

Optical Character Recognition

Industrial Mentor
Mr. Mahindra Rajan

Mentor
Dr. R. Rekha

Akil Rajendran - 19I204

Gokulnath P K - 19I217

Prithik Kumar R - 19I243

Sanjay Krishnan R - 19I254

Shevanth R - 20I437

Aathithya S - 20E102


CONTENTS

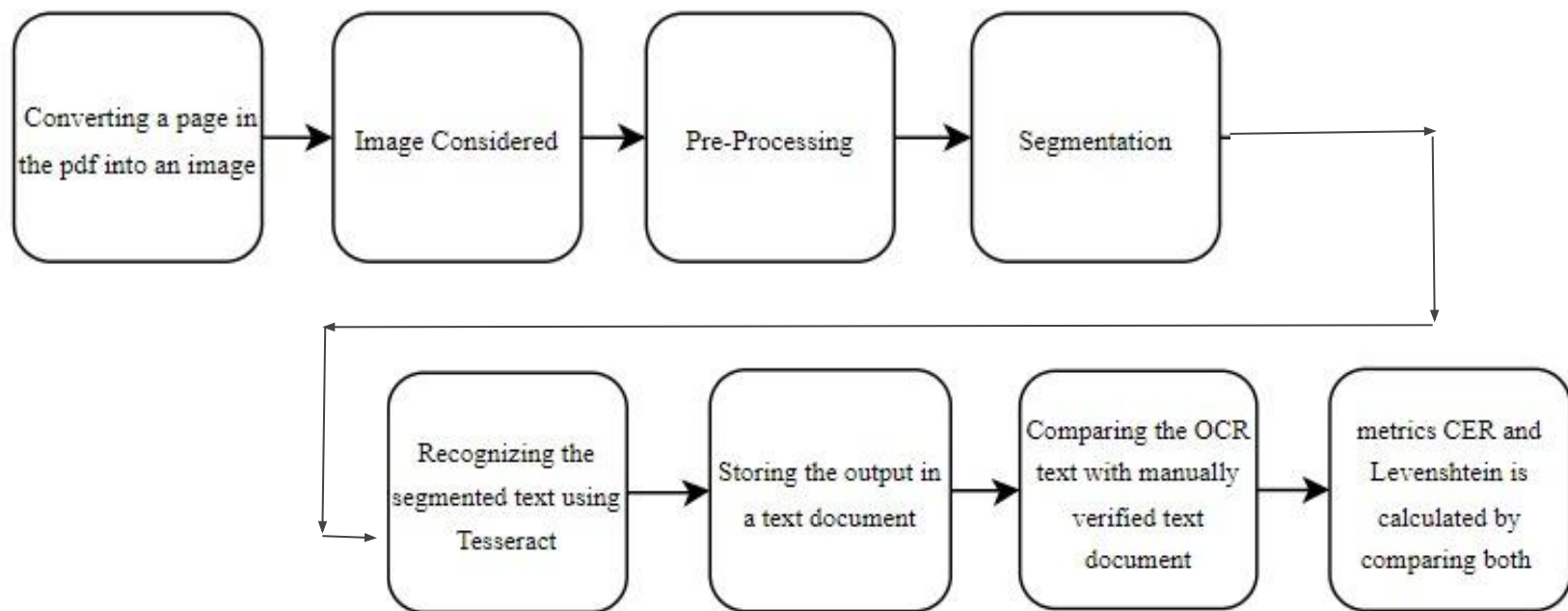
- Problem Statement
- Results till now
- Challenges faced
- Observation
- Upcoming plan

Problem Statement

Our problem statement is to build a Optical Character Recognition(OCR) system to recognize and digitize ancient Tamil books. To digitize the ancient tamil books we have used Google's Tesseract OCR system. The problem with Tesseract is that when documents have a lot of colours in it, the Tesseract fails to perform well. Hence our role is to fine tune the Tesseract OCR system.

Dataset taken : Abithana Chintamani (The Encyclopedia of Tamil Literature)





Work Done

- Tesseract OCR is applied on the scanned images that outputs text with an error rate of just ~4%.
- Evaluation metrics are used to determine the performance measure of the OCR system used.

Tesseract - <https://tinyurl.com/OCR-review2>



Challenges faced

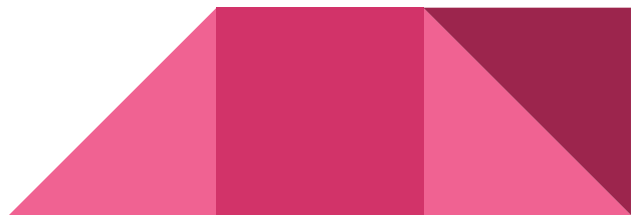
1. The images are read line by line and not column-wise - PSM values.
 - a. Tesseract has 14 Page Segmentation Modes in built.



```
$ tesseract --help-psm
```

Page segmentation modes:

- 0 Orientation and script detection (OSD) only.
- 1 Automatic page segmentation with OSD.
- 2 Automatic page segmentation, but no OSD, or OCR. (not implemented)
- 3 Fully automatic page segmentation, but no OSD. (Default)
- 4 Assume a single column of text of variable sizes.
- 5 Assume a single uniform block of vertically aligned text.
- 6 Assume a single uniform block of text.
- 7 Treat the image as a single text line.
- 8 Treat the image as a single word.
- 9 Treat the image as a single word *in* a circle.
- 10 Treat the image as a single character.
- 11 Sparse text. Find as much text as possible *in* no particular order.
- 12 Sparse text with OSD.
- 13 Raw line. Treat the image as a single text line, bypassing hacks that are Tesseract-specific.




Challenges faced

- Some words are not printed completely.
- Skewness.
- Some handwritten and overlapping letters are found.
- Better the image quality, better the images.



Upcoming Plans

- Updating the work done by applying the OCR implemented, for larger pdf files. (output format)
 - Considering the Mean value of the accuracy of each page in the pdf.
 - To divide the text area into blocks and try out few other page segmentation methods for better results.
 - Trying to clean the data more and arrive at a decent WER (word error rate) score.
- 

Thank you !