High Level Design

Image to Text Converter Web Application

Revision Number: 1.0

Last date of revision: 23/06/2021

1.1 Document Version Control

Change Record

Date	Version	Author	Description
23/06/2021	1.0	M.Akhilesh	Initial HLD-V1.0

Contents

1. Introduction
Abstract **3**
1.1. Document Version Control……..1
1.2. Why this High Level Design Document?.....4
1.3. Scope……....4
1.4. Definitions……4
2. General Description
2.1. Product Perspective……..5
2.2. Problem of statement……5
2.3. Proposed solution……5
3. Technical Requirements
3.1. Tools Used……6
3.2. Application Architecture
3.2.1. Main Design Features……..7
3.2.2. Database design……...7
3.2.3. Index page……8
3.2.4. Result page……..8
3.2.5. Screen Shot……8
4. General Constraints
4.1. Assumptions……..9
4.2.Event handling ……9
4.2.1.Initial step-by-step description……9
4.3.Error Handling……..9
4.4. Performance……10
4.5. Reusability……10
4.6. Application compatibility……...10
4.7. Resource utilization……11
4.8. Deployment……11
5**. Conclsion** **11**

6.References **11**

Abstract

Optical Character Recognition (OCR) is the process of extracting text from an image. The main purpose of an OCR is to make editable documents from existing paper documents or image files. Significant number of algorithms is required to develop an OCR and basically it works in two phases such as character and word detection. In case of a more sophisticated approach, an OCR also works on sentence detection to preserve a document's structure. It has been found that researchers put lots of efforts for developing a Bengali OCR but none of them is completely error free. To take this issue in consideration, the latest 3.03 version of Tesseract OCR engine for Windows operating system is used to develop an OCR for Bengali language. Moreover, 18110 characters and 2617 words are used to make the OCR's library. In this research, `Solaimanlipi' font and 200 input files are used to test the accuracy of OCR. It is found that for clean image files, the accuracy of the software is as high as 97.56%. It is to be noted that accuracy is measured as the percentage of correct characters and words.

1. Introduction

1.1. Why this High Level Design Document?

The purpose of this High Level Design (HLD) Document is to add the necessary detail tothe current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a referencemanual for how the modules interact at a high level.

The HLD will:

- present all of the design aspects and define them in detail
- describe the user interface being implemented
- describe the hardware and software interfaces
- describe the performance requirements
- include design features and the architecture of the project
- list and describe the non-functional attributes like:
    - security
    - reliability
    - maintainability
    - portability
    - reusability
    - application compatibility
    - resource utilization
    - serviceability

1.2. Scope:

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.3. Definitions:

- OCR--Optical Character Recognition
- Database –Collection of all the information monitored by the system.
- IDE – Integrated development environment
- AWS – Amazon Web Services.

2. General Description

2.1. Product Perspective

Optical Character Recognition, or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data.
Imagine you've got a paper document - for example, magazine article, brochure, or PDF contract your partner sent to you by email. Obviously, a scanner is not enough to make this information available for editing, say in Microsoft Word. All a scanner can do is create an image or a snapshot of the document that is nothing more than a collection of black and white or color dots, known as a raster image. In order to extract and repurpose data from scanned documents, camera images or image-only PDFs, you need an OCR software that would single out letters on the image, put them into words and then, words into sentences, thus enabling you to access and edit the content of the original document.

2.2.Problem of statement:

To create a solution for automating data extraction from printed or written text from a scanned document or image file and then converting the text into a machine-readable form to be used for data processing like editing or searching

2.3 . Proposed solution:

A common application of OCR technology is the automated conversion of an image based PDF, TIFF or JPG into a text based machine-readable file. OCR-processed digital files, such as receipts, contracts, invoices, financial statements and more, can be:

- Searched from a large repository to find the correct document
- Viewed, with search capability within each document
- Edited, when corrections need to be made
- Repurposed, with extracted text sent to other systems

3. Technical Requirements

This document describes minimum requirements for OCR i.e Hardware Requirement PC Computers with minimum capacity: **Processor: Pentium 200 MHz RAM: 32 MB Disk: 4 GB Form** modules are designed to operate in a batch processing, run under LAN and PC based platforms and take full advantage of the graphical user interface and 32 bit processing power available with Windows 95.

3.1 Tools Used

Python programming and frameworks such as Numpy, Pillow, Flask has been used to build to entire model.

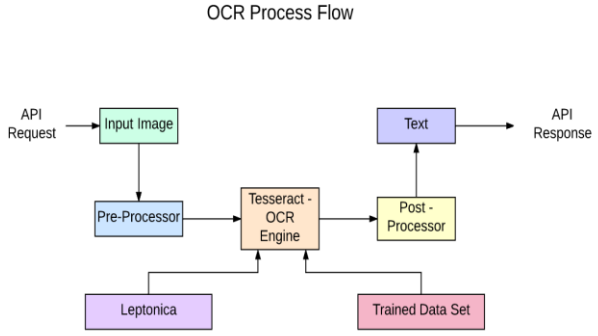
JavaScript

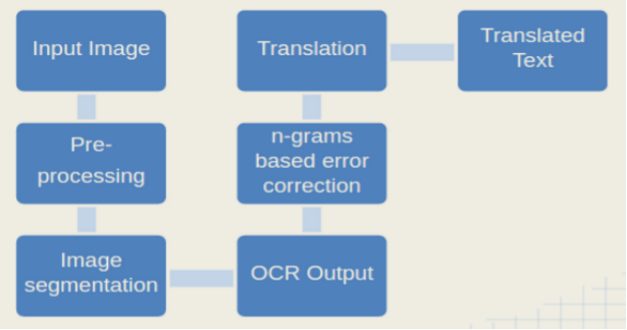
- PyCharm is used as IDE.
- AWS is used for deployment of the model.
- TabIeau/Power BI is used for dashboard creation.
- MySQL/MongoDB is used for retrieve and insert into database.
- Frontend is created using HTML/CSS/Javascript.
- Python Flask is used for backend development.
- GitHub is used for version control system.

3.2 Application Architecture:

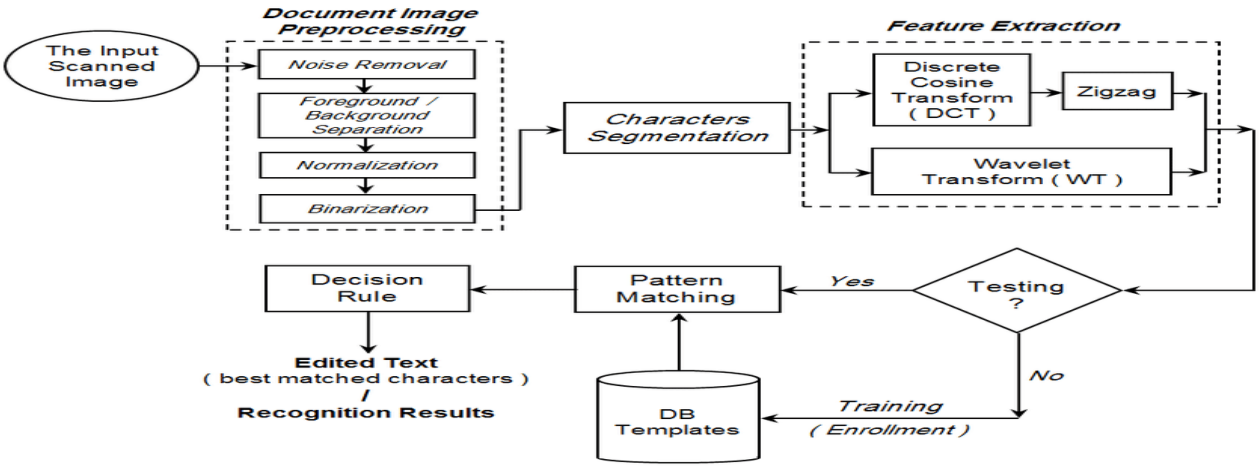
3.2.1.Main Design Features:

The main design features includes the user interfacedesign, external interface, the database, process relation. In order tomake these designs easier to understand, the design has been illustrated in attached diagrams (ER, Use Case, and Screen Shots).



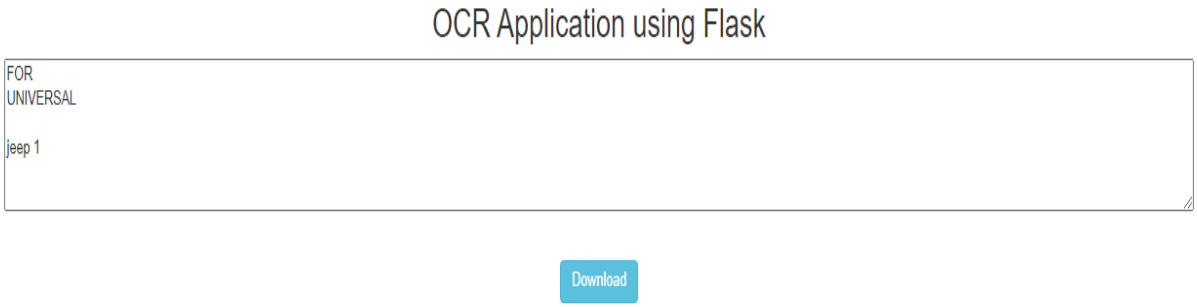


3.2.2.Database design:



3.2.3.Index page:

3.2.4.Result Page:



3.2.5.Screen Shot

Breakdown:

User:

- index.html – user have to choose a file to upload.
- Image display—Chosen image will be displayed at the center.
- Convert button—It is used to extract the text from image.

Result page

- Result.html—extracted text from image will be shown.
- Download button—Is used to download the text into ".txt" format

4.General Constraints:

OCR is often used to obtain text from image-only files for use in classifying them. However, there are several limitations of OCR that can result in inaccurate or missing text which makes text-based classification difficult

4.1.Assumptions:

Optical character Recognition (OCR) serves as a tool to detect information from natural images and transfer them into machine-coded texts, such as words, symbols and numbers. It is still a hot ongoing search area and some novel algorithms are publishing from time to time. It is pretty interesting and essential to recognize the characters in the image because it could help greatly in some certain area: auto plate number recognition, books and documents scanning, assistive technology for blind and visually impaired users , zip-code recognition needed for post offices and much more.

4.2.Event handling:

The system should log every event so that user will know what process is running internally.

4.2.1.Initial step-by-step description:

Application Log:

Any event logged by an application.

System Log:

Any event logged by the Operating System.

Security Log:

Any event that matters about the security of the system.

The reports will display the user's average usage up to the last time the system calculated it and their history.

4.3.Error Handling:

Should errors be encountered, an explanation will be displayed as to what went wrong. An error will be defined as anything that falls outside the normal and intended usage.

4.4.Performance:

Optical Character Recognition (OCR) is a technique, used to convert scanned image into editable text format. Many different types of Optical Character Recognition (OCR) tools are commercially available today; it is a useful and popular method for different types of applications. OCR can predict the accurate result depends on text pre-processing and segmentation algorithms. Image quality is one of the most important factors that improve quality of recognition in performing OCR tools. Images can be processed independently (.png, .jpg, and .gif files) or in multi-page PDF documents (.pdf). The primary objective of this work is to provide the overview of various Optical Character Recognition (OCR) tools and analyses of their performance by applying the two factors of OCR tool performance i.e. accuracy and error rate

4.5.Reusability:

The code written and the components used should have the ability to be reused with no problems. Should time allow, and detailed instructions are written on how to create this project, everything will be completely reusable to anyone.

4.6.Application compatibility:

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of thePython to ensure proper transfer of information.

4.7.Resource utilization:

When any task is performed, it will likely use all the processing power available until that function is finished.

4.8.Deployment:

5.Conclsion:

In conclusion, OCR is a very remarkable technology that holds a lot of potential. In this day and age, such tools are already quite advanced. However, Optical Character Recognition is going to look even better in the future.

6.References:

https://github.com/tesseract-ocr/tesseract/wiki/Downloads for Tesseract download.