Automated Layout Analysis for Manuscript Investigation

Abstract

Layout Analysis (LA) of documents is a main pre-processing step for OCR and handwritten text recognition, document classification, document annotation (for further processing), and information retrieval. LA refers to the process of transforming an image pixel representation into a higher level representation by forming groups of text lines, text blocks and/or graphical elements. Hence, layout analysis is a description of the visual appearance of documents (syntactic interpretation).

The LAs presented is a bottom-up approach, which is flexible since there is no need for incorporating global knowledge of a document's layout. Text elements are grouped to text lines based on their profile boxes. The text line clustering is using a global energy minimization which is robust to local skew and local deviations of the word's height or slant and noise. Furthermore, a grouping is possible even if documents lack global layout structure which is likely for document fragments.

Florian Kleber *received his PhD degree in Computer Science in 2014 at TU Wien, Austria. He gained experience in document analysis in several projects, dealing with the multi-spectral acquisition and restoration of ancient manuscripts and historical documents. Since 2014 he is working as a senior researcher at the Computer Vision Lab, Institute of Computer Aided Automation. Additionally, he is involved in lecturing at TU Wien, amongst others, Document Analysis, and is publishing several papers at international conferences, such as ICDAR and DAS, since 2008. His research interests are Cultural Heritage and Document Analysis Applications.*