A Comparative Study on OCR Tools

Carlos A.B. de Mello
Rafael D. Lins
Departamento de Informática, Universidade Federal de Pernambuco
Recife - PE, Brazil {cabm, rdl}@di.ufpe.br

Abstract

Tools for Optical Character Recognition (OCR), commercially available today, provide different recognition degrees depending on a number of factors. We analyse here the features of six of the most widely used "off-the-shelf" OCR softwares.

1 Introduction

The invention of paper in Egypt at about 4,000 B.C., represented one of the greatest revolutions of mankind, due to its practicity, portability and cost. It immediately replaced all other forms of information storage used (such as stone and wood carving and argil brick print). Today, it still is the form of media that accumulates the largest amount of information, although, it is not the most efficient one. Paper brings several disadvantages, such as the physical space necessary to store it, which can increase exponentially with the quantity of information.

Digital storage media provides more space efficient solution to information retrieval. The invention of image digitizers, such as *scanners*, made possible to store documents in a more efficient way and to protect their information of the wear and tear over time. Information stored in secondary memory devices (CD-ROMs, hard disks, Zip Disks, Jaz Drives, etc.), may be copied without loss of information. However, image files are greedy storage consumers. For example, a aheet of A4 paper (210 x 297 cm) image digitized at 200dpi (*dots per inch*) of resolution, 256 colors requires 4,113 Kbytes for storage in BMP file format (the standard Microsoft Windows image file format). The corresponding text file can be stored in less than 100 Kbytes.

A solution to the storage problem came with the development of computer systems that could translate image into text format. A non-automatic transposition is unacceptable because of the costs involved and the very low speed. An automatic process brings the problems of recognizing the characters present in documents and of translating them to ASCII. This translation procedure is called *Optical Character Recognition* (OCR). Besides saving storage space, there are many advantages of using text files such as the possibility of running searches for keywords. While it is very easy to do this with text files it is almost impossible for images. One of the difficulties of dealing with OCR's come with the choice of the best recognizing method and in the best setting of parameters for digitalization (resolution, brightness, contrast, number of colors, etc).

There are several commercial softwares available to perform optical character recognition. In this paper, we analyze ther recognition performance of the six most widely used commercial OCR tools. They are:

- Omnipage 9.0 (Caere Corporation)[14]
- Corel OCR Trace 8.0 (Corel Corporation)[16]
- SmartPage 2.1 (Recognita Corporation)[18]
- Wordlinx (OCRON Inc.)[19]
- TextBridge Pro 98 (Scansoft Inc.)[20]
- TypeReader Professional 4.0 (Expervision)[21]