iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5291

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Text Extraction from Images Using OCR

Jyothi E[1], K Tejaswini[2], Lakshmi Chintalapati[3], Mr. MD. Shafiulla[4]
[1, 2, 3]*Student,* [4]*Assistant Professor, Ballari Institute of Technology and Management, Ballari*

Abstract: ***Nowadays, there is an enormous demand in storing information available on papers, such as books or newspapers. There is an existing way to store information by scanning the desired text, but it will be stored as an image that won't help for further processing. For instance, if stored scanned text images, can't read the text word by word, or line by line; the text in these scanned images can't be reused unless we rewrite that whole content by ourselves. Detection of text from documents in which text is embedded in complex colored document images is a very challenging problem. There are a lot of potential users who want to extract the text from images, archiving documents etc. For this reason, user need an Optical Character Recognition (OCR). It aims at detecting textual regions from the document and separating it from the graphics portion. Getting information directly from applications forms and it saves a lot of time.***

I. INTRODUCTION

Text Extraction involves a computer system designed to translate captured or scanned documents into Machine editable text. OCR began as a field of research in artificial intelligence and computational Vision. It is developed by using the image processing and it used widely in the information from these image documents would give higher efficiency and ease of access if it is converted to text form. The process by which Image Text converted into plain text that computer can recognize its ASCII character is Text Extraction. The users can scan a document and have the text of that document in .txt or .doc..

II. LITERATURE SURVEY

This chapter shows various analyses and research made on this project and the results already published, taking into account the various parameters of the project and the extent of the project.

A. Papers Referred

1) *"Extracting Text from Image Document and Displaying ITS Related Information", K. N. Natei, journal of Engineering Research and Application:* Image Text is the text information embedded or written in image of different form. Image text can be found in captured images, scanned documents, magazines, newspapers, posters etc. These image texts are highly available nowadays and they are very important in representing, describing and transferring information which help peoples in communication, solving problems, availability, creation of new types of jobs, cost effectiveness, productivity, globalization and cultural gap etc. The information from these image documents would give higher efficiency and ease of access if it is converted to text form. The process by which Image Text converted into plain text that computer can recognize its ASCII character is Text Extraction.
2) *"Text Recognition using Image Processing", International journal of Advanced Research in Computer Science*: Text Recognition using OCR involves computer system designed to translate images of typewritten text into machine editable text or to translate pictures of characters into a standard encoding scheme representing them. OCR began as a field of research in artificial intelligence and computational vision. Text Recognition used in official task in which the large data have to type like post offices, banks, colleges etc., in real life applications where we want to collect some information from text written image.

III. EXISTING SYSTEM

Off-line handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. And, as of today, OCR engines are primarily focused on machine printed text and ICR for hand "printed" (written in capital letters) text. Pranob K Charles, V.Harish, M.Swathi ,CH. Deepthi : A Review on the Various Techniques used for Optical Character Recognition. In this paper various approaches used for the design of OCR systems are discussed. It presents the techniques that are slow which provide better results in nature and also the fast techniques that provide inefficient results. The first prominent piece of OCR software was invented by Ray Kurzweil in 1974 as the software allowed for recognition for any font. This software

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

used a more developed use of the matrix method (pattern matching). Essentially, this would compare bitmaps of the template character with the bitmaps of the read character and would compare them to determine which character it most closely matched with.

A. Disadvantages of Existing System

1) Sensitive to variations in sizing.
2) The distinctions between each individual's way of writing.
3) Inefficient Results with less probability .

IV. PROPOSED SYSTEM

In the devised model as shown the idea proposed is to take in a number of images of documents like identity proofs of individuals and classify them into classes, such as passport and license.Once the images are classified, they are subjected to the text extraction module. The text data are extracted from the classified images. The extracted credentials from the images are then stored in the database. Text Extraction Text extraction is implemented using the Tesseract OCR package which contains an optical character recognition (OCR) engine - libtesseract and a command line program –Tesseract. Tesseract includes a new neural net called Long Short-Term Memory (LSTM) based OCR engine, which focuses on line recognition and also recognizes character patterns. The LSTM network is the units of Recurrent Neural Networks. The Python-Tesseract is an optical character recognition (OCR) tool in python used for text extraction.

V. DESIGN

A. Block Diagram

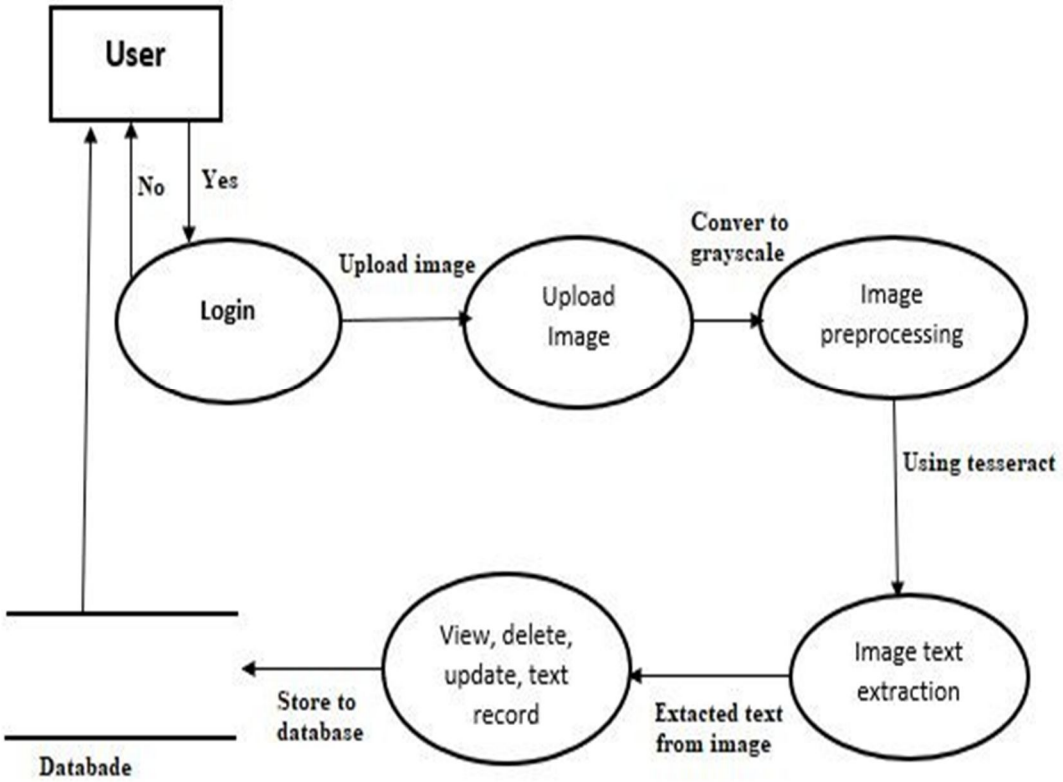


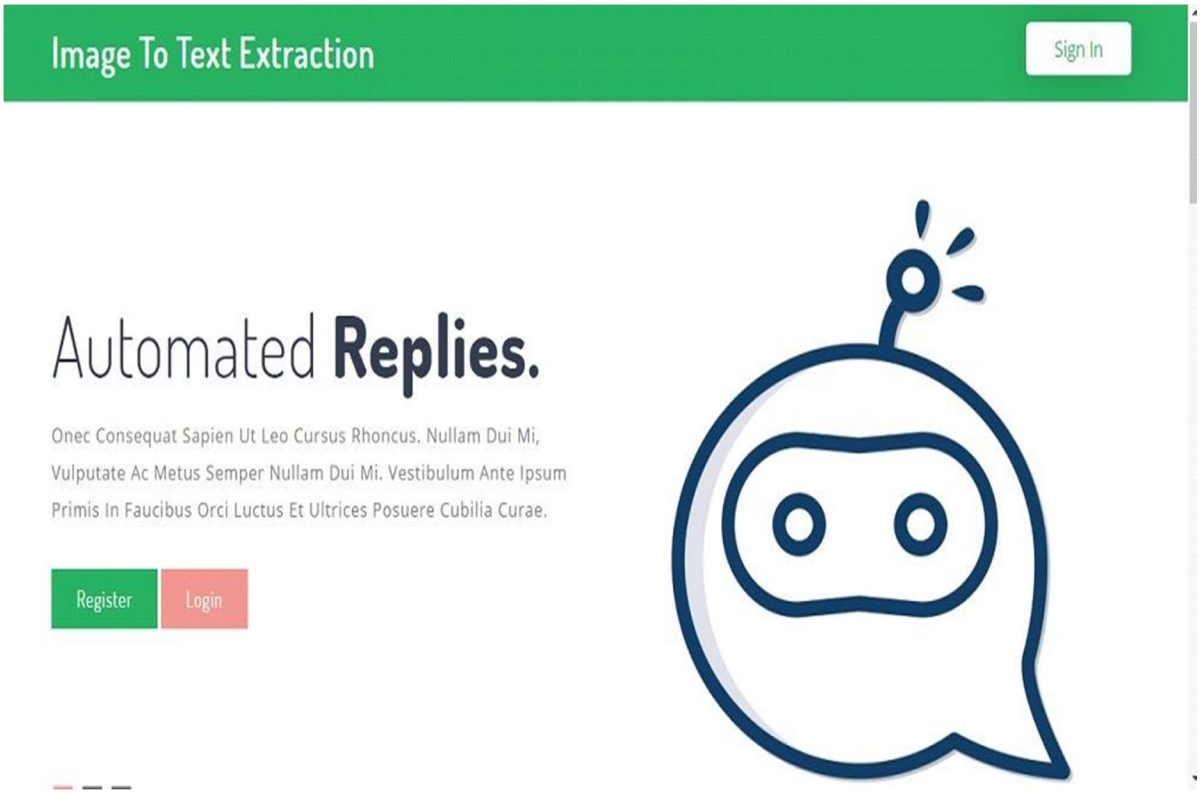
Fig1: Block Diagram

The data flow of this project is users upload the image then image scaling or gray scale conversion takes place and store the record. once it is stored so we can view, delete, update, and record the text. And it is stored in the database.

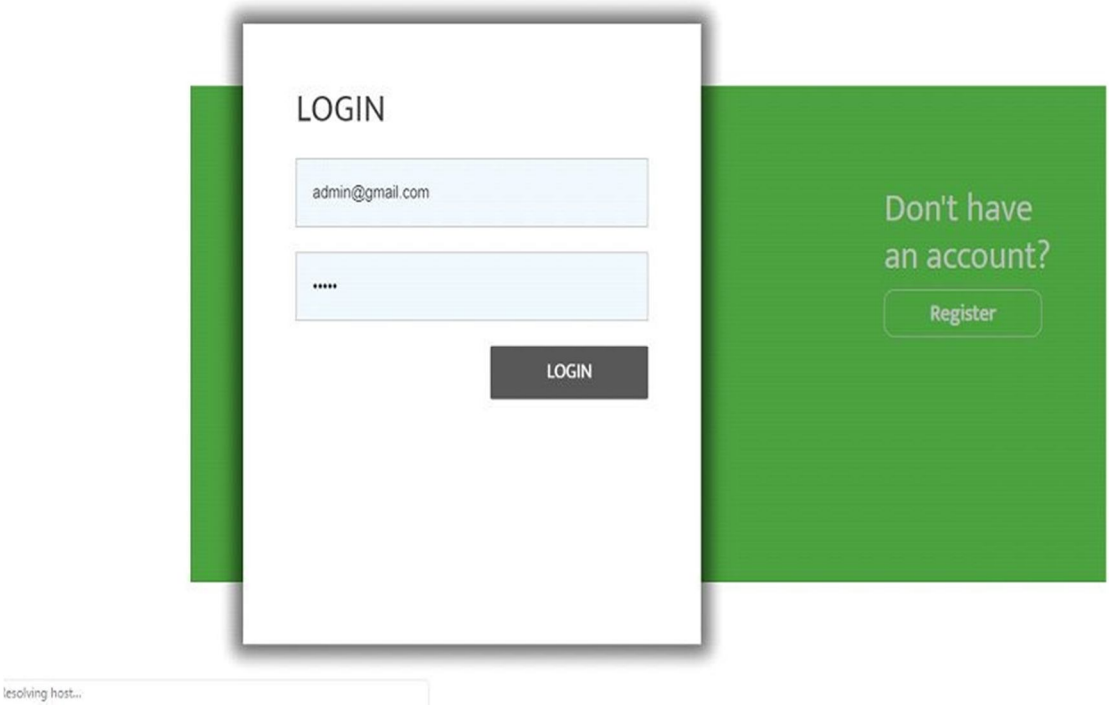
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

VI. SNAPSHOTS

A. Home Page

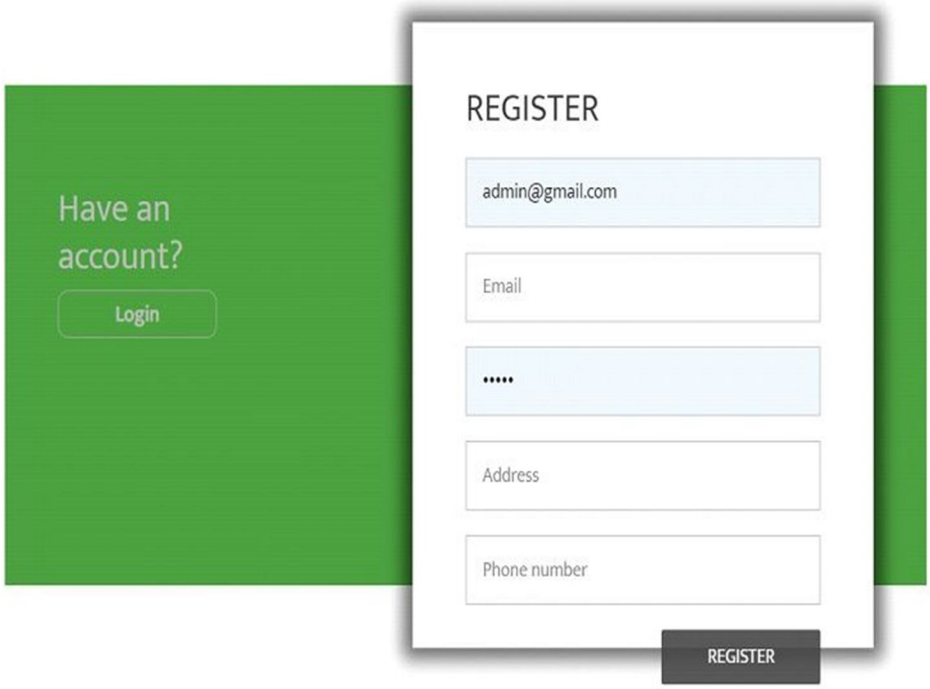


B. Login Page

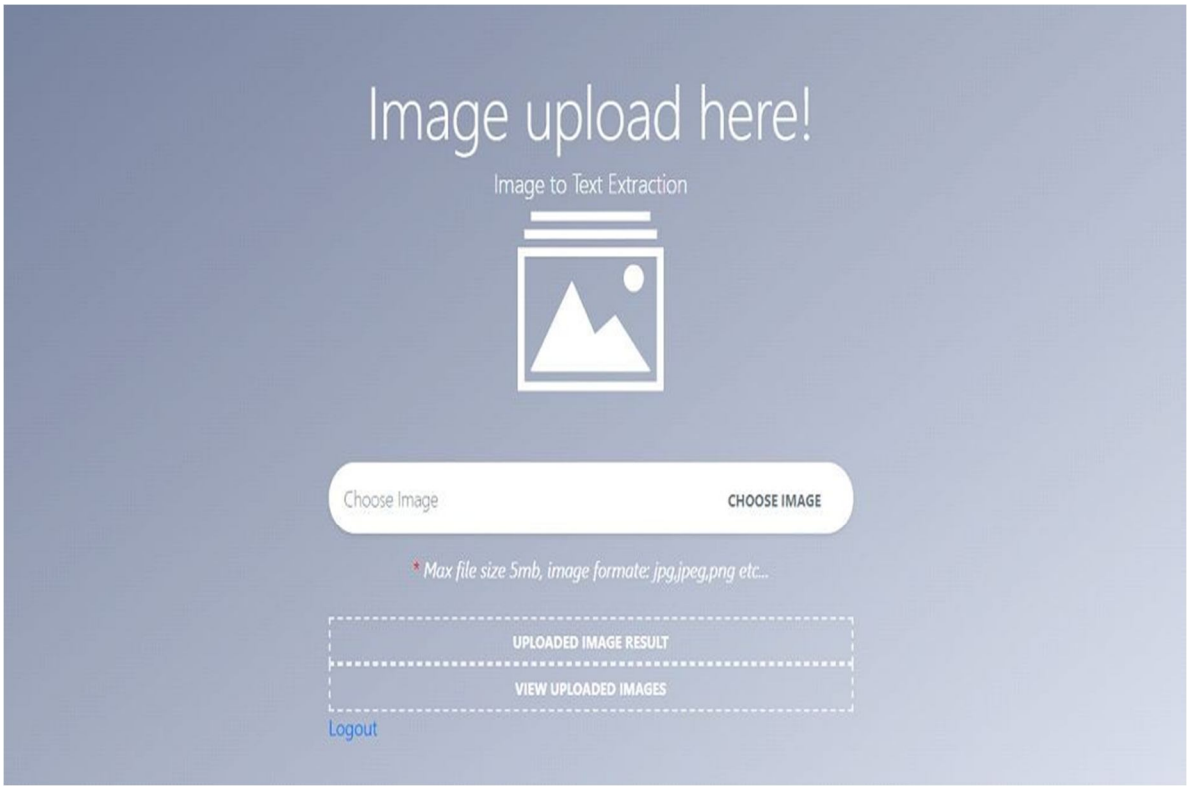


ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

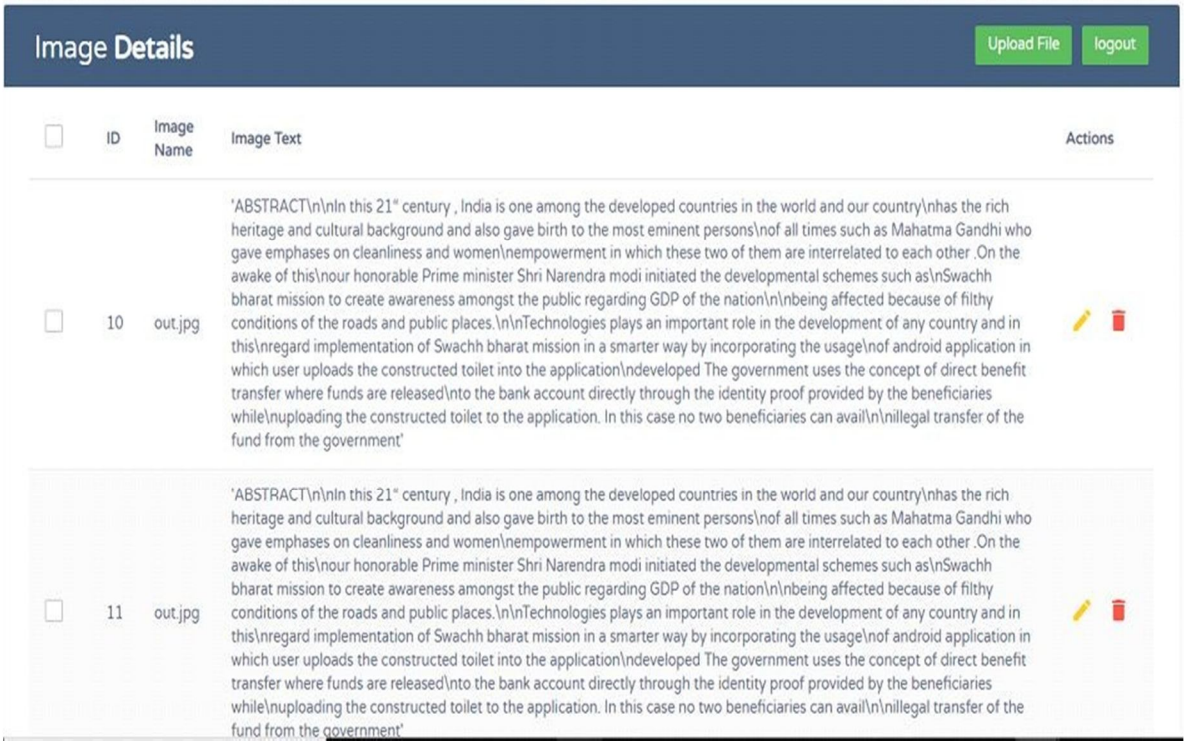
C. Register Page



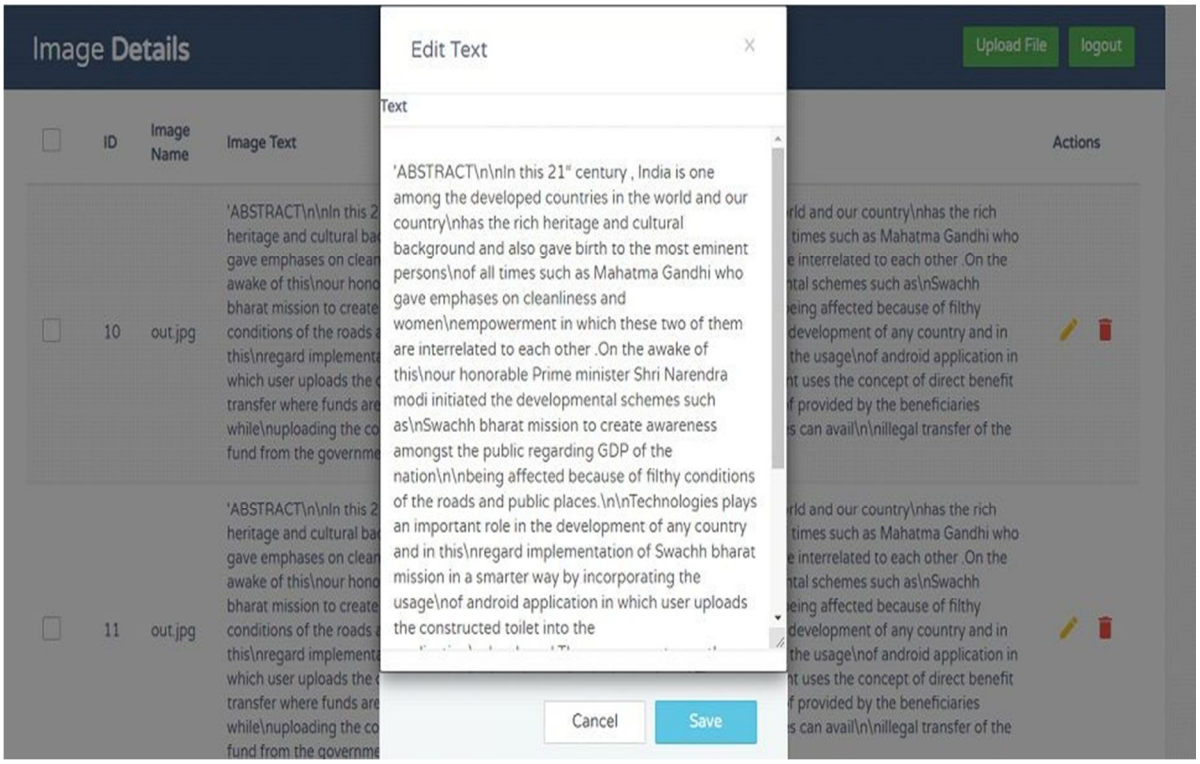
D. Upload Image



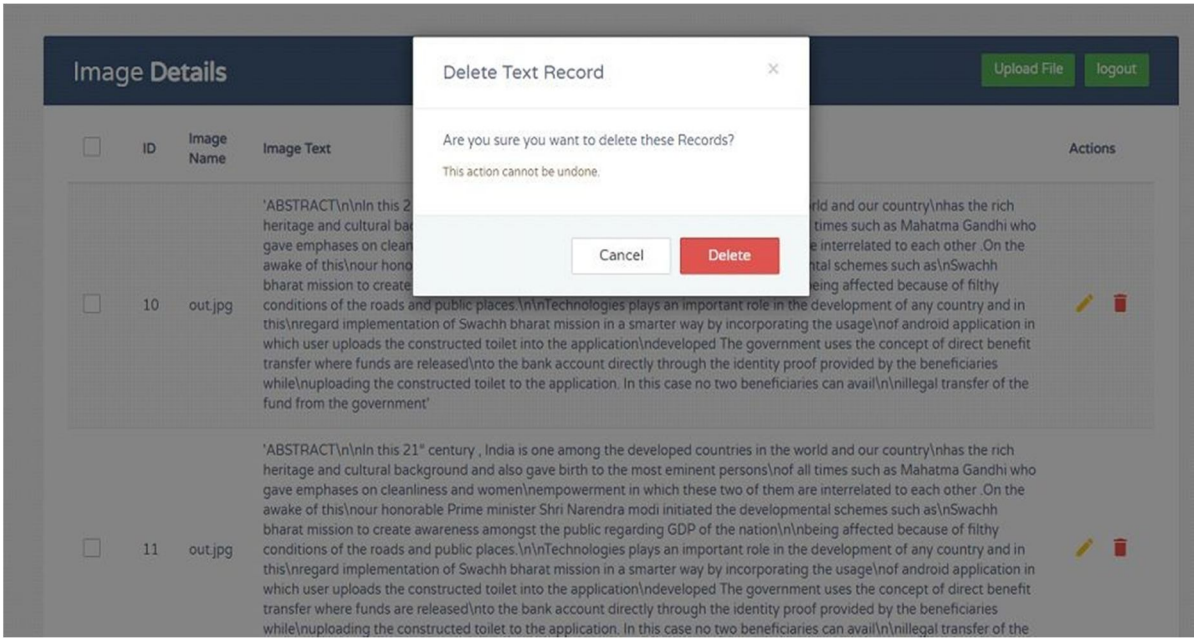
E. View Text Record



F. Edit Text Record



G. Delete Text Record

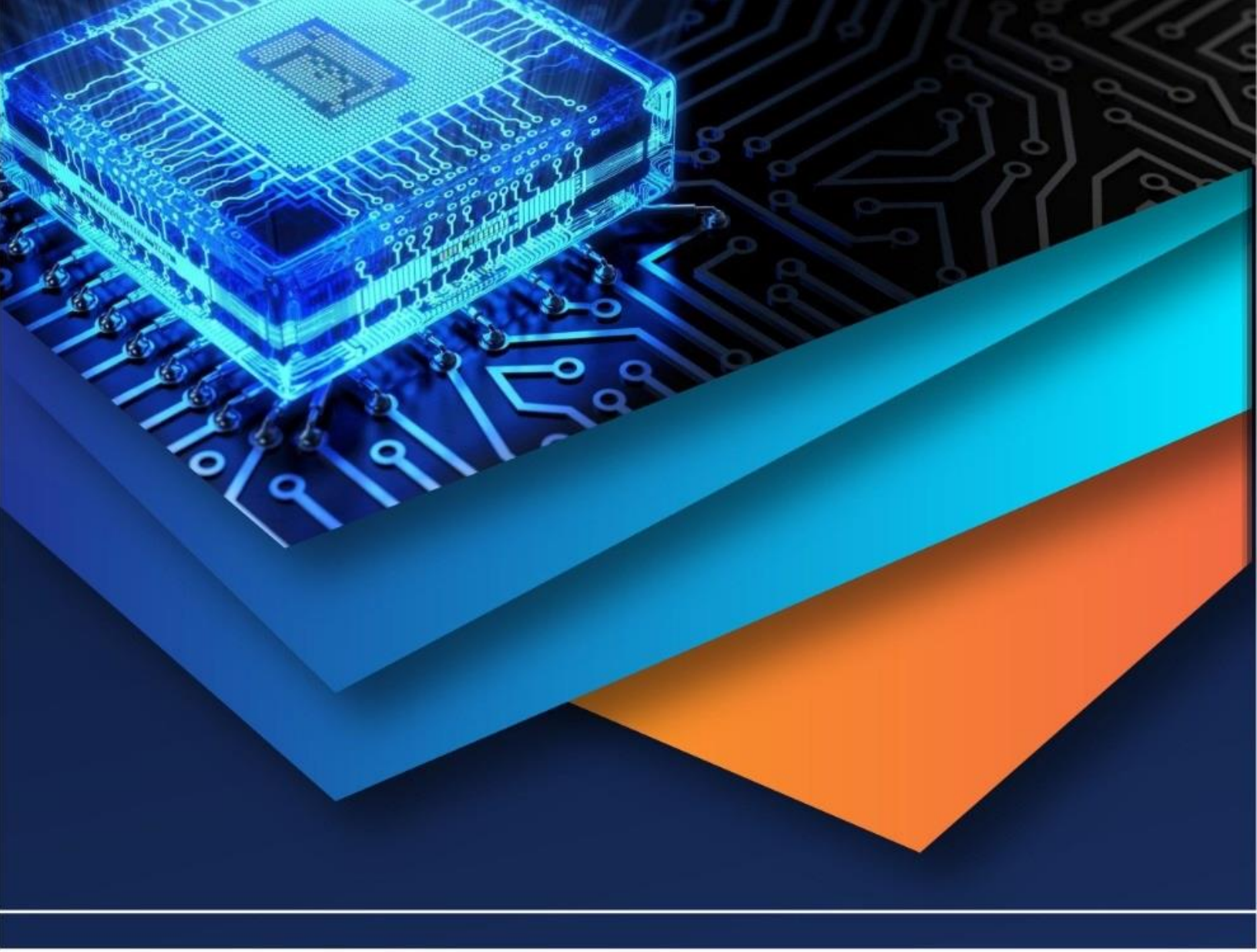


VII. CONCLUSION

The application which is proposed in this report can be effective which will save a lot of time and money of common people. The main objective behind developing this application to automate and make a system that provides a reliable and an efficient way of recognizing text in scanned text documents, text images, and any picture in order to reuse itlater. This report attempts to summarize what is the project all about, technologies used, database being used, the main cause for which the application is to be developed and the techniques which are required to make the project functional.

REFERENCES

[1] "Extracting text from image document and displaying its related information", K.N. Natei journal of Engineering Research and Application (ISSN : 2248-9622, Vol. 8, Issue5 (Part -V) May2018.
[2] K. Gaurav and Bhatia P. K., "Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition", 2nd International Conference on Emerging Trends in Engineering & Management, ICETEM, 2013.
[3] J. Pradeep, E. Srinivasan and S. Himavathi, "Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.
[4] "Text Recognition using image processing", International journal of Advanced Research in Computer Science by Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmakar (Vol-8,No.5,May-June 2017).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY