Documentation: OCR System for Cadastral Map Digitization

1. Introduction:

The project aims to develop a system that uses optical character recognition (OCR) based system to extract handwritten numerical data from scanned cadastral map images. The main goals of the approach are to precisely recognise numerical and decimal values and filter non-numeric text from photos.

2. Approach:

The workflow involves three stages:

a. Image Preprocessing:

- Convert images to grayscale.
- Remove noise using Non-Local Means Denoising.
- To improve text areas, use Otsu's Binarization thresholding.
- Handle alpha channels (transparency) if present.

b. OCR Text Extraction:

- Make use of EasyOCR's pre-trained, numeric data-optimized 'number-dense` model.
- Text regions can be extracted from processed photos.

c. Image Postprocessing:

- Use regex to filter results so that only legitimate numbers are retained.
- Handle edge cases.
- Output results in CSV format with image names and extracted numbers.

3. Tools and Models:

a. OpenCV
b. EasyOCR
c. Regex
d. CSV
e. OS

4. Key Features:

- **Noise Consideration:** Takes into account stains, smudges, and irregular lighting.
- **Decimal Extraction:** Extracts decimals with accuracy.
- **Error Handling:** Logs errors and skips invalid images.

5. Sample Output:

1. **Image:** VW_DTN01_L_00007_101000000000000000002038853600000.front.JPG
   Extracted Numbers: "3", "8", "33", ..., "9.5"
2. **Image:** VW_DTN01_A_00002_101000000000000000002051033000000.front.JPG
   Extracted Numbers: "4", "19", "2", ..., "006341"

GitHub Link: Please click here