09 - Preliminary Project Report

Koidala Surya Prakash
ee18btech11026@iith.ac.in

VVS Pavan Kumar
ma18btech11010@iith.ac.in

Veggalam Sai Sudeep
ee18btech11045@iith.ac.in

Vedala Sai Ashok
ee18btech11044@iith.ac.in

Abstract

Lawyers decides the fitment of a collateral by comparing certain fields in the document. This procedure can be laborious and can be automated by an intelligent system which knows what fields to look for in a document and prepare a summary for each document thus making it easier for a lawyer for verification . We plan to use our image processing and deep learning skill set to build an end-to-end solution that can cater our needs.

1. Introduction

- Even to this date the process of collateral verification in banking for sanctioning loans requires the lawyer to manually verify the chain of transactions along with any issues related to the collateral and prepare a Title Search Report in which the lawyer provides the legal opinion to the bank on whether or not to sanction the loan.

- While preparing a Title Search Report the lawyer manually collects all the sale deeds related to the history of the collateral (for the past 15 to 30 years). Based on these he/she approves whether a collateral can be mortgaged or not, while verifying and evaluating the trustworthiness(by auditing finances) and thus approves a final fitment certification.

- The process of manually verifying all the sale deeds related to the chain of transfers involves reviewing if the document has any pending litigations; ensuring that the document is properly registered and the previous owner has waived all other claims and ensure the details of the owner appearing in a document matches in all other documents as well. This process can be quite tiresome but can be automated with image and text processing.

2. Problem Statement

We plan to assist the lawyer in making a Title Search Report by generating a summary for each and every sale deed. We mainly identified three sub-problems to generate a summary report from sale deed provided in PDF format.

One of the key challenges to be faced is that the sale deeds do not have a well structured format , thus information will be represented in different ways across different documents . Thus we need to ensure that our solution should work for any such document.

- Preprocessing (involves denoising ,removing unwanted artifacts such as logos, watermarks, signatures and stamps etc...)

- OCR (to convert the denoised PDF into .docx or .txt file)

- Named Entity Recognition (to extract relevant information* from OCR output and create summary for lawyer)

  *Relevant information involves current owner's name, registration ID, date of sale or purchase etc...

3. Literature Review

This problem is being solved in association with a legal tech startup : **Lending Katalyst** which is trying to assist lawyers by developing collateral valuation solutions.

An extensive reading has been done to understand OCR tools which includes learning common image preprocessing techniques for improving the performance of an OCR model.

Explored deep learning models that can assist in removing artifacts (especially signatures) that can be recognized by an OCR model.

4. Tentative tools / frameworks to be used

- Classical Image processing algorithms for preprocessing and denoising.

- Object detection tools such as YOLOv5 for artifact detection.
- Python’s Tesseract engine : Pytesseract for OCR.
- NLTK or SpaCy for Named Entity Recognition.

5. References

A good book to understand OCR was : https://www.pyimagesearch.com/ocr-with-opencv-tesseract-and-python/