An overview of text-image alignment methods

Thomas Breuel
University of Kaiserslautern

4 June 2013

I will give a brief overview of the current capabilities of automated text-image alignment techniques. Text-image alignment is a key part of OCR training, so many approaches to OCR already support it. Alignment can broadly be divided into two steps: finding text-line images in document images and identifying the corresponding text, then aligning each text-line image with the corresponding text; I will describe approaches to both problems. I will also describe the special challenges posed by historical documents, including differences in line breaks, differences in orthography, and unusual character sets and ligatures.