**Review-1 Report**

**Improving Text Extraction Accuracy with Image Preprocessing**

Submitted in partial fulfillment of the requirements for the award of degree

**BACHELOR OF ENGINEERING**

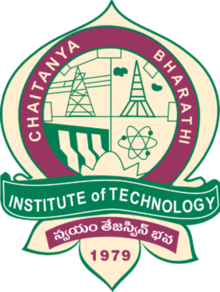
**in**

**COMPUTER SCIENCE AND ENGINEERING**

**by**

**KOMALI BEERAM (160118733067)**

**SOUMYA VEMURI (160118733081)**



**Department of Computer Science and Engineering,**

**Chaitanya Bharathi Institute of Technology (Autonomous),**

**(Affiliated to Osmania University, Hyderabad)**

**Hyderabad, TELANGANA (INDIA) –500 075**

**[2020-2021]**

**CERTIFICATE**

This is to certify that the project titled “**Improving Text Extraction Accuracy with Image Preprocessing**” is the bonafide work carried out by **KOMALI BEERAM (160118733067) and SOUMYA VEMURI (160118733081)**, students of B.E.(CSE) of Chaitanya Bharathi Institute of Technology(A), Hyderabad, affiliated to Osmania University, Hyderabad, Telangana(India) during the academic year 2020-2021, submitted in partial fulfillment of the requirements for the award of the degree in **Bachelor of Engineering** (**Computer Science and Engineering** ) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

|  |  |
| --- | --- |
| **Supervisor**  Dr. T. Sridevi | **Head, CSE Dept.**  Dr. Y Ramadevi |
| **Place:** Hyderabad, Telangana, India  **Date:** 25th October, 2021 | |
|  | |

**DECLARATION**

We hereby declare that the project entitled “**Improving Text Extraction Accuracy with Image Preprocessing**” submitted for the B.E (CSE) degree is our original work and the project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

**KOMALI BEERAM (160118733067)**

**SOUMYA VEMURI (160118733081)**

**Place:** Hyderabad, Telangana, India

**Date:** 25th October, 2021

**ABSTRACT**

Digitization allows us to immortalize a physical entity by creating a digital representation of it on our devices. It saves us time in manually sifting through physical storage units such as albums and notebooks and provides us with programs to manage and secure our data. We often take images of Receipts or Invoices, Identity Cards, and nutritional labels to save a copy of their details. This can be taken a step further by automating the process of information extraction and documentation.

Advancements in computer vision have provided us with the expertise to create tools for text detection and extraction. But it is still an ongoing challenge because documents with unstructured layouts, poor image quality, and noise around the text yield very low accuracy in text extraction results. Conquering this challenge would require the image to be highly enhanced through pre-processing techniques such as Brightness Correction, Contour Detection, Skewness Correction, Morphology, and Binarization. A mechanism made from the best combination of image pre-processing techniques prior to text extraction can improve text accuracy to a large extent.

*Keywords: Computer Vision, Image pre-processing, Text Extraction*

|  |  |  |
| --- | --- | --- |
|  | Title Page | i |
|  | Certificate of the Guide | ii |
|  | Declaration of the Student | iii |
|  | Abstract | iv |
|  |  |  |
| **1.** | **INTRODUCTION** | **1** |
|  | * 1. Problem Definition including the significance and objective | 1 |
|  | * 1. Methodologies   2. Outline of the results   3. Scope of the project   4. Organization of the report | 1  2  2  2 |
| **2.** | **REFERENCES** | **3** |

**Table of Contents**

1. **INTRODUCTION**
   1. **Problem Statement**

Historical OCR engines have their accuracy lying between 70-80% for a high-quality image at page level. That means in a page of 100 words 70-80 words are accurate. This will lead to significant inaccuracies if used on a large volume of sensitive documents.

Through this project we would like to increase the accuracy of results to at least 80-90% by implementing an ideal combination of image preprocessing techniques prior to text extraction.

* 1. **Methodology**

1. **Document Scan and File Upload**

*User scans their document and uploads it to the system interface. Documents include invoices, receipts, nutrition labels, book covers etc.*

1. **Image Preprocessing Pipeline**

*Scanned documents undergo the best combination of image preprocessing techniques for maximum image text enhancement. Image Preprocessing Techniques include*

* 1. *Brightness Correction*
  2. *Contour Detection*
  3. *Scaling of Image*
  4. *Skewness Correction*
  5. *Noise Removal*
  6. *Binarization*

1. **Text Extraction**

*The image goes through an Optical Character Recognition (OCR) engine that’s been built to recognize printed text in paper documents, handwritten characters, and text elements in the image.*

1. **Information Extraction**

*Information Extraction involves extracting meaningful information from raw text data into a structured format.*

* 1. **Outline of Result**

For performance of recognition, we will simulate the proposed text recognition system to various characters in English language and calculate the recognition rate or accuracy which is expected to rise from approximately 70% to 90%.

* 1. **Scope of Project**

The scope of our project on a grid infrastructure is to provide an efficient and enhanced software tool for the users to perform document image analysis, document processing by reading and recognizing the characters in research, academic, governmental, and business organizations that are having large pool of documented, scanned images. Irrespective of the size of documents and the type of characters in documents, the product is recognizing them, searching them, and processing them faster according to the needs of the environment.

* 1. **Organization of Report**

The report is divided into two parts. Each part deals with different aspects of our project. Each part has various chapters explaining in detail.

* Part 1: Introduction

This part summarizes about the project in brief. It includes the problem statement, methodology, and outline of our results and the scope of the project.

* Part 2: References

This part has a consolidated list of papers that we referred to develop the project.

Although the document may be read from front to back for a complete understanding of the project, it was written in sections and hence can be read as such. For an overview of the document and the project itself, refer to Introduction.

**2. REFERENCES**

[1]. HaoranLi and Yang Zhang (2021) “Event Info Extraction from Flyers”

[2]. Anchal Garg and Rishabh Mital (2020) “Text Extraction using OCR: A Systematic Review”, ICIRCA-2020, IEEE Xplore Part Number: CFP20N67-ART

[3]. Costin-Anton Boiangiu, Dan Sporici and Elena Cusnir (2020) “Improving the Accuracy of Tesseract 7.0 OCR Engine Using Convolution-Based Preprocessing”

[4]. Kritika Handa, Mahika Sharma, Rishika Jaiswal and Sanjeev Kumar (2020) “Improve OCR Accuracy with Advanced Image Preprocessing using Machine Learning with Python”, IJITEE, ISSN: 2278-3075, Volume-9, Issue-7