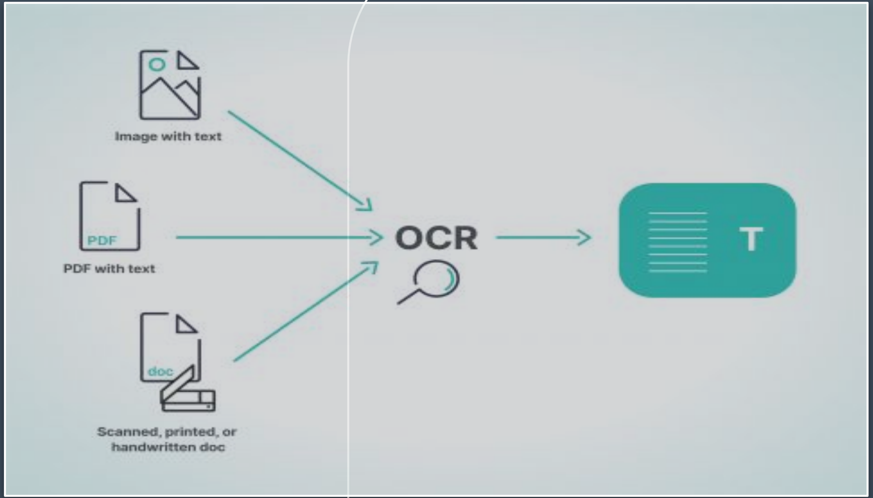
Text Extraction For Bangla Land Records Using OCR

United International University

Machine Learning

Paper Presentation
Section : A

Radowan Ahmed
(011201420)



Introduction

1. This project aims to solve that problem by using OCR with image preprocessing to accurately detect and extract text from land record images.

2. This can support fraud detection, digitization, and secure storage of official documents.

3. These records are crucial in legal, governmental, and property-related matters. However, manual verification is time-consuming, error-prone, and vulnerable to fraud.

Background

01

Automated text detection

02

Fraud detection in land management systems

03

No availability of bangla text detection for land records

04

Crucial in legal, govt and property related matters

Background

05

Enabling future classification of genuine or forged records

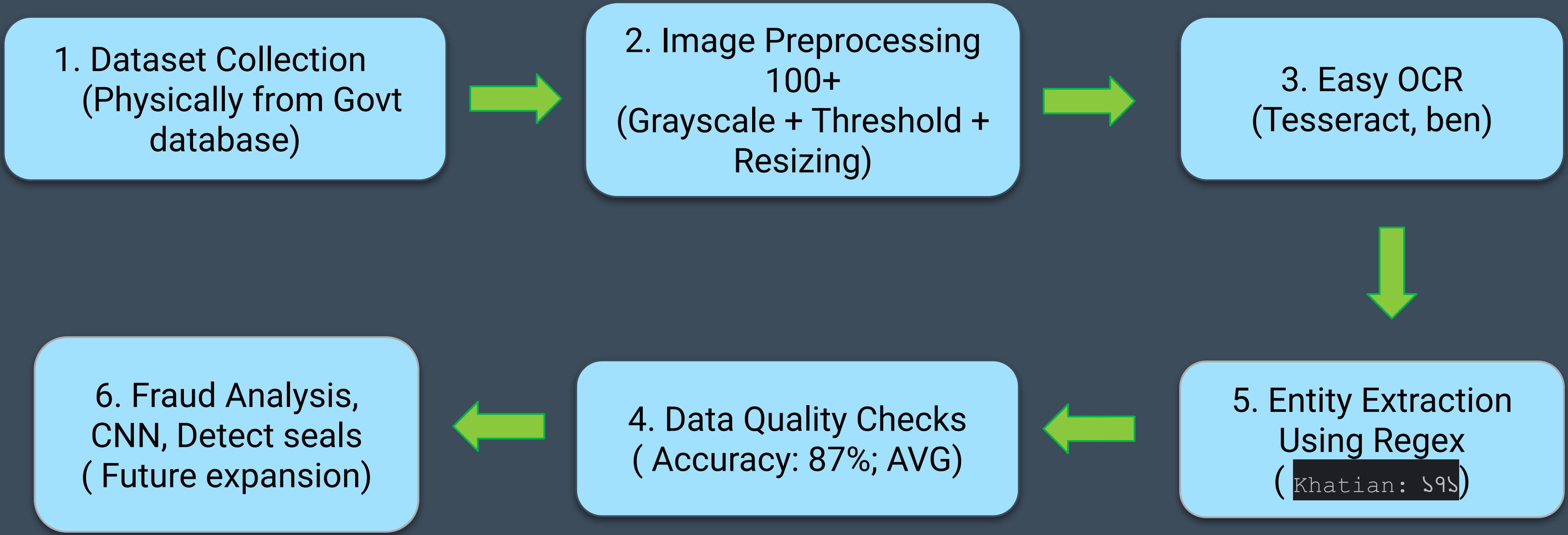
06

Manual data extraction from land records is time-consuming and error-prone. (segmentation)

Aims and objectives

1. The primary objective is develop an automated system for detecting and extracting Bangla text from scanned land record images to support digital verification and fraud analysis.
2. To apply Optical Character Recognition (OCR) techniques for documents such as *Khotiyan*.
3. To preprocess and enhance scanned image quality using techniques like grayscale conversion and thresholding to improve OCR accuracy.
4. Validate the extracted text for completeness, accuracy and consistency to ensure dataset integrity.
5. Foundation for fraud detection, enabling future classification of genuine versus manipulated records.

Methodology



1. Dataset Collection
(Physically from Govt database)

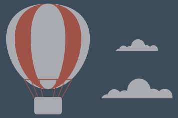
2. Image Preprocessing
100+
(Grayscale + Threshold + Resizing)

3. Easy OCR
(Tesseract, ben)

5. Entity Extraction Using Regex
(`Khatian: ১৭১`)

4. Data Quality Checks
(Accuracy: 87%; AVG)

6. Fraud Analysis, CNN, Detect seals
(Future expansion)



Tools

Tools & Libraries Used

❑ Python

❑ OpenCV – for image preprocessing

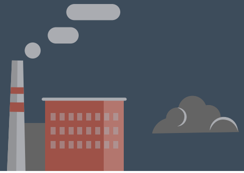
❑ pytesseract – for Bangla OCR

❑ Regex – for entity extraction

❑ PIL – for image handling

Output

```
========= IMG_2901.JPG =========
খতিয়ান: 104
দাগ: 557
মালিক: মো. হাবিবুর রহমান
```



Challenges & Future Works

Challenges

- Poor image quality
- Misrecognized Bangla characters
- Layout variation in documents

Improvements & Future Work

- Apply deep learning (CRNN/CNN) for better OCR
- Detect seals or signatures using contours/CNN
- Add fraud detection: duplicate/missing records
- Export data to structured format (CSV/JSON/SQL)

References

[1] R. Smith, "An Overview of the Tesseract OCR Engine," *Proc. Ninth Int. Conf. Document Analysis and Recognition (ICDAR)*, Curitiba, Brazil, 2007, pp. 629–633.

[2] M. T. Islam, M. S. Hossain, and M. A. Kabir, "Improving Bangla OCR performance using custom training with Tesseract 4," *Int. J. Comput. Appl.*, vol. 178, no. 11, pp. 5–9, May 2019.

[3] T. Paul and I. H. Sarker, "Bangla handwritten digit recognition using deep learning," *Pattern Recognit. Lett.*, vol. 145, pp. 69–76, Mar. 2021.

[4] M. M. Alam, S. Rahman, and M. A. Azim, "Binarization techniques to improve OCR performance on Bangla scripts," *Bangladesh J. Inf. Technol.*, vol. 5, no. 1, pp. 15–21, 2020.

Thank you