A Brief Description of the DSPH Method

We propose a document image segmentation method namely the DSPH (Document Segmentation with Probabilistic Homogeneity) method which has the following steps:

1. Binarization: a combination of two different configurations of Sauvola's method is used.
2. Text and non-text classification: we introduce the text homogeneity and classify text and non-text components using a probabilistic map which is computed by exploiting text homogeneity at regional and component layers.
3. Text region extraction: text lines and text blocks are extracted by exploiting text homogeneity at progressively higher layers, and are refined to address versatile text structures such as drop capitals and heterogeneous text blocks. Text paragraphs are subsequently identified through a within text block segmentation based on the analysis of relative locations of the text lines.
4. Non-text structure extraction: by following a seed-grow-extraction (SGE) process separators are first extracted from non-text components. Remaining non-text components are classified heuristically into images, graphics, charts and 'reverse-video' text. Similar processes that all follow the SGE paradigm are performed to extract these regions.