Part B

Page Analysis

Historically, documents in printed form have been organized into pages in a way that facilitates sequential browsing, reading and/or search. The structure of these documents varies by genre; there is no doubt that there is an implicit organization that helps readers navigate the content in order to enable the transfer of information from the author to the reader. Scientific articles first provide basic metadata about the subject and authors, phone books provide indexing information and are ordered alphabetically, business correspondence follows established protocols, and tables organized implicitly in row and column structures provide an efficient representation of information.

As authors, humans have a canny ability to tweak the basic rules of organization to enhance the experience of the reader, and as readers we are able to use this information implicitly to comprehend what we are reading. While ideally we should be teaching our document analysis systems to be able to parse structures and use them to their advantage, traditionally page analysis has simply been a method to divide up content in a way that enables content analysis algorithms to perform more efficiently. Pages are divided into zones and these zones labeled and analysis by content-specific routines. Zones of text are divided into lines, lines of text into words, etc. As part of interpretation, the logical relationships between zones or regions on the page can be analyzed for reading order or as functional units of organization.

▸Part B (Page Analysis) focuses on the fundamentals of this process and is divided into three chapters. In ▸Chap. 5 (Page Segmentation Techniques in Document Analysis), Koichi Kise addresses the first of these problems, page segmentation. Prior to the widespread deployment of word processing systems, textual documents tended to be fairly simple with text and illustrations falling into overlapping, homogeneous regions that could be segmented with basic image analysis techniques. However, as more complex layouts evolved and the possibility of processing less constrained handwritten material make this problem especially challenging. Robust solutions are essential and this chapter provides a firm baseline for understanding what is required.

▸Chapter 6 (Analysis of the Logical Layout of Documents) addresses the problem of analyzing the logical layout of documents. As previously suggested, this problem is often highly dependent on the type of document, so assumptions are made about genre. Andreas Dengel and Faisal Shaifait have a significant history with this problem and provide an in-depth discussion of the issues and solutions of logical analysis. They provide some important practical recommendations for those needing to design or incorporate their own logical analysis capabilities.

▸Chapter 7 (Page Similarity and Classification) takes a step back and examines the higher level of the classification of pages and the computation of a page similarity measure. Simone Marinai motivates this chapter as a step required to take input from a heterogeneous workflow and classify the document with respect to genre in order to allow subsequent genre-specific analysis. The related problem of determining similarity addresses both content and structural features. As previously suggested, layout information often provides important clues and user are more and more interested in grouping or searching document collections by document type. Providing robust similarity measures is an essential step toward this goal.