Project Synopsis

on

**OPTICAL CHARACTER RECOGNITION**

Submitted as a part of the course curriculum for

**Bachelor of Technology**

in

**Computer Science**



**Submitted by**

ANOOP KUMAR (1900290120017)

HIMANCHAL K GAUTAM (1900290120046)

**Under the Supervision of**

Ms. AKANKSHA

ASSISTANT PROFESSOR

**KIET Group of Institutions, Ghaziabad**

**Department of Computer Science**

**Dr. A.P.J. Abdul Kalam Technical University**

**2022-2023**

**DECLARATION**

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature of Students:

Name: ANOOP KUMAR & HIMANCHAL Kr. GAUTAM

Roll No.: 1900290120017 & 1900290120046

Date:

**CERTIFICATE**

This is to certify that Project Report entitled “**OPTICAL CHARACTER RECOGNITION**” which is submitted by **ANOOP KUMAR & HIMANCHAL KR. GAUTAM** in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidate’s own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

**Date: 25/12/2021 Supervisor Signature**: Ms. Akanksha (Assistant Professor)

**ACKNOWLEDGEMENT**

It gives us a great sense of pleasure to present the synopsis of the B. Tech Mini Project undertaken during B.Tech. Third Year. We owe a special debt of gratitude to Prof. AKANSHA, Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his/her constant support and guidance throughout our work. Her sincerity, thoroughness, and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavors have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Dr. P. K Singh, Head of the Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Date: 25/12/2021

Name ANOOP KUMAR & HIMANCHAL Kr. GAUTAM

Roll No: 1900290120017 & 1900290120046

**ABSTRACT**

In many different fields, there is a high demand for storing information on a computer storage disk from the data available in printed or handwritten documents or images to later re-utilize this information by means of computers. One simple way to store information in a computer system from these printed documents could be first to scan the documents and then store them as image files. But to re-utilize this information, it would be very difficult to read or query text or other information from these image files. Therefore, a technique to automatically retrieve and store information, in particular text, from image files is needed. Optical character recognition is an active research area that attempts to develop a computer system that automatically extracts and processes text from images. The objective of OCR is to achieve modification or conversion of any form of text or text-containing documents such as handwritten text, and printed or scanned text images, into an editable digital format for deeper and further processing. Therefore, OCR enables a machine to automatically recognize text in such documents. Some major challenges need to be recognized and handled to achieve successful automation. The font characteristics of the characters in paper documents and the quality of images are only some of the recent challenges. Due to these challenges, characters sometimes may not be recognized correctly by the computer system. To overcome these challenges, I will be using a tesseract engine that reads words automatically and approximately accurately words. The output of this project will be that it will convert image files data to excel files and compare two excel files.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
|  | Page No. |
| TITLE PAGE .................................................................................................................... | 1 |
| DECLARATION .............................................................................................................. | 2 |
| CERTIFICATE …........................................................................................................... | 3 |
| ACKNOWLEDGEMENT.................................................................................................. | 4 |
| ABSTRACT...................................................................................................................... | 5 |
|  |  |
| CHAPTER 1 INTRODUCTION | 7 |
| 1.1.          Introduction ……………………................................................... |  |
| 1.2 Problem Statement.……………………...................................... |  |
| 1.2.          Objective………………………………………………………… |  |
| 1.3.          Scope……………………………………………………………. |  |
| CHAPTER 2 LITERATURE REVIEW…………………………………………….... | 8 |
| CHAPTER 3 PROPOSED METHODOLOGY …………………………………........ | 9 |
| 3.1 Flowchart |  |
| 3.2 Algorithm Proposed | 10 |
| CHAPTER 4 TECHNOLOGY USED ……….……………………….………………. | 12 |
| CHAPTER 5 DIAGRAMS …........................................................................................... | 12 |
| CHAPTER 6 CONCLUSION …....................................................................................... | 13 |
| REFERENCES….............................................................................................................. | 14 |

**Chapter1: INTRODUCTION**

Optical character recognition (OCR) is the electronic identification and digital encoding of typed or printed text using an optical scanner and specialized software. Using OCR software allows a computer to read static images of text and convert them into editable, searchable data. OCR typically involves three steps: opening and/or scanning a document in the OCR software, recognizing the document in the OCR software, and then saving the OCR-produced document in a format of your choosing.

OCR can be used for a variety of applications. In academic settings, it is oftentimes useful for text and/or data mining projects, as well as textual comparisons. OCR is also an important tool for creating accessible documents, especially PDFs, for blind and visually impaired persons.

* 1. **PROBLEM STATEMENT:**
* Many times, we need to write text from some hard paper like a magazine, newspaper, or some images. They could all involve you spending hours retyping manually and correcting typos. Or you could take a more modern approach and convert all of them into a digital format with fully editable text in a matter of minutes.
* Using OCR, we can capture the image of that page and give that to the OCR engine, OCR will convert the text from that image.
* It can be used at airports, banks, Railways, etc. as this project will convert image data to MS-excel files.

1.2 **OBJECTIVES:**

The research objective aims to create a tool that, when supplied with an input image, will be able to extract alphabets, digits, and symbols from it and convert it to an MS Excel file. The process will be easier to implement on printed data because that is easier to analyze, but the same system can be built for handwritten notes as well.

This tool can also be used to compare the difference between two Excel files.

* 1. **FUTURE SCOPE:**
* In the future this system can be expanded further: -Can Deploy on serve.
* System/website can also be used to store excel data in a database so the user can use his previous excel file. It satisfies the user requirement, easy to operate, and expandable user-friendly website; it is easy to understand by the user and operator

**Chapter 2: LITERATURE REVIEW:**

* 1. **A Detailed Analysis of Optical Character Recognition Technology**
* One simple way to store information on the computer system from these paper documents could be to first scan the documents and then store them as image files.
* The font characteristics of the characters in paper documents and the quality of images are only some of the recent challenges.
* We discuss major challenges and main phases of optical character recognition such as pre-processing, segmentation, normalization, feature extraction, classification, and post-processing in detail which needs to be considered during the implementation of any application related to the OCR and in the last section of our paper some OCR applications.
* In this paper, we have proposed an organization of these methods, algorithms, and techniques.
* Firstly, we discussed the major challenges of OCR, then we discussed in detail the main important phases, architecture proposed algorithms, and techniques of OCR, we highlight that for designing any application related to the OCR, one must pay great attention to each phase to obtain high accurate character recognition rate, but still, we cannot propose comprehensive algorithms for each phase because it depends upon datasets, application specifics, and parameter specifics.
* Although the state-of-the-art OCR enables text recognition with high accuracy, we think that there could be many more practical applications of OCR.
  1. **Survey on Character Recognition using OCR Techniques**
* Optical Character Recognition is a process by which we convert the printed document or scanned page to ASCII character that a computer can recognize.
* A typical handwriting recognition system consists of several steps, namely: preprocessing, segmentation, feature extraction, and classification, and several types of decision methods, including statistical methods, neural networks, and structural matching.
* In this paper, we will discuss how artificial neural network genetic algorithms and fuzzy logic can be used in optical character recognition for the use of character recognition.
* Several techniques that are used for optical character recognition have been discussed which use correlation and neural networks.
* The paper presents a brief survey of the applications in various fields along with experimentation in a few selected fields.
* The proposed method is extremely efficient to extract all kinds of bimodal images including blur and illumination.
  1. **Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic** **Literature Review (SLR)**
* An OCR system depends mainly on the extraction of features and the discrimination/classification of these features.
* Statistical data from selected studies are presented in Section III.
* It also enhances the consistency of the review and reduces the researchers' biases. This is because researchers must present a search strategy and the criteria for the inclusion or exclusion of any study in the review.
* Therefore, considering the quality of research studies, we believe that this systematic review can be used as a reference to find the latest trends and to highlight research directions for further studies in the domain of handwritten OCR.
* We believe that application areas of handwritten OCRs will further increase in the coming years. This is to be noted that this number of studies only includes research articles that are related to our research questions.
* A total number of selected studies are 176, and out of these 172 studies, the English language has the highest contribution of 53 studies in the domain of handwritten character recognition, 44 studies related to the Arabic language, 37 studies are on the Indian scripts, 23 on the Chinese language, 118 on the Urdu language, while 14 studies were conducted on the Persian language.
* We explored that some techniques perform better on one script than on another, e.g., the multilayer perceptron classifier gave better accuracy on Devanagari, and Bangla numerals, but gave average results for other languages.
* It was also observed that researchers are increasingly using Convolutional Neural Networks for the recognition of handwritten and machine-printed characters. This is since CNN-based architectures are well suited for recognition tasks where input is an image.
* Published research studies have proposed various systems for OCR but one aspect that needs to improve is the commercialization of research.
  1. **A Complete Optical Character Recognition Methodology for Historical**

**Documents**

* In and holistic word recognition approach for historical documents are presented based on scalar and profile-based features and on matching word contours respectively.
* In this paper, an offline recognition system for either machine-printed or handwritten historical documents is presented.
* It consists of a pre-processing stage where documents are converted into binary images, a top-down segmentation technique that extracts the characters, the creation of a database by the extracted characters, and a recognition stage where the database is used for converting any document into the text file.
* One set of 10 printed documents TrainSet1 and one set of 10 handwritten documents TrainSet2.
* To evaluate the performance of the segmentation procedure described in Section 3 for text line detection and word segmentation, we manually marked and extracted the ground truth on these train sets.
* From these training sets a typewritten database TWDB and a handwritten database HWDB were created respectively following the methodology discussed in Section 4.
* show the text files in ASCII format for a printed and handwritten document using the TWDB and the HWDB databases.
* Finally, 5 printed and 5 handwritten documents, different from the ones used in the training phase, formed two test sets.
* In this paper, a complete OCR methodology that assists the recognition of historical documents is presented. This methodology can be applied to either machine-printed or handwritten documents.
  1. **Identifying Text in Images Using OCR Testing**
* There are many applications in which the automatic detection and recognition of text embedded in images is useful.
* To identify the text in each image and extract the text in a clear background.
* It is expected that some strokes only cover fragments of the corresponding characters. This experiment demonstrates the performance Note that only the text which is horizontally aligned is counted.
* After the text detection and clean-up process, the text will be printed in a clear background. Thus, using the OCR tool text is retrieved successfully and displayed in a notepad.
  1. **An overview of Tesseract OCR Engine**
* Optical character recognition is the mechanical or electronic conversion of images of typed, handwritten, or printed text into machine-encoded text.
* It was further developed at HP between 1984 and to1994.
* In late 2005, HP released Tesseract for open source and is now available at Google.
* A more font-sensitive adaptive classifier that is trained by the output of the static classifier is used to obtain greater discrimination within each document, where the number of fonts is limited.
* Tesseract uses the same features and classifier as the static classifier for Adaptive classification.
* Since the adaptive classifier learns during the first run, it can only make significantly less contribution near the top of the page if deployed during the first run.

1. **METHODOLOGY:**

* Forms containing character images can be scanned through the scanner and then the recognition engine of the OCR system interprets the images and turn images of handwritten or printed characters into ASCII data (machine-readable characters).
* Therefore, OCR allows users to quickly automate data capture from forms, eliminate keystrokes to reduce data entry costs, and still maintain the high level of accuracy required in forms processing applications.

3.1 **FLOWCHART & ALGORITHM USED:**

TESSERACT OCR Engine will be used in this project. Tesseract is an open-source optical character recognition engine. Tesseract began as a Ph.D. research project in HP Labs, Bristol. It was further developed at HP between 1984 and to1994. It was modified and improved in 1995 with greater accuracy. In late 2005, HP released Tesseract for open source which is now available at Google.

**The architecture of the Tesseract OCR engine.**

1. **TECH STACK USED:**

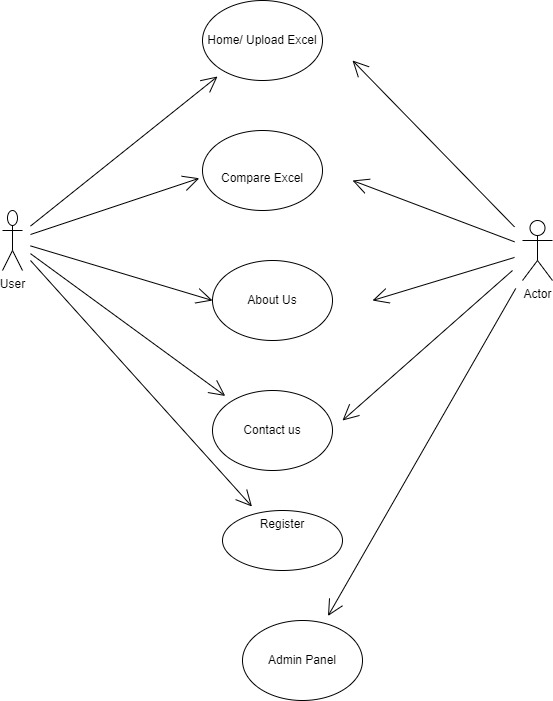
**FRONT END TOOL:** HTML, CSS, JAVASCRIPT.

**BACKEND TOOL:** Python, Django, Machine Learning & some other python library

**FRAMEWORK/LIBRARY:** Font Awesome, Google Font, Bootstrap

**DATABASE:** SQLite3

**5.** **ER DIAGRAM/USE CASE DIAGRAM/DFD**





**Block diagram of Word Recogniser.**

6. **CONCLUSION:**

Numerous algorithms, methods, and techniques have been proposed for optical character recognition in scene imagery, yet there are not enough literature surveys in this field. In this paper, we have proposed an organization of these methods, algorithms, and techniques. It is hoped that this comprehensive survey will provide insight into the concepts involved, and perhaps provoke further advances in the area.

This project converts image data, printed document data, passbook images data, etc. to an excel file, and it will also compare and tell the difference between two to excel files.

**REFERENCES:**

[1] Hamad, Karez Abdulwahhab, and K. A. Y. A. Mehmet. "A detailed analysis of optical character recognition technology." *International Journal of Applied Mathematics Electronics and Computers* Special Issue-1 (2016): 244-249.

*[2]* Tyagi, Karishma, and V. Rastogi. "Survey on Character Recognition using OCR Techniques." *International Journal of Engineering Research and Technology* 3.2 (2014): 2533-2536.

*[3]* Memon, Jamshed, et al. "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)." *IEEE Access* 8 (2020): 142642-142668.

[4*]* Vamvakas, Georgios, et al. "A complete optical character recognition methodology for historical documents." *2008 The Eighth IAPR International Workshop on Document Analysis Systems*. IEEE, 2008.*,*

[5] Ponnarasi, M. Jeslin Benita, and PG STUDENT M. TECH. "Identifying Text in Images Using OCR Testing."

[6] Smith, Ray. "An overview of the Tesseract OCR engine." *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE, 2007.