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]

Table 1 below summarizes some important features of these OCR tools.

Amongst the softwares tested only OmniPage and Corel OCR Trace work with greyscale images. Omnipage, however, converts the greyscale image to monochromathic before performing the recognition, while Corel OCR Trace is the only one of the tested softwares capable of processing in greyscale.

	Omnipage	*Corel*	*SmartPage*	*Wordlinx*	*TextBridge*	*TypeReader*
Image Types (colors)	Greyscale and B&W	Greyscale and B&W	B&W	B&W	B&W	B&W
Rotation	Yes	Yes	No	No	Yes	No
Dictionary	Yes	Yes	No	Yes	Yes	Yes
Dictionary changes	Yes	No	No	Yes	Yes	Yes
Maximum Resolution Tested	250 dpi	250 dpi	150 dpi	250 dpi	250 dpi	250 dpi
Input File Formats	Several	Several	Only TIFF	Several	Several	Only TIFF
OCR Technique	Neural Networks	Not Specified	Not Specified	Not Specified	Neural Networks	Not Specified
Font Selection	No	Yes	No	No	No	Yes
Multiple Layouts	Yes	Yes	No	No	Yes	Yes

Table 1. Commercial OCR Softwares main features

It has to be noticed that Smartpage works only with resolutions less than 150 dpi.

2 Experimental Results

For our initial experiments we have used a text extracted from an Internet site and three other images of a technical paper on Computer Science All the texts were printed with a Hewlett-Packard Laserjet 5L printer with 300 dpi of resolution and printed in a Chamex [15] A4 white paper with 210x297mm and 75 g/m2. The documents have in average 3714.75 characters and 624.25 words.

The documents were digitized with a Hewlett-Packard Scanjet 4C scanner with maximum resolution of 720 dpi. Brightness was set to 123% and contrast to 133%. For digitalization, we used the HP Deskscan II (the software that came with the device itself). The documents were scanned from left to right and top to bottom.

The boundary between the ink and the paper is not so clear in greyscale images as it is in monochromatic ones. When the image is digitized it is affected by the Gibbs phenomenon [5]. Real world objects are not bounded in frequency. Digitalization is done using a sampling frequency chosen so that it takes into account the Nyquist rate [3]. This scheme bounds the digitized objects in the frequency. We can see this phenomenon as a "ring" that envolves the image whenever there is a strong variation of hues. Figure 1 below zooms into it. This one was digitized in greyscale (8 bits) and with 200 dpi of resolution.

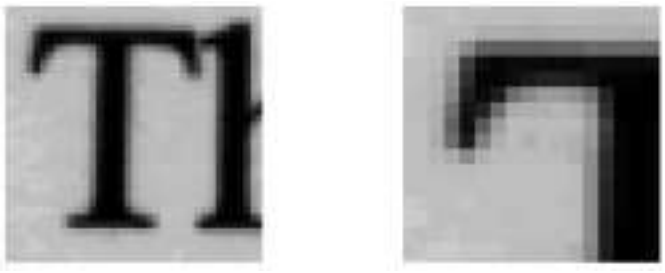


Figure 1: Gibbs phenomenon - original image and zooming

2.1 Resolution

The image of a text file is obtained by a scanner, in general. The resolution set for image digitalization influence the recognition capability of OCR tools. A higher digitalization implies proportional increase on space consumed but may not yield higher degrees of textual recognition.

All documents were scanned with several different resolutions, in greyscale and monochromatic to check the sensitivity of the six softwares. The digitalization device for the experiments herein was a Hewlett-Packard flatbed scanner with maximum resolution of 720dpi. The average hit rate[1] of the softwares for the four documents can be seen on tables 2 and 3, below. No dictionary lookups was used in any case. These hit rates are exactly what the OCR program generated with no external help. Both greyscale and monochromatic images were generated by scanning, without any color processing.

dpi	*Omnipage*	*Corel*	*SmartPage*	*Wordlinx*	*TextBridge*	*TypeReader*
75	5.3248 %	0.000 %	0.000 %	10.2583 %	9.0695 %	0.000 %
100	42.6984 %	87.0991 %	62.2227 %	58.3072 %	89.4951 %	89.9835 %
150	98.9890 %	96.9371 %	88.2299 %	68.5209 %	99.2113 %	99.0894 %
200	99.5297 %	99.3187 %	-	66.1667 %	66.1667 %	99.4135 %
250	99.5238 %	99.0501 %	-	70.5298 %	70.5298 %	99.4737 %

Table 2. Average hit rate for different resolutions for Black and White images

SmartPage does not work with resolution higher than 150 dpi. A degradation was found with the increase of the resolution from 200 dpi to 250 dpi in Omnipage and in Corel OCR Trace. This can be explained due to the addition of noise by the digitalization process.

Wordlinx presented the worst hit rates amongst all tested softwares. Even with higher resolutions its results were below the acceptable with almost 30% of error rate. Up to 150 dpi, TextBridge and TypeReader reached the best rates with almost 100% of accuracy. These softwares, however, have the disadvantage of not working with greyscale images.

As shown on Table 1, only Omnipage and Corel OCR Trace work with grey scale images. Their average accuracy in these kind of images can be found in Table 3.

	Omnipage	*Corel*
75 dpi	23.7169 %	20.8 %
100 dpi	48.5133 %	33.2367 %
150 dpi	97.7649 %	60.2649 %
200 dpi	98.6341 %	97.3096 %
250 dpi	99.6689 %	99.2135 %

Table 3. Softwares hit rate in different resolutions for Greyscale images

Even though OmniPage and Corel OCR Trace have similar response in higher resolution images, Omnipage converged to better results faster than Corel OCR Trace and it also presented better results in lower resolutions.

2.2 Sensitivity to Rotation

The OCR process is also sensitive to document rotation. The hit rate of OCR tools varies with the degree of rotation of the document. While some softwares deal with the rotation of the image with little loss of information, others do not work properly.

After digitization, OCR softwares must check if the image is rotated or not. For this a skew detection process is made to check for undesired degrees of rotation. This is necessary because some techniques used in OCR programs require inputs to be correctly aligned. The skew detector can determine the angle of rotation of a document image and use this information to reorient the image to align it.

One way to perform skew detection is with the use of *Hough transforms* [3, 10]. This procedure can check if there is a rotation in the image and its angle, allowing the software to perform a new rotation to make the image straight.

[1] The number of words correctly transcribed

Image rotation is treated only by Omnipage, Corel OCR Trace and Text Bridge Pro. These results are shown on Table 4 where the image was clockwise rotated. For this experiment we just used one of the images (the one extracted from Internet).

	Omnipage	*Corel*	*TextBridge*
1 degree	98.634106 %	99.254967 %	99.25497 %
2 degree	98.634106 %	96.771523 %	99.1308 %
3 degree	98.509934 %	33.236755 %	99.006623 %

Table 4. Software's hit rate with different rotations for a image with 250 dpi

Omnipage has dealt with up to 12 degrees of rotation with a small error rate. The same did not happen with the other softwares which had a high degradation while working with images with more than 3 degrees (for Corel OCR Trace) and 10 degrees (for TextBridge) with less of 10% of degradation in these higher degrees.

2.3 Brightness

The next of our experiments has the objective to find the best brightness value for digitization in order to achieve the best performance for each software. For this purpose, we have worked with 200 dpi black and white images. Two softwares were not used in this test. Smartpage because of its limitation to resolution higher than 150 dpi and TypeReader which, in its demo version (the one used here), do not allow the storage of the text file generated (in the previous experiments we have generated this file from a screen dump but this scheme is not so easily done now).

Herein the documents were digitized with different brightness values from 1 to 255 totalizing 50 images of each text. These images were then used as input to the OCR tools. Their response was compared with the original file and the percentual average error rate was plotted as can be seen in figures 2 and 3 next.

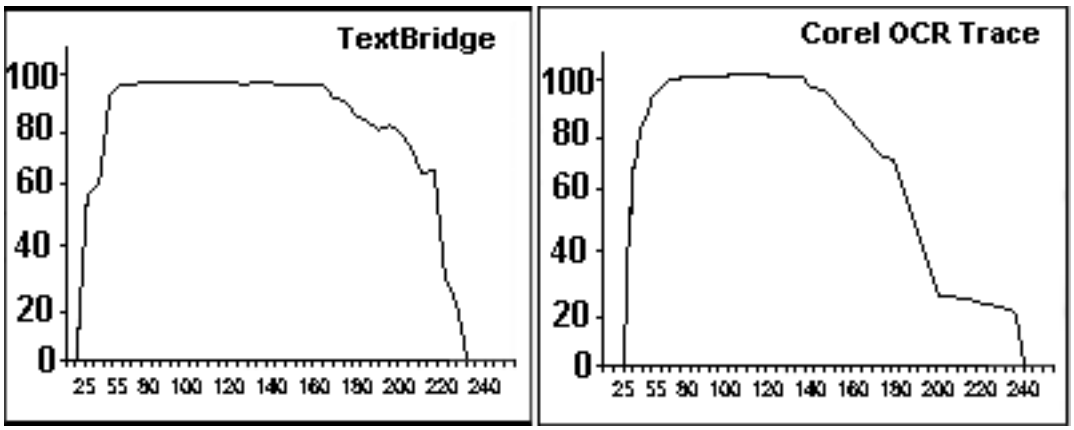


Figure 2: Brigthness *versus* Accuracy

Some points of degradation can be noticed in the accuracy graphics of figures 2 and 3. For Textbridge, for example, a degradation was noticed with brightness values of 50 and 55. We can see details of the zooming into images generated at these values in figure 4 below.

The word displayed in the figure would be correctly translated to "History". The image scanned with 50 of brightness generated the translation "History" with Textbridge. The same software translated the word as "HIstwy" with the image scanned with brightness set to 55. The same occurred with some of the other softwares with other images producing a decrease in the curve of the accuracy.

The average brightness values which produced the best performance for each software are presented on Table 5.

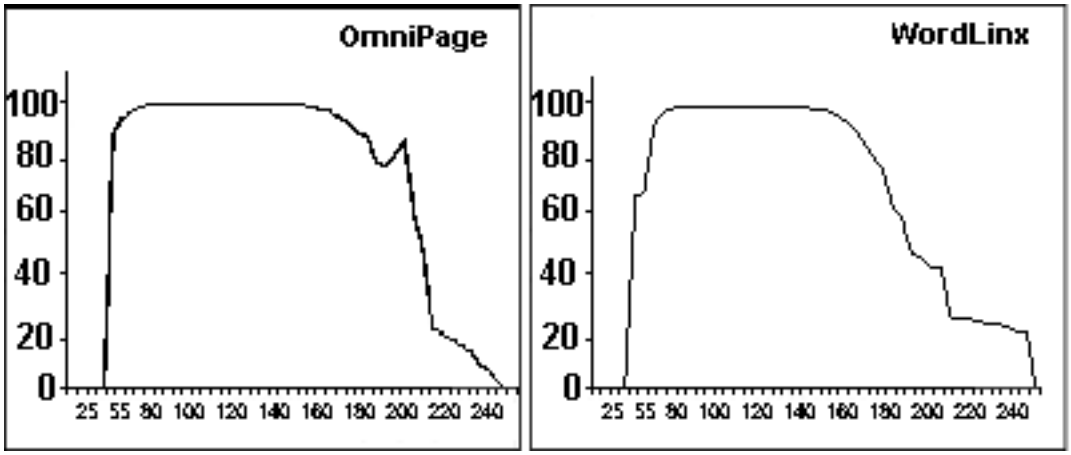


Figure 3: Brigthness *versus* Accuracy

Software	*Best Brightness Value*
TextBridge	90
Corel OCR Trace	115
Omnipage	105
WordLinx	105

Table 5. Best brightness values for recognition

The average best value of brightness is 103.75 which is less than the medium of the range (from 0 to 255) of the scanner. The accuracy of the softwares was better at lower brightness values (darkest ones) than higher ones. The error rate reached in these values is different for all the softwares. We have only shown the best settings in average for each of them.

Figure 4: Zooming into images generated with brightness sets of 50 and 55

2.4 Background Colors

For this experiment we have used four different types of colored papers. The papers had background colors with predominance of blue (RGB: 153, 229, 255), green (RGB: 153, 255, 229), Pink (RGB: 255, 178, 255) and yellow (RGB: 255, 255, 229). Each text was printed (with the same printer as before) in these papers and they were scanned with 200 dpi resolution. Only OmniPage and Corel OCR Trace were capable of working with the images. As the other softwares work only with monochromatic images, they could not open the images. Even though Corel OCR Trace and Omnipage could work with the colored images, only Corel produced acceptable results as **Omnipage could not recognize any character** of the text in all four images of each text. We can see next the average hit rate of Corel OCR Trace in three of the documents.

Background Color	*Corel's Hit Rate*
Blue	99.52%
Green	97.75%
Pink	97.99%
Yellow	77.37%

Table 6. Corel OCR Trace's average hit rates for different background colors

The best hit rates were obtained with a blue background color and the worst ones with a yellow background color. In the worst case the error rate was around 30% with a background 255 (red), 255 (green) and 229 (blue) for one of the test images.

2.5 Historical Documents

For the second phase of our experiments, we have chosen four images of historical typed letters from Joaquim Nabuco [2] belonging to Nabuco's Project [22]. These images are a good representative of a large part of this file which encompasses over 6,500 documents. All of they are from the end of the nineteenth century and were scanned with 200 dpi resolution and in greyscale. Appendix A shows a zooming into one of these images.

Again, only Corel OCR Trace and Omnipage were tested. The other softwares do not work properly with greyscale or colored images. These images have very particular characteristics which make them extremely special ones. Because of the age, some of these documents are very degradated. The paper of the late 19th century is particularly fragile due to the excess of acid used in the chemical process of its whitening and, in some cases, the ink has faded.

For the four images tested here, table 7 shows the average hit rate of Omnipage and Corel OCR Trace.

Image	*Corel OCR Trace*	*Omnipage*
d0023	87.026%	81.803%
d0064	91.677%	88.687%
d0077	89.612%	87.785%
d0097	98.71%	98.46%

Table 7. Hit rate for historical typographic documents

Corel OCR Trace has reached the better result getting in the best case a hit rate of more than 98%. Although, with a larger sampling space the error rate of this software was about 20% which is unacceptable for the Project's purpose. This has given the motivation to the Nabuco Project to develop a new OCR software which can reach the results expected for the project.

3 Error Analysis

It is important to observe that the comparison between the OCR response and the original text is not as simple as it seems. In the brightness experiment, with the differences of brigthness values, several kind of errors were present in the OCRs' outputs. A non-automatic process of comparison was unacceptable due to its cost and low confidence rate. An automatic system, however, will have to deal with all possible errors to generate a correct response. We have detected in our experiments the following classes of errors in the output texts:

1. Substitution of one character for another (as "*m*" for "*rn*")
2. Substitution of one character for more than one character (as "*rn*" for "*m*")
3. Substitution of more than one character for just one
4. Loss of characters (supression)

[2] Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil, Brazilian ambassador to London (b.1861-d.1910)

5. Merge of words with no loss of characters (white spaces supression)
6. Merge of words with loss of characters
7. Loss of complete lines of text
8. Insertion of characters
9. Noise

Item 1 above is the most simple error to detect. It does not cause any changes in the original word length; just a substitution of non-blank characters. We must observe that we are not working here with the correction of the word (this could easily be done with a dictionary). Our interest is in the error detection. As it is the most simple error to detect, it is also the most desirable. When Brightness is set around the value that generated the best OCR's responses (table 5), this is the most common error, followed by the inclusion of characters (due to the noise of the documents, in general). Smartpage and WordLinx were the softwares that had the higher rates of change of characters error reaching more than 60% of their errors from this class. The other softwares had a more accentuated division of the error between all the classes, observing the growing of the inclusion of characters at higher brightness sets.

The worst case to detect is the deletion of one (or more than one) complete line of text. The system must decide if there was really an exclusion or if the line was erroneously recognized in all of its characters. Although this is the most difficult class of error to detect this is also the rarest one. It was detected just in low brightness sets and only with Smartpage and Wordlinx.

The merge of two words with loss or not of some character is also a very commom error widely found in lower or higher brightness settings. Amongst these two classes, the most commom is the merge of two words without the loss of any character.

As we mentioned previously, we are only interested in the detection of errors generated by the OCR tool for comparison between its output and the original file. No attempt in error-correction was made.

4 Conclusions

Although optical character recognition is applied in a very large number of commercial softwares it has not reached a very satisfactory point yet. Many variations of several algorithms are present in these softwares, and some others are being tested. Another point to be considered is that some of these softwares are better suited to certain types of images, whereas others produce better results when dealing with other types of images. In this paper, we have shown that Omnipage from Caere Corp. has reached the best recognition rates in almost all the cases analyzed. This performance was measured relatively to the software's characteristics and the results obtained with low resolution of digitization and different degrees of rotation. It was the most complete of the tested softwares.

However, when used in one of the images tested the software did not perform well. In that case Corel OCR Trace provided better results. It is noteworthy to mention that this image was very different than the others used to assess the software's performance. The image extracted from the Internet has a clear background and a good definition of background and foreground color. It does not happen when we are working with historical documents where there is a very subtle difference between these frequency levels because of the presence of Gibbs frequencies [3]. This is just one of the reasons that led us to the development of a new OCR software. It is not just the fact that we can work with historical images but we will also have a dictionary based on the portuguese language of the begining of the century which can be modified by the user any time.

References

[1] I. Aleksander e H. Morton. *An Introduction to Neural Computing.* International Thomson Computer Press, 1995.

[2] R. Gonzalez and P. Wintz. *Digital Image Processing.* Addison Wesley Publishing Company, 1987.

[3] S. Haykin. *Neural Networks A Comprehensive Foundation.* IEEE Press, 1994.

[4] A.C. Kak e M. Slaney. *Principles of Computerized Tomography.* IEEE Press, 1988.

[5] L.R. França Neto, C.A.B. Mello and R.D. Lins. *Filtering Techniques for Digital Images of Historical Documents.* XV Brazilian Symposium of Telecommunications, Recife, Brazil,September, 1997.

[6] R.D. Lins, M.S. Guimarães Neto, L.R. França Neto and L.G. Rosa. *An Environment for Processing Images of Historical Documents.* Microprocessing & Microprogramming, pp. 111-121,North-Holland, January, 1995.

[7] C.A.B.Mello, L.R.França Neto e R.D.Lins. *A New Technique for Compressing Static Images.* Proceedings of the XVI Computation Brazilian Society Congress, Brazil, 1996, pages 37-48.

[8] C.A.B.Mello, L.R.França Neto e R.D.Lins. *A New Algorithm for Monochromatic Image Compression.* 23rd Euromicro Conference, Budapest, Hungria, 1997.

[9] J.R. Parker. *Algorithm for Image Processing and Computer Vision.* John Wiley and Sons, 1997.

[10] K.Sayood. *Introduction to Data Compression.* Morgan Kauffman Publishers, Inc., 1996.

[11] I.H. Witten, A.Moffat and T.C. Bell. *Managing Gigabytes - Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, 1994.

[12] *Adobe Systems Inc. URL: http://www.adobe.com*

[13] *Caere Inc. URL: http://www.caere.com*

[14] *Chamex. URL: http://www.chamex.com.br*

[15] *Corel Corp. URL http://www.corel.com*

[16] *Jasc Software. URL: http://www.jasc.com*

[17] *Recognita Corp. URL: http://www.recognita.com*

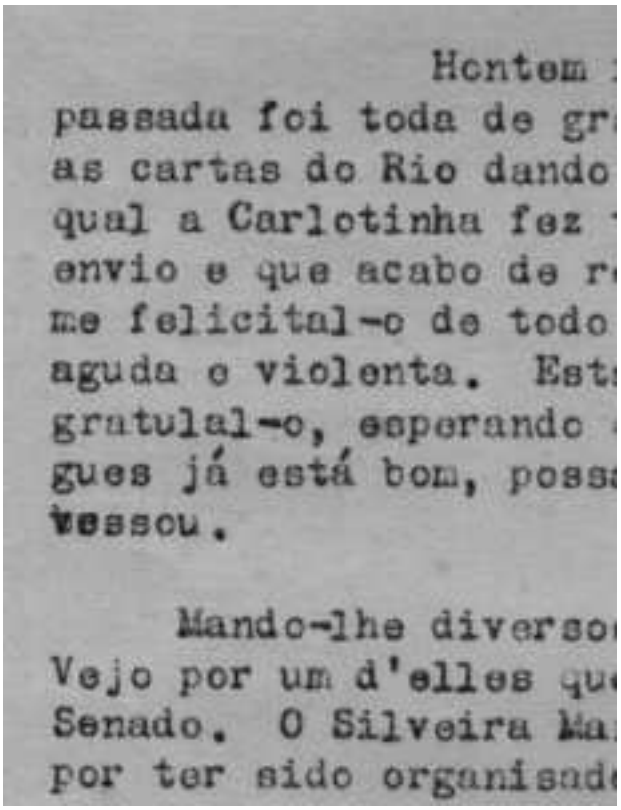
[18] *OCRON Inc. URL: http://www.newsoftinc.com*

[19] *Xerox. URL: http://www.xerox.com*

[20] *Expervision. URL: http://www.expervision.com*

[21] *Nabuco Project. URL: http://www.di.ufpe.br/~nabuco.*

Appendix A
Zooming into letter d0023 from Joaquim Nabuco's bequest


Hontem
passada foi toda de gr
as cartas do Rio dando
qual a Carlotinha fez
envio e que acabo de r
me felicital-o de todo
aguda e violenta. Est
gratulal-o, esperando
gues já está bom, poss
vessou.

Mando-lhe diverso
Vejo por um d'elles qu
Senado. O Silveira Ma
por ter sido organisad