Text/Name Extraction from Scanned Document

BY X.WEI

1 Method overview

The task contains 2 tasks: extract text from scanned image, and extract participant names (and maybe more structural infromation in the future) from the text. The code/notebook is in this git repo.

1.1 Text extraction

For text extraction, in the spirit of enginnering rather than academic, I use `pyocr` (which uses tesseract-ocr in the backend) to do the job.

With the `pyocr` module, we can not only extract raw text, but also extract lineboxes/wordboxes, which contain in additional **position information** of the recognized texts.

1.2 Name extraction

I have tried to process the raw text extracted from `pyocr`, then use `langid` to detect language then use `nltk` for named entity recognition (see this notebook), but this work out poorly as there lacks the pretrained NER model for french/german.

Then I tried to first identify paragraphs by merging wordboxes returned by `pyocr`, after some tuning, the paragraph separation work out pretty well (see here). Then I just take the string before the semicolon as participant name. With this technique I can extract text as paragraphs, and the participant names are correctly extracted. The tuning of paragraph separation can be found in this notebook.

2 Run the code

To run the code in `text_extract.py`, you have to install `tesseract`, and install the python module: `opencv, pyocr, matplotlib, Pillow`.

After installing these, you should be able to run the code, the ocr step taks some time (~ 1min), then the code will output paragraphs of text extrated, and ouput the participant names.

3 Discussion

The ad-hoc solution for participant name extraction works well for the test image, but maybe this cannot generalize to all scanned documents, but the paragraph separation still gives a track of extracting information. Maybe should exploit more of the boxes information, for example: box shape, box positon, font size inside box, etc. The box information might help us in finding structural elements of the document.

Another thing to explore is the vertical/horizontal lines in image, with these lines maybe we can segment the image into several parts, and tesseract might work better on separate parts of the image.