DOCUMENT READER - IMAGE TO TEXT

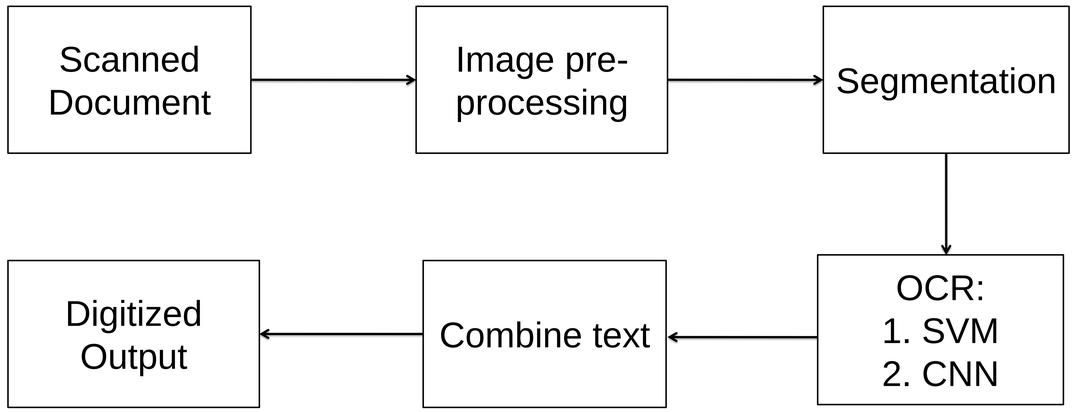
ANURAG JAIN (MT19AI014), SIDDHARTH YADAV (MT19AI011)

INDIAN INSTITUTE OF TECHNOLOGY JODHPUR

PROBLEM STATEMENT

Document Reader is a well known problem in the field of computer vision, pattern recognition, where the objective is to convert images with English text into editable text files. The problem statement is not as trivial as it looks, as most of the time images are not very well captured, the text alignment may not be horizontal, the characters may be blur, etc. The main objective is, to extract all the text from scanned documents so that it can be used for searching purpose and create an overlay so text can be copied and pasted from the images saved as text document.

APPROACH



- Image Pre-Processing will include binarization of the image, slant and shear or skew correction of the text.
- In segmentation we will first extract text lines, followed by extraction of words, and finally character segmentation from words.
- OCR will be performed at character level.

DATASET

- For training of OCR, 62 classes dataset will be used (A-Z, a-z, 0-9) and other classes may be added based on requirement.
- Annotated images of words are freely available online.

- Labeled text document image is not readily available. We will create a small 30 images dataset comprising of 15 good quality and 15 noisy images.

MOTIVATION

Searching in or to analyze huge text manually becomes very difficult and sometimes even for extracting very small amount of information we need to search huge text, which doing manually is very difficult. In contrast digital text processing is very easy to analyze and search for relevant content. The effort and time complexity will be reduced exponentially if we generate the editable text version of the images containing text.

EVALUATION METRICS

- Character level analysis:
  - mAP@5
  - Accuracy %
- Word level analysis:
  - Accuracy %
  - 1-deviation Accuracy
  - 2-deviation Accuracy
- Document level analysis:
  - How much % of the text is correctly recognized
  - Total words missed

TIMELINE

MID SEM SUBMISSION :

- Pre-Processing of images
- Word level segmentation
- Dataset Creation

FINAL SUBMISSION:

- Character level segmentation
- Training of OCR
- Evaluation of results
- Comparison of algorithms

REFERENCES

•A. Kae, G. Huang, C. Doersch and E. Learned-Miller, "Improving state-of-the-art OCR through high-precision document-specific modeling," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 1935-1942.

•Sen Maitra, Durjoy & Bhattacharya, Ujjwal & Parui, Swapan. (2015). CNN based common approach to handwritten character recognition of multiple scripts. 1021-1025. 10.1109/ICDAR.2015.7333916.