Low Level Design

Image to Text Converter Web Application

Revision Number: 1.1

Last date of revision: 24/06/2021

1.0 Document Version Control:

Date	Version	Author	Description
24/06/2021	1.0	M.Akhilesh	First Draft
25/06/2021	1.1	M.Akhilesh	Added Workflow chart
25/06/2021	1.2	M.Akhilesh	Added user I/O flowchart
25/06/2021	1.3	M.Akhilesh	Added overview and updated userI/O flowchart
26/06/2021	1.4	M.Akhilesh	Restructure and reformat LLD

Contents

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 Low Level Design Document?

The purpose of this Low-Level Design (LLD) Document is to add the necessary detail to the 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 reference manual for how the modules interact at a high level.

The main objective of the project is to detect the text from scanned documents, camera images or image-only PDFs.

OCR can help:

- In business like Let's look at one mock scenario. You need to renew a vehicle registration with your state government such as Name, Address, vehicle identification number, vehicle model, make and year.
- Using OCR we can extract information from a written prescription can be sent over to a pharmacy via an app that might read the prescription using OCR, saving the patient and pharmacy time.

1.2. Scope:

The LLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The LLD uses non-technical to mildly-technical terms whichshould be understandable to the administrators of the system. This software system will be a Web application This system will be designed to extract text from scanned documents, camera images or image-only PDFs.

1.3 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

1.4 Risks:

Document specific risks that have been identified or that should be considered.

1.5 Out of Scope:

Delineate specific activities, capabilities, and items that are out of scope for the project.

2. Technical Specifications

2.0 Image pre-processing:

OCR software often pre-processes images to improve the chances of successful recognition. The aim of image pre-processing is an improvement of the actual image data. In this way, unwanted distortions are suppressed and specific image features are enhanced. These two processes are important.

- **Binarisation** – Convert an image from color or greyscale to black-and-white (called a "binary image" because there are two colors). The task of binarisation is performed as a simple way of separating the text (or any other desired image component) from the background.
- **Line removal** – Cleans up non-glyph boxes and lines Layout analysis or "zoning" – Identifies columns, paragraphs, captions, etc. as distinct blocks.
- **Script recognition** – In multilingual documents, the script may change at the level of the words and hence, identification of the script is necessary, before the right OCR can be invoked to handle the specific script.
- **Character isolation or "segmentation**" – For per-character OCR, multiple characters that are connected due to image artifacts must be separated; single characters that are broken into multiple pieces due to artifacts must be connected. Segmentation of fixe dpitch fonts is accomplished relatively simply by aligning the image to a uniform grid based on where vertical grid lines will least often intersect black areas. We have to Normalize aspect ratio and scale.

2.1 Character Recognition:

There are two basic types of core OCR algorithm, which may produce a ranked list of candidate characters.

- **Matrix matching**:

Involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching", "pattern_recognition", or "image_correlation". This relies on the input glyph being correctly isolated from the rest of the image, and on the stored glyph being in a similar font and at the same scale.

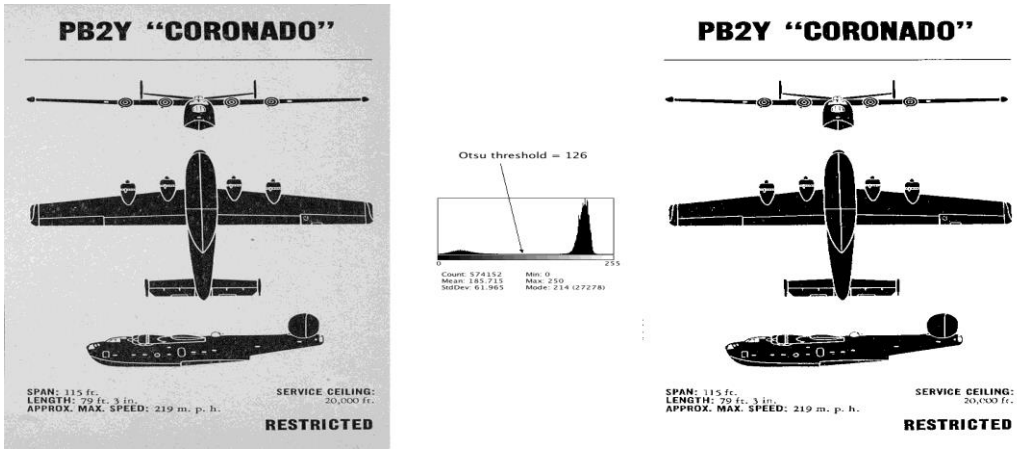
- **Feature extraction**:

Decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. The extraction features reduces the dimensionality of the representation and makes the recognition process computationally efficient. Software such as Cuneiform and Tesseract use a two-pass approach to character recognition.

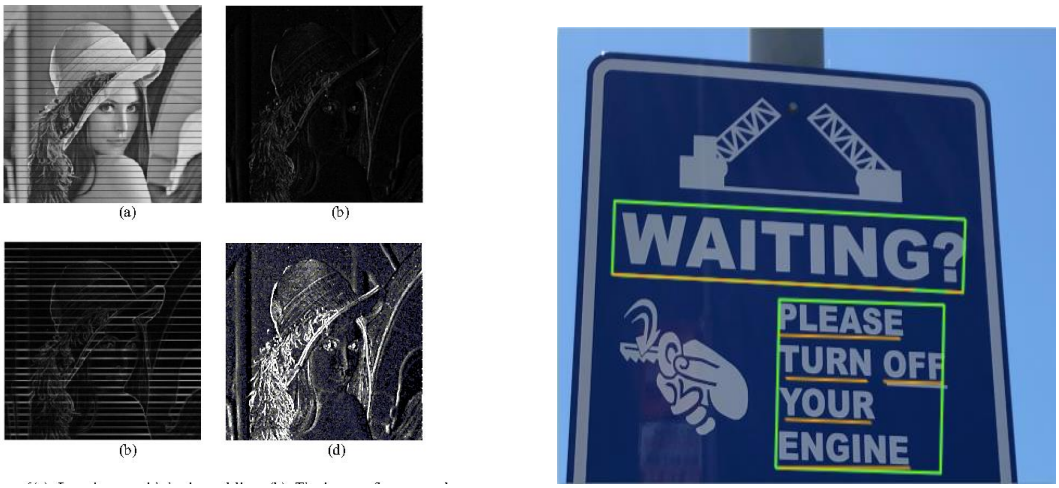
2.2 Post-processing:

OCR accuracy can be increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. Tesseract uses its dictionary to influence the character segmentation step, for improved accuracy.

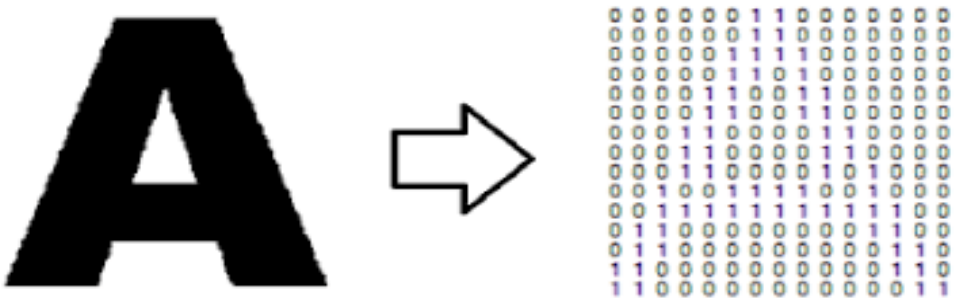
Binarisation :



Line removal and Script recognition:



Character isolation:



2.3 Input Schema:

Feature name	Data type	size	Null/Required
Image type	scanned documents, camera images or image-only PDFs.		Required

2.4 Loggings:

We should be able to log every activity done by the incidents.

- The System identifies at what step logging required
- The System should be able to log each and every system flow.
- Developers can choose logging methods. You can choose database logging/ File logging as well.
- System should not be hung even after using so many loggings. Logging justbecause we can easily debug issues so logging is mandatory to do.

2.5 Database:

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

- The User chooses the activity dataset.
- The User gives required information.
- The system stores each and every data given by the user or received on request to the database. Database you can choose your own choice whether MongoDB/ MySQL.

3 Deployment:

4 Technology Stack:

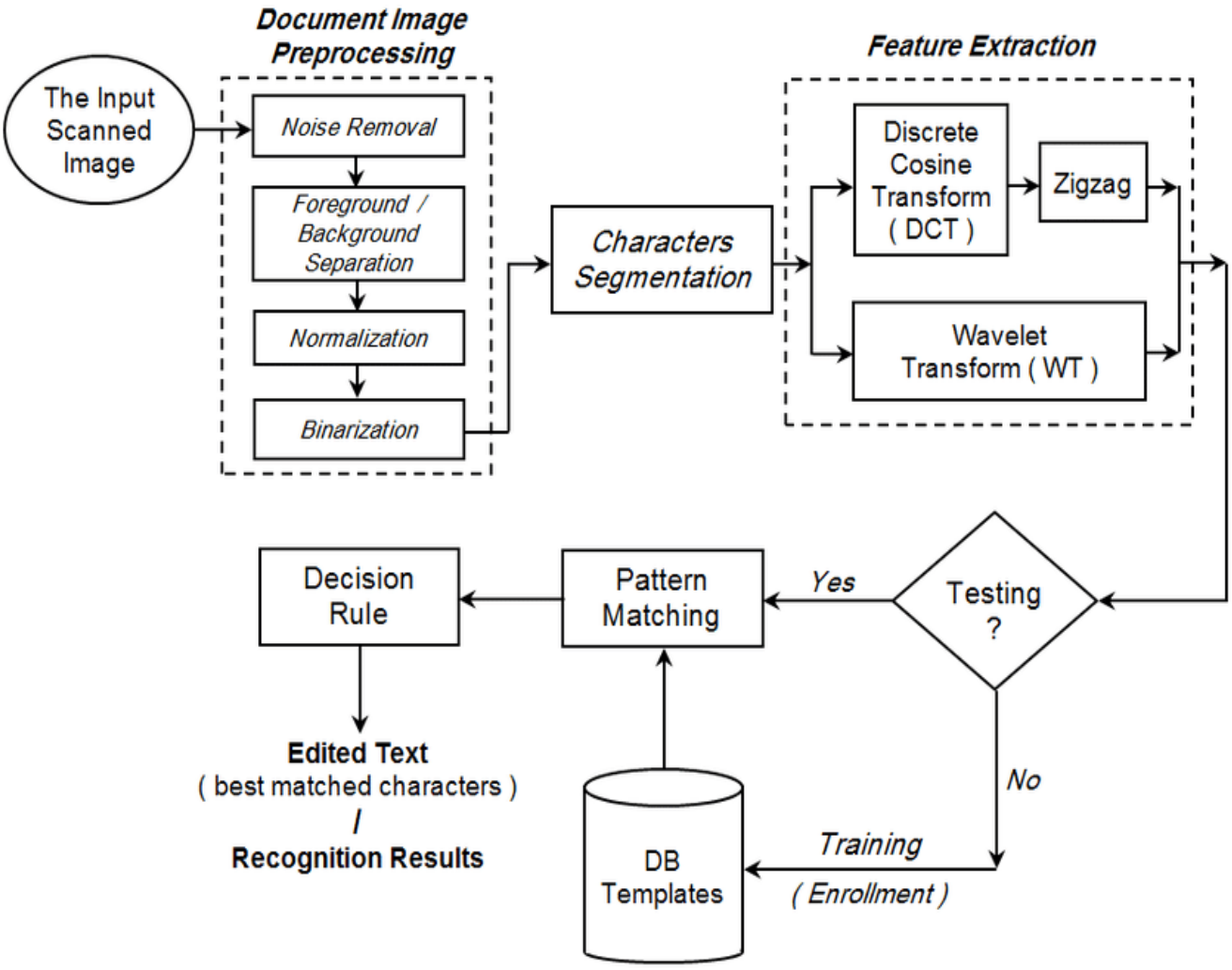
Front End	**HTML/CSS/JavaScript**
Backend	**Python Flask**
Database	**MongoDB/MySql**
Deployment	**AWS, Azure**
Dashboard	**Tableau/Power BI**
Project design platform	**Pycharm**
version control	**GitHub**

5 Proposed solution:

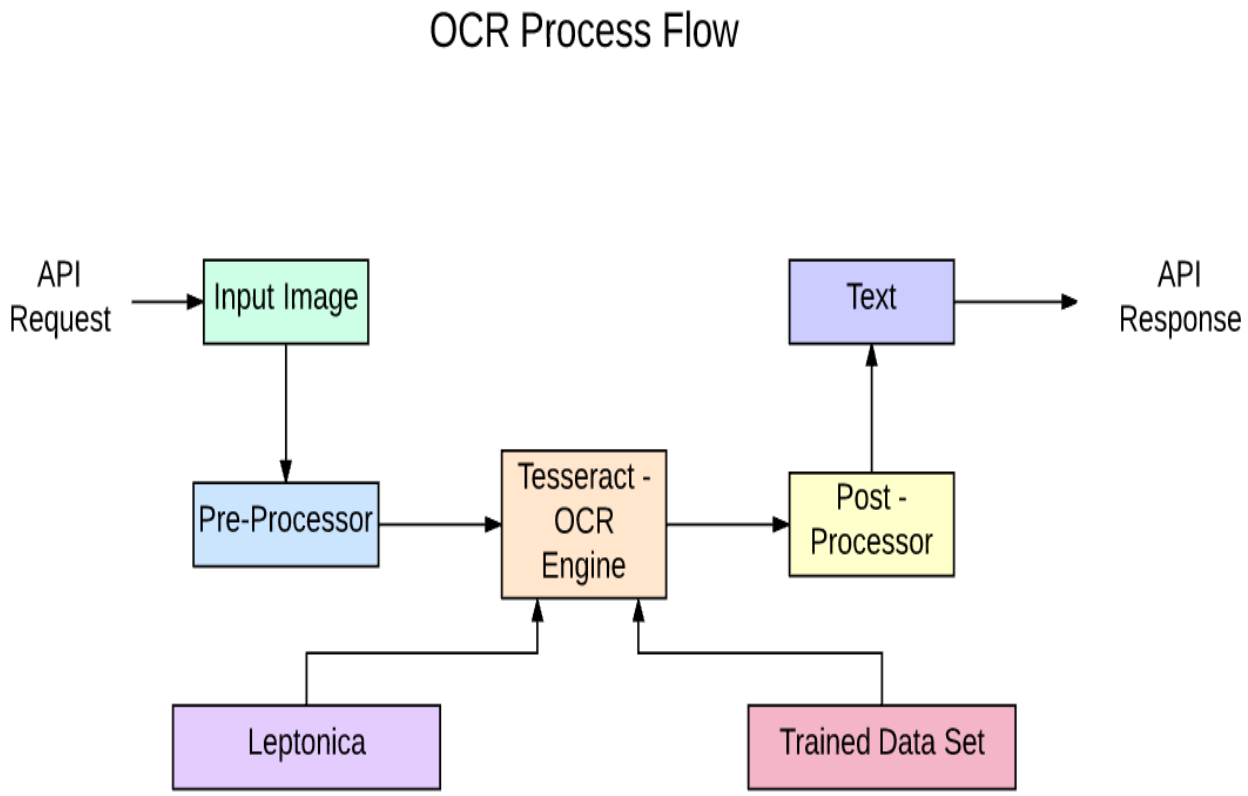
A common application of OCR technology is the automated conversion of an image based PDF, TIFF or JPG into a text based machine-readable file. OCR-processed digital files, such as receipts, contracts, invoices, financial statements and more, can be:

- 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

6 Model Workflow:



7 User I/O Workflow:



8 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.

9 Conclusion:

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.