Optical Character Recognition(OCR)

(Recognition of Handwritten Text)

**Introduction:**

The goal of this project is to use the existing OCR technology(Tesseract OCR) in the context of recognizing handwritten texts and improving the accuracy of the OCR Engine and also adding/improving language accuracy.

**What is Tesseract OCR:**

Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract printed text from images. It supports a wide variety of languages.Tesseract is compatible with many programming languages and frameworks through wrappers.The latest version of Tesseract includes a new neural network subsystem configured as a text line recognizer

**How it Works:**

To recognize an image containing a single character, we typically use a Convolution Neural Network (CNN). Text of arbitrary length is a sequence of characters, and such problems are solved using RNN(Recurrent Neural Network) and LSTM(Long Short-Term Memory) is a popular form of RNN.The input image is processed in boxes (rectangle) line by line feeding into the LSTM model and giving output.

**Existing System:**

While Existing OCR’s and the Tesseract engine does capture and recognizes the characters well it is still far from accurate and that is just for the printed texts and it is still far from accurately recognizing handwritten texts or characters. Sometime the output will have very poor quality if the input images are not preprocessed correctly and especially screenshots.

**Problem Definition:**

Errors include misreading letters, skipping over unreadable letters, or combining text from adjacent columns or image captions. While many factors affect the performance of OCR, the number of errors depends on the quality and form of the text, including the font used. However, even with high-quality documents, OCR can make mistakes because there are a variety of document formats, fonts, and styles for each character. The limitations that prevent OCR from reaching 100% accuracy. And when it comes to Handwritten Characters it is extremely hard to get the desired accuracy because each and every individual has different style of handwriting and cursive text which can be an intricate process to implement.

**Proposed System:**

The scope of this system is to improve the accuracy of the recognition of the Handwritten Character recognition.The research on handwriting recognition also leverages the dynamic motion created during the handwriting process to identify characters. While the main problem with handwriting recognition is the variety of character styles, OCR’s accuracy in this area is constantly and slowly improving. This System will emphasize on improving that by using the existing tools provided and by creating large number of data files and storing them into a trained data file which will be implemented to recognize any particular piece of data which will be taken as an input.