

Optical Character Recognition

Mentor

Dr. R. Rekha

Akil Rajendran - 19I204
Sanjay Krishnan R - 19I254

Gokulnath P K - 19I217
Shevannth R - 20I437

Prithik Kumar R - 19I243
Aathithya S - 20E102


CONTENTS

- Problem Statement
- Proposed Approach
- Work done
- Technical Challenges
- Upcoming plan

Problem Statement

Literature has been one of our important resources to understand history and culture of ancient civilizations. But most of them are in paper form which makes it hard for maintenance. Hence our problem statement is to build a Optical Character Recognition (OCR) system to recognize and digitize ancient Tamil books.

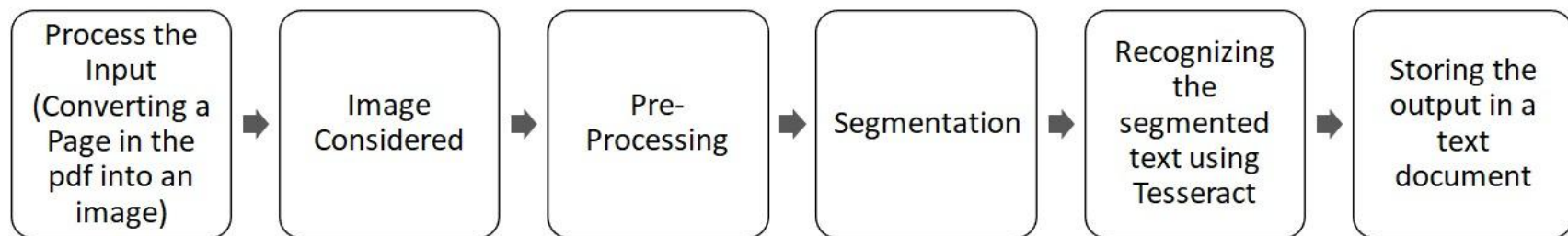
Dataset taken : Abithana Chintamani (The Encyclopedia of Tamil Literature)



Proposed Approach

1. Considering one page at a time.
2. Converting the page into a binary image.
3. Segmenting the letters and slicing them into letter blocks.
4. Recognizing the letters using fine tuned Tesseract Model.
5. Storing the recognized text in a document.





Work Done

<https://colab.research.google.com/drive/1vwRBLCRF0lhFtahPaMI4MAJmRg-b2ZrI?authuser=1>



Challenges faced

1. The Dataset considered has highly noisy content.
2. Some words are not printed completely.
3. The Tamil used in the dataset is largely varying from regular tamil text.
4. The content is split by a vertical line.
5. Some handwritten and overlapping letters are found.



Upcoming Plans

- Fine tuning the Tesseract Model for this purpose.
- Updating the work done, from a single image into a pdf file.

