can extract text without decompression.

3. EXISTING SYSTEM:

In the current world there is growing demand for the users to convert printed documents into electronic documents for maintaining the security of their data. Hence the basic Text recognition system was invented to recognize and convert the data available on papers into computer processable documents, So that the documents can be editable and reusable. The existing system/previous system of Text recognition on a grid infrastructure is just a text recognition system without grid functionality. That is the existing system deals with the homogeneous character recognition or character recognition of single language.

The drawback in the early text recognition system is that they only have the capability to recognize and convert only the images of English or only of single language. That is the older Text recognition system is Unilingual.

4. PROPOSED SYSTEM:

Our proposed system is Extraction of text from an image using Tesseract OCR engine, on a grid infrastructure which is a character recognition system that supports recognition of the characters of multiple languages. This feature is what we call grid infrastructure which eliminates the problem of heterogeneous character recognition and supports multiple functionalities to be performed on the image. In this context, Grid infrastructure means the infrastructure that supports group of specific set of languages. Thus, Extraction of text from an image using Tesseract OCR engine on a grid infrastructure is multi-lingual.

The benefit of proposed system that overcomes the drawback of the existing system is that it supports multiple functionalities such as editing and searching. It also adds benefit by providing heterogeneous characters recognition.

This system recognizes the characters approximately based on their trained data values.

Tesseract OCR engine is used to match the

recognized character with the trained dataset.

5. MODULES:

PRE-PROCESSING MODULE

The paper document is generally scanned by the optical scanner and is converted in to the form of a picture. At this stage we have the data in the form of image and this image can be further analyzed so that's the important information can be retrieved. The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a pre-processor to smooth the digitized characters.

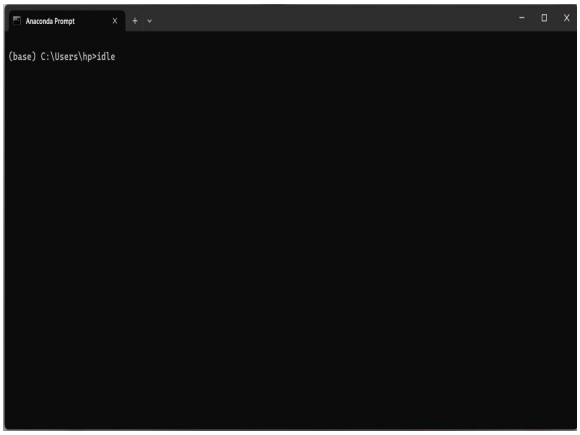
TEXT RECOGNITION MODULE

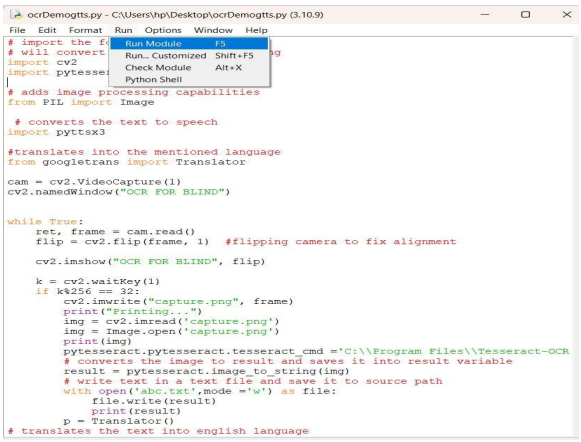
The preprocessing the image to make it suitable for the Convolutional Neural Network (Deep Neural Network) this module can be used for text recognition in output image of pre-processing module and give output data which are in computer understandable form. Hence in this module following techniques are used.

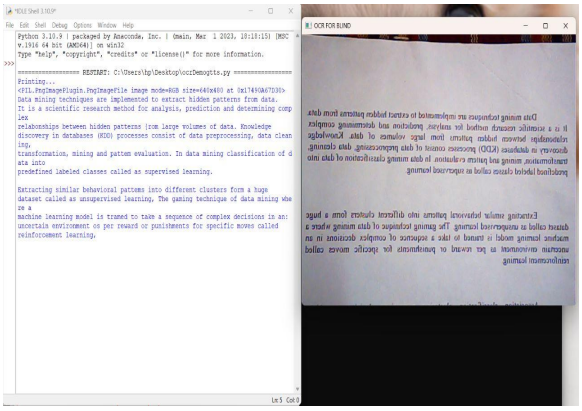
POST-PROCESSING MODULE

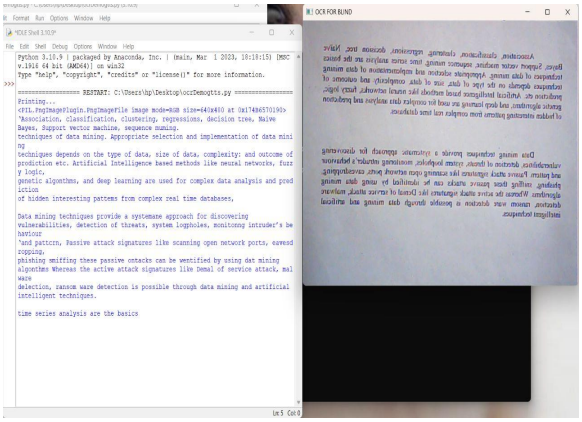
Post processing attempts to increase the quality of a mask image. Post processing is performed with the help of Morphology. The morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image. It is the final step of recognition system being discussed. It prints the corresponding characters which were recognized in the structured text form which is done by the calculation of equivalent ASCII value using recognition index of the test samples.

6. RESULTS









7. CONCLUSION

In this paper, a model has been created with GTTS (Google text to speech) - based OCR application using Tesseract OCR engine. The application successfully recognizes text of various fonts from different input images and transforms the OCR output into an audio output, which is extremely accurate. We have implemented Tesseract Engine because it is the most powerful and open source software, and also it does not require any license/investment. It will assist visually impaired people so that they can read the document using this aiding methodology. Experiments have been performed by evaluating visual comparison of OCR-Test Cases and good results have been achieved by calculating the precision of the model. The model has been a far better methodology in OCR in any input document or image and transcribing it into audio output. The proposed system eases the digital experience of people having learning disabilities, reduced vision or those with literary difficulties.

8. REFERENCES

1. L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In ACCV, pages 770–783, 2010.

2. Reza Sarshogh and Keegan Hines, ”Computer Vision Methods for Extracting Text from Images”, Capital One Tech.

3. T.Som, Sumit Saha," Handwritten Character Recognition Using Fuzzy Membership Function", International Journal of Emerging Technologies in Sciences and Engineering, Volume 5, December 2011.

4. L. A. Zadeh. Fuzzy sets, Information Control 8 (1965) 338–353.

5. Gur, Eran, and ZeevZelavsky, "Retrieval of Rashi Semi-Cursive Handwriting via Tesseract OCR engine ," IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012

6. Thomas Natsvhlager, "Optical Character Recognition", A Tutorial for the Course Computational Intelligence.

7. Andrei Polzounov,ArtsiomAblavatski ,Sergio Escalera, Shijian Lu, JianfeiCai "Wordfence: Text Detection In Natural Images With Border Awareness".

8. D.Trier ,A.K.Jain ,T.Taxt , "Feature Extraction Method for Character Recognition-A Survey" ,Pattern Recognition