International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 749-753

OCR Of Handwritten Forms

Prabhu A[1], Pratyasha Panda[2], N. Prerana[3], Saquib Ahmed Khan[4]
[1] *Associate Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal Rd, Hyderabad, Telangana, India.*
[2] *UG Student, Department of Computer Science and Engineering, CMR Technical Campus, Medchal Rd, Hyderabad, Telangana, India.*
[3] *UG Student, Department of Computer Science and Engineering, CMR Technical Campus, Medchal Rd, Hyderabad, Telangana, India.*
[4] *UG Student, Department of Computer Science and Engineering, CMR Technical Campus, Medchal Rd, Hyderabad, Telangana, India.*

ABSTRACT
Given the presence of written documents in human transactions, Optical Character Recognition (OCR) of documents have valuable and sensible value. Optical Character Recognition could be a science that permits the translation of varied forms of documents or pictures into readable, editable, searchable and understandable information. The purpose is to develop user-friendly applications which carry out conversion of image, pdfs into text forms. The OCR takes an image/pdf as the input, gets text from that image/pdf and then as a result the user can view the converted document. This system can be useful in various applications like banking, schools, railway and other industries. It is mainly designed to save time and labor cost.
Keywords: *Recognition, Tesseract OCR, Optical Character Recognition (OCR), Template.*

Date of Submission: 01-06-2022 Date of acceptance: 12-06-2022

I. INTRODUCTION

With everybody moving towards digital revolution, there has been a hike of applications written by humans of vital data that makes it an issue once storing on-line. Human written recognition tends to be tough as a result each human has its own approach of writing that is restricted and distinctive. The variations in handwriting designs creates drawbacks for the Optical Character Recognition engines, that square measure usually trained on laptop fonts,not handwriting fonts.

Character Recognition may be a common technique of digitizing written texts in order that it is electronically emended, searched, employed in machine processes like computational linguistics, text-to-speech, key knowledge and text mining. This technique may be a straightforward approach to convert the input text into text kind, which springs from concepts of various researchers who have given their valuable contribution in developing algorithms for OCR. Tesseract OCR by Google provides such an engine to convert written forms into text forms.

A. TESSERACT OCR

Tesseract is an open source optical character recognition (OCR) platform. OCR extracts text from pictures and documents and outputs the document into a new searchable document, PDF, or most alternative standard formats. Tesseract is extremely customizable and can be used in different languages and on any OS. It uses algorithms like Otsu for image segmentation and Hough rework for feature extraction for doing varied other jobs.

B. OPENCV

OpenCV, is an open-source computer vision and Machine Learning library, is employed to tell apart and understand faces, objects, follow eye movements, track camera activities, hunt nearly identical photos from an image data info, the prompt technique makes use of those OpenCV options.

II. PROJECT OBJECTIVES

The objective is to provide a system where we don't have to do paperwork or use some software that does not fulfill all the requirements. The newly introduced system will provide easy access to the system and it will contain user friendly functions with an attractive interface. This will help in assembling data as per user

choice of templates and store them as excel sheets. the application will provide an easy way of conversion without the requirement of signing up.

III. LITERATURE SURVEY

The review method was adopted by measuring the analysis in the last few years for extraction of information. The thirty one analysis articles were reviewed to cover the review of character recognition technique.
Example-
Mohammed Z. Khedher, Gheith A. Abandah, and Ahmed M. Al-Khawaldeh 2005. This paper describes that Recognition of characters greatly depends upon the features used. Several features of the handwritten Arabic characters are selected and discussed. An off-line recognition system based on the selected features was built. The system was trained and tested with realistic samples of handwritten Arabic characters. Evaluation of the importance and accuracy of the selected features is made. The recognition based on the selected features give average accuracies of 88% and 70% for the numbers and letters, respectively. Further improvements are achieved by using feature weights based on insights gained from the accuracies of individual features[1].

MAJIDA ALI ABED HAMID ALI ABED ALASADI [2005]. This manuscript considers a new approach to Simplifying Handwritten Characters Recognition based on simulation of the behavior of schools of fish and flocks of birds, called the Particle Swarm Optimization Approach (PSOA). We present an overview of the proposed approaches to be optimized and tested on a number of handwritten characters in the experiments. Our experimental results demonstrate the higher degree of performance of the proposed approaches. It is noted that the PSOA in general generates an optimized comparison between the input samples and database samples which improves the final recognition rate. Experimental results show that the PSOA is convergent and more accurate in solutions that minimize the error recognition rate[2].

IV. METHODOLOGY

There are differents algorithms that the tesseract OCR consists and following are examples of such -

A. HOUGH TRANSFORM

The Hough transform is an approach which can be used to insulate features of a particular shape within an image. Hough transform is most generally used for the discovery of regular angles similar as lines, circles, spheres, etc. The main advantage of the Hough transform approach is that it's tolerant of gaps in point boundary descriptions and is fairly unchanged by image noise.

B. OTSU ALGORITHM

It starts with processing of the input image where histogram of the image is obtained which is done by minimization of the variance of each classes for segmenting the images and the threshold T value is calculated which then replaces image pixels into white in the regions where saturation is greater than T and into the black otherwise. The computation equation can be described as:

$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t)$, where $w_1(t), w_2(t)$ [3] are the probabilities of the two classes divided by a threshold T, whose value is within the range from 0-255.

V. WORKING

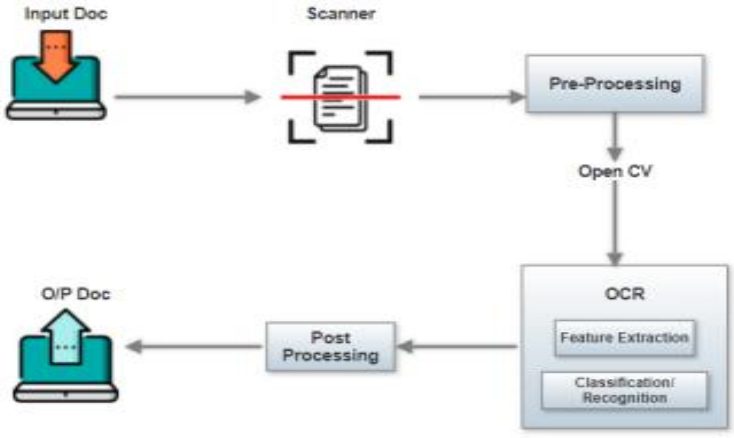

Figure 1: Project architecture

When a user logs into the website, they can choose their file from their directories and upload. The user can also select templates of their choice from the document for generating excel sheets. It will return the resulting text form or excel sheet to the user.

A. IMAGE CAPTURING
This step consists of scanning. In scanning, a digital image of typewritten or hand-written or image document is captured. To perform this method, optical scanners are used which are easily accessible from the market. Optical scanners convert coloured levels into gray-levels for machine understanding purposes.
B. PRE-PROCESSING
Preprocessing of an image is performed to enhance the possibilities of successful recognition. Normally noise filtering, smoothing filing and thinning are performed during this step. Image that was scanned during image capturing might contain a precise amount of noise.
Techniques of preprocessing include:
De-skew: aligns the document properly either few degrees in clockwise or counterclockwise direction therefore to form lines of text absolutely horizontal or vertical.
DE speckle: removing spots i.e. positive or negative spots and smoothing of edges.
Line removal: removes gratuitous lines.

C. FEATURE EXTRACTION
In this part, features of each character are extracted. The extracted features from input character ought to permit classification in an exceedingly distinctive approach. Different types of features such as the image itself, geometrical features and statistical features can be used.

D. CLASSIFICATION AND RECOGNITION
The aim of creating a part of the recognition system is the classification stage. Here the matrix containing the image of input characters is directly matched with the set of paradigm characters representing every possible category. The category of prototype giving the simplest match is allotted to the pattern.

E. POST-PROCESSING
At the final stage, it prints the recognized characters in structured text form. We get a set of individual characters, but these characters in themselves do not contain meaningful information. But once grouped in the form of string or word they can convey a meaning.

VI. RESULT

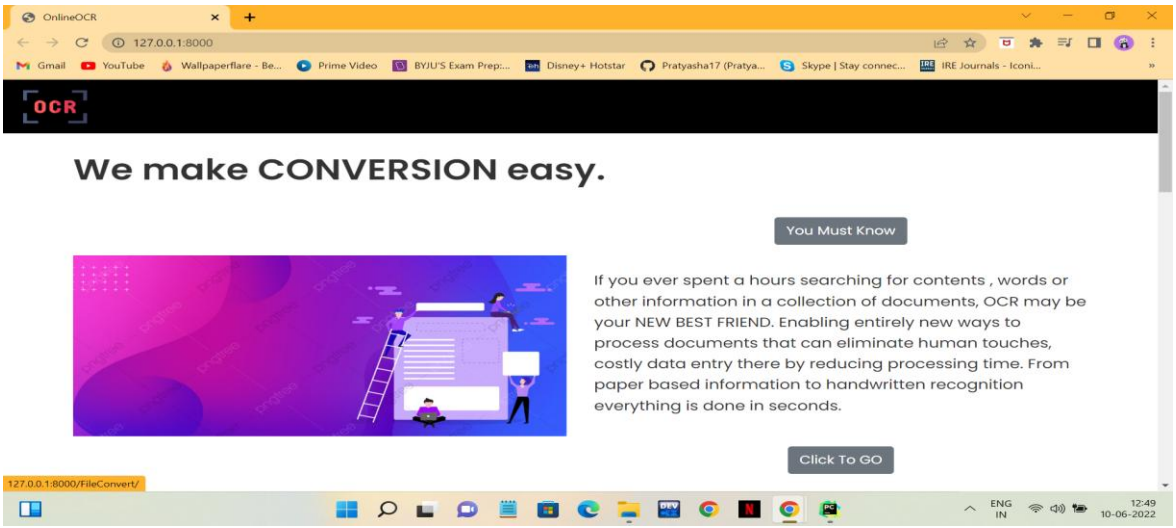

Figure 2: Website

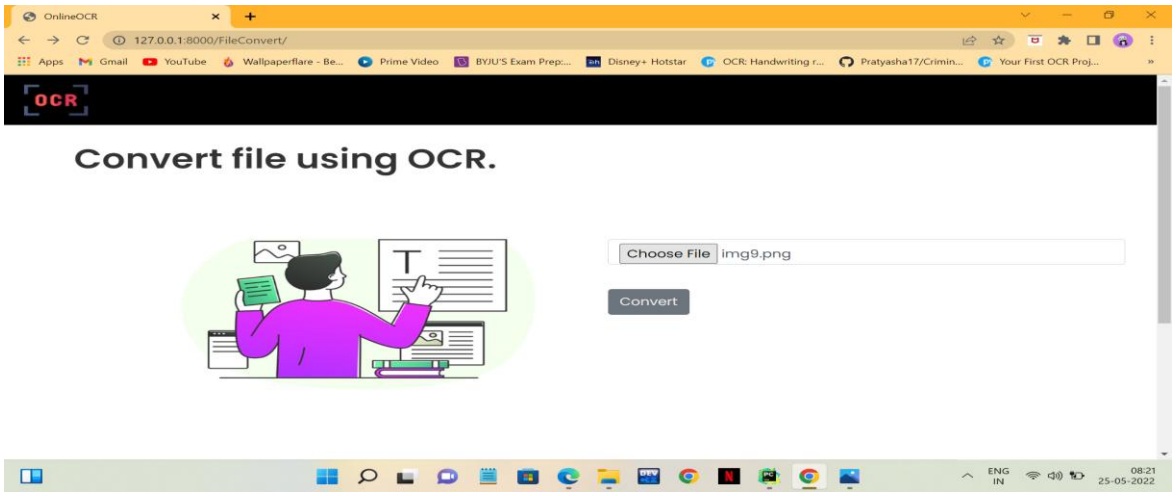

Figure 3: Choosing image for conversion

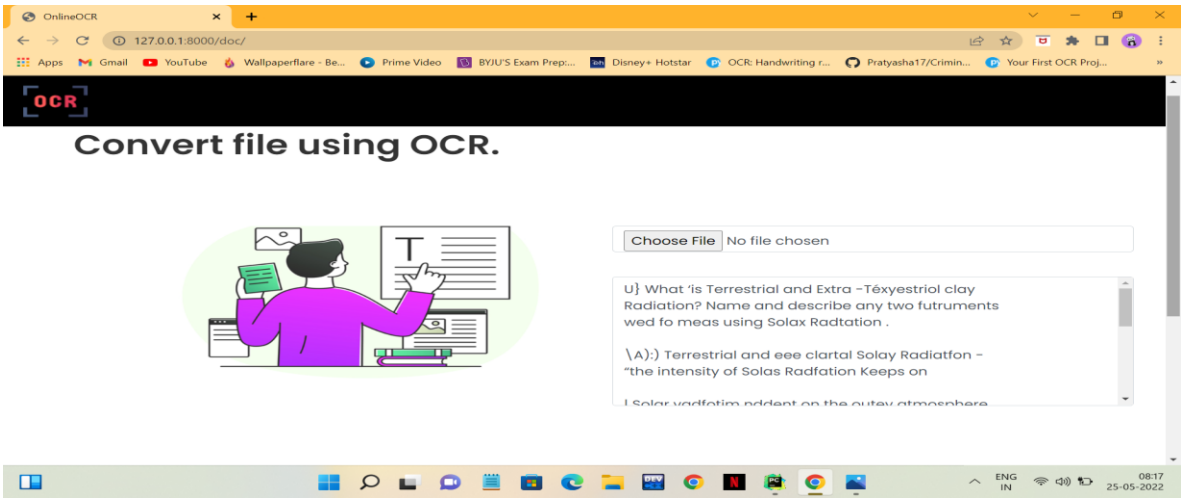

Figure 4: Converted into text

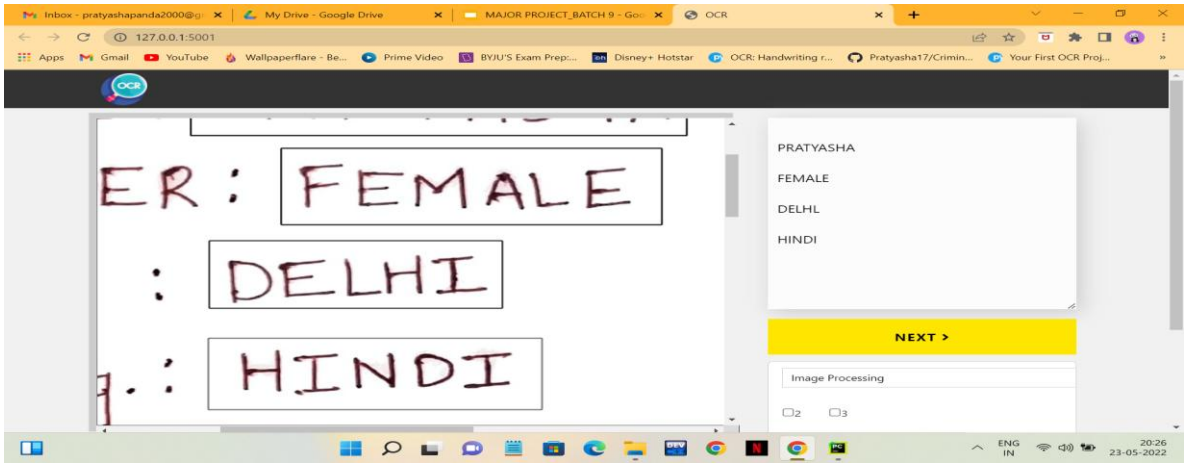

Figure 5: Selecting templates

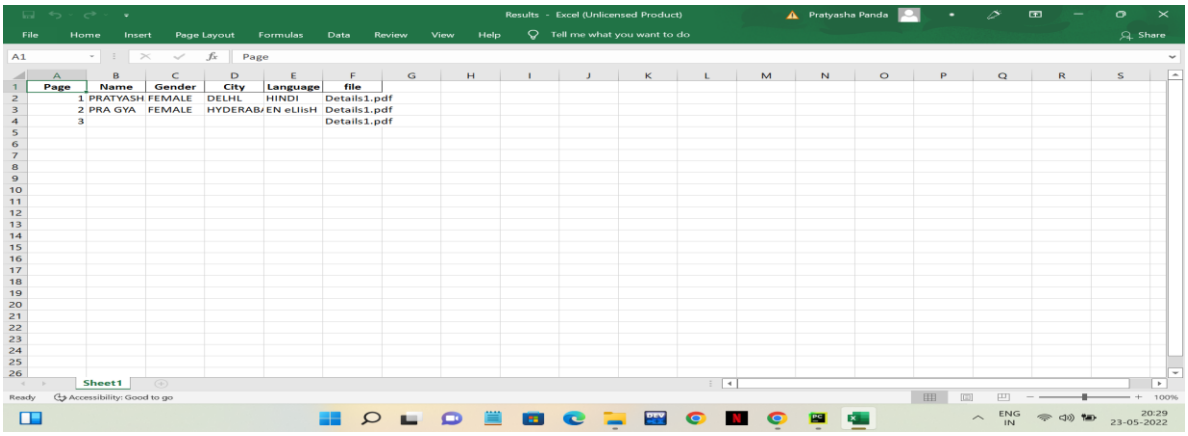

Figure 6: Resulted excel sheet

VII. FUTURE SCOPE

This platform will provide individuals, organizations, industries or businesses to get their data in one form for easy usage. This will help in getting large data into text form with higher quality and in minimum time. OCR's can be introduced in different languages for people to access easily. Also, it will be able to recognize even the blurriest pictures with high efficiency which will ease the human work and can be used in the advancement of robots or to make robots use OCR technique to read papers.

VIII. CONCLUSION

This project converts unstructured information into structured information. . The project provides an adequate plan on developing a full-fledged application satisfying the user necessities. The system is extremely versatile and resilient. Validation checks induced have greatly reduced errors. Provisions are created to upgrade the software system. The application has been tested with live information and has provided a roaring result. Thus the software system has been established to figure expeditiously.

ACKNOWLEDGMENT

We are thankful to our Project Guide Dr. Prabhu A and Project Coordinator Mr. J. Narasimha Rao, for their valuable guidance, genuine suggestion and constant encouragement during preparation of project paper work without which completion of this project would be a difficult task.

REFERENCES

[1]. Mohammed Z. Khedher, Gheith A. Abandah, and Ahmed M. Al-Khawaldeh, "Optimizing Feature Selection for Recognizing Handwritten Arabic Characters", published under PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY, vol. 4, Feb 2005, ISSN 1307- 6884.

[2]. Majida Ali Abed, Hamid Ali Abed Alasadi, "Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach", published under EUROPEAN ACADEMIC RESEARCH, Vol. 1, Issue 5, Aug 2013 ISSN 2286-4822.

[3]. https://learnopencv.com/otsu-thresholding-with-opencv/

[4]. S. Madhusudhan, R. Venkat Tarun, Kishore Sarangi, "Digital Image Processing", published under INTERNATIONAL PEER-REVIEWED, Vol. 5, Issue 1, July 2021, ISSN 2456-8880.

[5]. Ankit Kumar Singh, Aman Gupta, Aman Saxena, "Optical Character Recognition: A Review", published under JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH, Vol. 3, Issue 4, April 2016, ISSN 2349-5162.

[6]. Swati Tomar, Amit Kishore,"A Review: Optical Character Recognition", published under INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, April 2018, ISSN 2277-9655.