IIT Tirupati Navavishkar I-Hub Foundation Coding Challenge – 1

Title: Develop an Optical Character Recognition (OCR) based Characters and Numbers Detection Model for Cadastral Maps.

Background: Cadastral maps are detailed land mapping documents used by governments and land authorities to represent property boundaries, land ownership, and administrative divisions. These maps often contain critical alphanumeric information such as plot or survey numbers, landowner names, village or ward identifiers, and other survey-related codes. Traditionally available in printed or scanned formats, these maps are not machine-readable, which poses challenges for digitization and integration with modern geospatial systems. The use of Optical Character Recognition (OCR) for automatic detection of characters and numerals from cadastral maps is essential for converting these legacy documents into digital formats. Accurate recognition of plot numbers and landowner details supports land ownership verification, reduces disputes, and facilitates integration with property tax databases and spatial decision support systems. This task enables real-time applications such as mobile land survey tools, smart urban planning, disaster relief management, and agriculture monitoring by linking geotagged land parcels to relevant textual information.

Description of Challenge: The main aim of this task to extract all the possible characters and numbers from cadastral maps which involves several tasks but not limited to:

1. Preprocessing of Cadastral map images
2. Text and numbers region detection
3. Characters and Numbers Segmentation
4. Optical Character recognition for recognizing the segmented characters and numbers
5. Post-processing

Input Data: the input data is given in the image format; sample image is shown in Figure.1;

Output Data:

An excel sheet or CSV sheet which includes all the extracted characters and numbers as follows;

Characters	Numbers
Konal, Gonal, ……	76, 20, 22, 24, ….

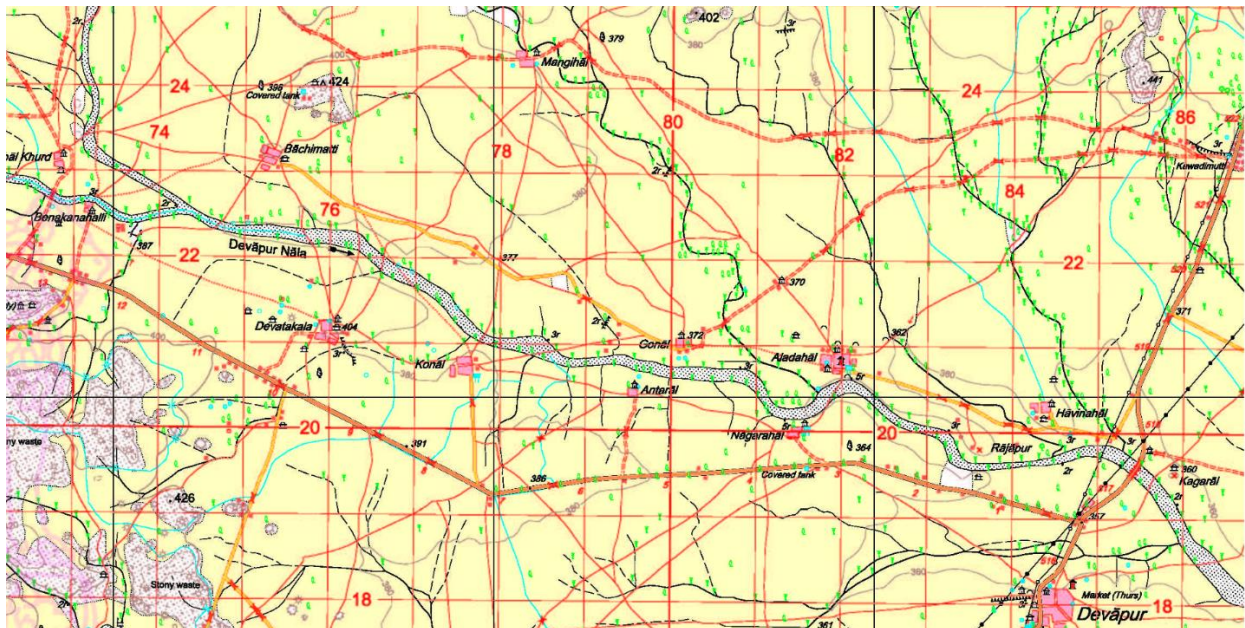Figure.1 Cadastral Map (https://surveyofindia.gov.in/pages/educational-map-series)

The characters and numbers present in the input images are as follows:

Characters	Numbers
Benakanahalli, Devapur Nalla, cover tank, Devatakala, Mangihal, Gonal, Aladahal, Covered tank, Antaral, Rajapur, Stony waste, Antaral, Nagarahal, Devapur, kagaral, kawadimutt, Konal.	74, 24, 387, 13, 12, 11, 10, 76, 22, 396, 424, 404, 379, 372, 362, 364, 20, 426,18, 391, 386, 377, 364, 361, 402, 400, 84, 24, 521, 519, 518, 517, 516, 357, 371, 522, 360, 379

Hints for better Performance:

1. Apply preprocessing techniques to filter out noise and blur if exists in the image.
2. Machine learning or Deep Learning algorithms can be leveraged.
3. Apply image conversion methods at the pre-processing level to extract better features.

Deliverables:

A. **Output Files:** A CSV or Excel sheet with all extracted characters and numerical information.

B. **Code Repository:** A complete code repository which consists of entire OCR system which includes preprocessing, text detection, characters and number segmentation, OCR and post processing. (Preferred format is .zip)

C. **Report:** A summarized report which explain the complete details about the methods, tools and performance analysis. (Preferred format is .pdf)