1

Reading Systems: An Introduction to Digital Document Processing

Lambert Schomaker

1.1 Introduction

Methods for the *creation and persistent storage* of text [10] have existed since the Mesopotamian clay tablets, the Chinese writings on bamboo and silk as well as the Egyptian writings on papyrus. For *search and retrieval*, methods for systematic archiving of complete documents in a library were developed by monks and by the clerks of emperors and kings in several cultures. However, the technology of *editing* an existing document by local addition and correction of text elements has a much younger history. Traditional copying and improvement of text was a painstakingly slow process, sometimes involving many man years for one single document of importance. The invention of the pencil and eraser in 1858 was one of the signs of things to come. The advent of the typing machine by Sholes in 1860 allowed for faster copying and a simultaneous on-the-fly editing of text. The computer, finally, allowed for a very convenient processing of text in digital form. However, even today, methods for generating a new document are still more advanced and mature than are the methods for processing an existing document.

This observation may sound unlikely to the fervent user of a particular common word-processing system, since creation and correction of documents seems to pose little problems. However, such a user has forgotten that his or her favourite word-processor software will only deal with a finite number of digital text formats. The transformation of the image of an existing paper document − without loss of content or layout − into a digital format that can be textually processed is mostly difficult and often impossible. Our user may try to circumvent the problem by using some available software package for optical-character recognition (OCR). Current OCR software packages will do a reasonable job [16] in aiding the user

to convert the image into a document format that can be handled by a regular word-processing system, provided that there are optimal conditions with respect to:

- image quality;
- separability of the text from its background image;
- presence of standard character-font types;
- absence of connected-cursive handwritten script;
- simplicity of page layout.

Indeed, in those cases where strict constraints on content, character shape and layout do exist, current methods will even do quite a decent job in faithfully converting the character images to their corresponding strings of digital character codes in ASCII or Unicode. Examples of such applications are postal address reading or digit recognition on bank cheques.

On the other hand, if the user wants to digitally process the handwritten diary of a grandparent or a newspaper snippet from the eighteenth century, the chances of success are still dim. Librarians and humanities researchers worldwide will still prefer to manually type ancient texts into their computer while copying from paper rather than entrusting their material to current text-recognition algorithms. It is not only the word processing of arbitrary-origin text images that is a considerable problem. Even if the goal can be reduced to a mere search and retrieval of relevant text from a large digital archive of heterogeneous text images there are many stumbling blocks. Furthermore, surprisingly, not only are the ancient texts posing problems. Even the processing of modern, digitally created text in various formats such as web pages with their mixed encoded and image-based textual content will require "reverse engineering" before such a digital document can be loaded into the word processor of the recipient.

Indeed, classification of text within an image is so difficult that the presence of human users of a web site is often gauged by presenting them with a text fragment in a distorted rendering, which is easy on the human reading system but an insurmountable stumbling block for current OCR systems. This weakness of the artificial reading system thus can be put to good use. The principle is known as "CAPTCHA: Completely Automated Public Turing Tests to Tell Computers and Humans Apart" [5]. During recent years, yet another exciting challenge has become apparent in pattern-recognition research. The reading of text from natural scenes as recorded by a camera poses many problems, unless we are dealing with a heavily constrained application such as, e.g. the automatic recognition of letters and digits in snapshots of automobile licence plates. Whereas licence-plate recognition has become a "mere" technical problem, the camera-based reading of text in man-made environments, e.g. within support systems for blind persons [8], is only starting to show preliminary results.

Non-technical users will often have difficulties in understanding the problems in digital-document processing (DDP) and in optical character recognition. The human reading process evolves almost effortlessly in the