2

Document Structure and Layout Analysis

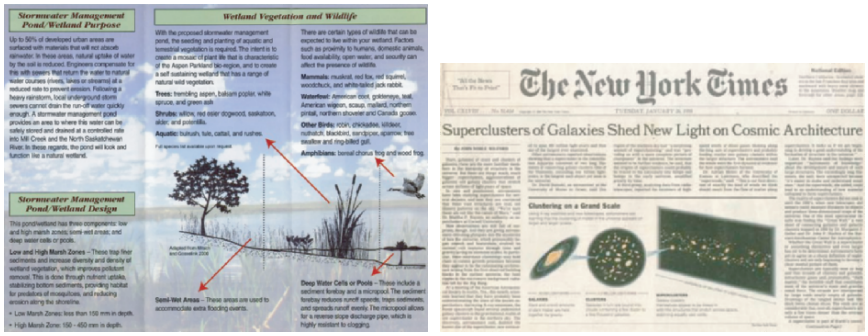
Anoop M. Namboodiri and Anil K. Jain

2.1 Introduction

A document image is composed of a variety of physical entities or regions such as text blocks, lines, words, figures, tables and background. We could also assign functional or logical labels such as sentences, titles, captions, author names and addresses to some of these regions. The process of *document structure and layout analysis* tries to decompose a given document image into its component regions and understand their functional roles and relationships. The processing is carried out in multiple steps, such as preprocessing, page decomposition, structure understanding, etc. We look into each of these steps in detail in the following sections.

Document images are often generated from physical documents by digitization using scanners or digital cameras. Many documents, such as newspapers, magazines and brochures, contain very complex layout due to the placement of figures, titles and captions, complex backgrounds, artistic text formatting, etc. (see Figure 2.1). A human reader uses a variety of additional cues such as context, conventions and information about language/script, along with a complex reasoning process to decipher the contents of a document. Automatic analysis of an arbitrary document with complex layout is an extremely difficult task and is beyond the capabilities of the state-of-the-art document structure and layout analysis systems. This is interesting since documents are designed to be effective and clear to human interpretation unlike natural images.

As mentioned before, we distinguish between the physical layout of a document and its logical structure [3]. One could also divide the document analysis process into two parts accordingly.



(a) (b)

Fig. 2.1. Examples of document images with complex layouts.

2.1.1 Physical Layout and Logical Structure

The *physical layout* of a document refers to the physical location and boundaries of various regions in the document image. The process of *document layout analysis* aims to decompose a document image into a hierarchy of homogenous regions, such as figures, background, text blocks, text lines, words, characters, etc. The algorithms for layout analysis could be classified primarily into two groups depending on their approach. Bottom-up algorithms start with the smallest components of a document (pixels or connected components) and repeatedly group them to form larger, homogenous, regions. In contrast, top-down algorithms start with the complete document image and divide it repeatedly to form increasingly smaller regions. Each approach has its own advantage and they work well in specific situations. In addition, one could also employ a hybrid approach that uses a combination of top-down and bottom-up strategies.

In addition to the physical layout, documents contain additional information about its contents, such as titles, paragraphs, captions, etc. Such labels are logical or functional in nature as opposed to the structural labels of regions assigned by layout analysis. Most documents also contain the notion of *reading order*, which is a sequencing of the textual contents that makes comprehension of the document easier. Languages such as Arabic, Chinese, etc. can have different reading directions as well (right-to-left, top-to-bottom). The set of logical or functional entities in a document, along with their inter-relationships, is referred to as the *logical structure* of the document. The analysis of logical structure of a document is usually performed on the results of the layout analysis stage. However, in many complex documents, layout analysis would require some of the logical information about the regions to perform correct segmentation.

Most document images contain noises and artefacts that are introduced during the document generation or scanning phase. In order to make the analysis algorithms more robust to this noise, the layout and structure