1 Text Line Extraction - DC1

1.1 Motivation

Since we chose to process the first document class we decided to go with a run length smearing approach, because machine written text doesn't have much outliers which could be a problem for the algorithm. Luckily we already had some very good results with this approach. The algorithm showed some problems when the lines were too narrow. Although the smearing worked, the connected component algorithm sometimes interpreted two lines as one.

After we got near perfect result on our own document class we decided to also try the histogram approach but it didn't yield better results for the problems mentioned before so we decided to stick with the smearing approach

1.2 Method description

1.2.1 RLSA

First we tried to do horizontal and vertical smearing and then combine the two images with a logical and but results showed that a simple horizontal smearing yielded the best results. The algorithm is written as such that we can provide a limit of white pixels between two black ones so that if it exceeds that limit the pixels in between get colored black. We also added a threshold for the maximum amount of white pixels so that the side-space of the paper doesn't get colored black. After some tests we figured that the values 3 and 50 accordingly were best for our needs.

1.2.2 Connected component

The Connected component algorithm builds on the output of the RLSA. It takes the smeared image and draws a bounding box for each connected component. This method works satisfactorily for our document class. There are only some minor problems with list bullet points and some dots over some i's, because these don't get connected with the RSLA. A additional vertical smearing for each line could solve this issue. To adjust potential binarization errors, the bounding box can be enlarged a adjustable amount of pixels.

1.3 Results

1. **DC1**: near perfect results, some problems with bullet points because they are farer away from the text than our upper white pixel threshold.

2. **DC2**: mixed results but relatively good for the amount of possible things that could have gone wrong. Problem with big starting letters and side notes. Also some problems with lines that were interpreted as one because they were to close together or lines that were interpreted as two because the space between words was too big.

3. **DC3**: With the threshold of 50 almost every word gets interpreted as own component but with some tweaking in this department we achieved quite good results since the hand writing was very regular.

4. **DC4**: Problems with lines that are too close and side notes. Otherwise the results weren't that bad.

5. **DC5**: Problem that side notes and main text get interpreted as one line. Also due to big starting letters sometimes two lines get interpreted as one.

1.4 Conclusion

The method is very reliable in regular text and yields almost perfect results with machine written text. Even some distortion is manageable. The biggest drawback of the method is that when letters overlap lines above or beyond they get interpreted as one line. Some problems mentioned above could surely be avoided by further tweaking the threshold values.

The method could be further improved if we would not smear the whole picture at once but in stripes so that we could differ between side notes and main text for example. Furthermore a polygonal approach in the connected component algorithm would better the results with handwritten text to some extent.