DOI: http://dx.doi.org/10.5769/C2011012

OCR errors and their effects on computer forensics

Mateus de Castro Polastro[a,b], and Nalvo F. Almeida[c]
(a)Department of Electrical Engineering, University of Brasilia, Brasilia/DF, Brazil
(b) Technical-Scientific Sector, Brazilian Federal Police, Campo Grande/MS, Brazil
(c) College of Computing, Federal University of Mato Grosso do Sul, Campo Grande/MS, Brazil

Abstract — The use of Optical Character Recognition (OCR) technology is an alternative when it is desired to search by keywords in image documents. In the field of computer forensics, this technology was recently incorporated into the version 3.1 of Access Data Forensic Toolkit (FTK). In this paper, we propose a method to evaluate the effects of OCR errors on information retrieval in Portuguese and English texts using this FTK feature. The method is described in detail and tools and public data were used. The experiments results showed that keywords search hits in OCRed texts are directly affected by the type of degradation suffered by the images. Success rates in searches of the English texts were around 95% and below 80% in Portuguese texts.

Keywords — computer forensics; OCR; image degradation; keyword search; FTK.

1. Introduction

The concept of paperless office began in 1975 [1] and was intended to reduce or eliminate the use of paper in offices. That did not happen as expected and today still remains a myth [2], since the use of paper continues to be intense. Several factors contribute to the continued use of paper. For example, the ease that people have to read and understand printed text and low-cost and increase quality of printers and photocopiers [3] has lead people to print digital documents. On the other hand, other factors lead to the need for the use of documents in digital format. The cost to physically store paper documents is very high compared to digital documents, in addition to difficulty of retrieving information stored in them [2]. Printed documents have their access restricted to where they are, while digital documents have a great facility of movement. Digital documents can be transferred to various places on the planet with a simple mouse click. Also, while a printed paper can be used only by one person at a time [2], a digital document can be easily replicated to other people using the Internet.

With the increased use of computers, easy Internet access and low prices of devices for scanning documents, the presence of scanned documents is becoming more common. Although many documents are born digital, they are often printed at any given time and then digitized to be stored or transferred to other locations. This movement brings with it a problem for information retrieval. A document, when printed and scanned, becomes stored in image format, making it impossible to perform keyword search without the application of optical character recognition (OCR).

This global trend is also perceived in computer forensics that works primarily with computer equipment from government agencies, businesses and residences. One of the techniques used in computer forensics that helps in the treatment of large amounts of data is the use of keywords search, because it allows information of interest to be found quickly. However, in cases of printed documents that are digitized again, these searches are inefficient because they are not able to find text in images. Thus, this work is often done manually by looking image by image.

In 2010, AccessData FTK (Forensic Toolkit) [4], one of the most used programs in the computer forensics field brought into version 3.1 the ability to recognize characters in images documents via OCR. However, to the best of our knowledge, there was no work in the literature to evaluate the efficiency of this functionality of FTK. Thus, this study aims to answer the following questions:

- Is it possible to perform keyword search in images using OCR technology on computer forensics?
- How does keyword search in images behave with different types of images degradation?
- Does the OCR functionality of the FTK provide equivalent results for different language texts?

This paper is organized as follows: section 2 presents some important aspects related to the OCR technology, some image defects and how this technology is inserted into the field of computer forensics. The description of the experiment is detailed in Section 3. Section 4 presents and analyzes the results. The conclusion and future work are presented in section 5.

2. Background

Important aspects related to OCR, as well as the use of this technology in the field of computer forensics will be discussed below.

2.1. OCR and its Accuracy

Optical Character Recognition or simply OCR, is the process of automatic conversion of characters present in the images to text format. The OCR has emerged in the early '50s and has been the subject of many studies and has advanced greatly in performance and accuracy [5].

OCR process can be summarized in four steps [6]. The first step is the *acquisition*, which is the act of digitize the printed

document, usually done through a scanner. The second is the *image processing*, which will analyze aspects like skew, image segmentation, filtering, among others, depending on the program used for OCR. The third step is the *characters identification*, where the image obtained in previous step is analyzed, usually through artificial intelligence techniques. The last step is the *output file generation*. Depending on the output format (txt, rtf, etc.), figures and text fonts identified in the previous steps can be reproduced.

The *acquisition* stage is largely dependent on the user and when can be inserted various image defects. Some image defects commonly inserted during the acquisition phase are:

- Skew: while scanning, the document can be positioned at an angle different than zero degree on the scanner optical reader, creating a skewed image that can compromise the OCR result. Some OCR tools are able to detect and correct certain types of skews without causing damage in the detection process [7];
- Salt-and-pepper noise: it consists in the presence of white pixels (salt) in dark regions and dark pixels (black) in bright regions, which may occur due to variation of the surface of the material or lighting and also in the process of converting analog to digital [8], as occurs in the scanners. For example, an image that has such a degradation of 1% means that 1% of the image pixels are randomly converted to black or white ones;
- Low Resolution: in general, higher resolutions provide greater accuracy. However, according to a study [9], resolutions higher than 300 dpi will not cause relevant improvements in accuracy rates.

To illustrate these image defects, Figures 1 (a), 1 (b) and 1 (c) display the same text applying these three different defects.

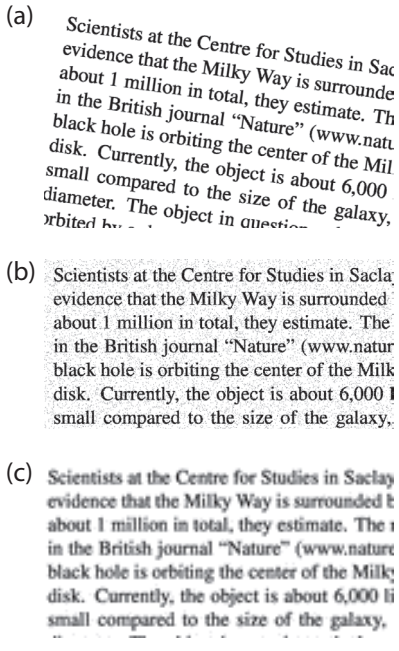

Fig. 1. (a) skewed image; (b) image with salt-and-pepper noise added; (c) image with low resolution.

In order to evaluate the accuracy of OCR tools under various types of images, many experiments were conducted in the past. The measure of accuracy of OCR tools can be done in various ways, such as calculating the number of errors per page, number of errors in words, total number of errors, among other options. To be able to carry out these measures, it is first necessary to compare the original text with the OCRed text, and then count the differences between them. One of the ways to accomplish this task is by calculating the edit distance between the texts. The edit distance is a measure that counts the number of operations (insertion, deletion or replacement) required to transform one string into another [10].

Between 1992 and 1996 the Information Science Research Institute at the University of Nevada, Las Vegas, USA, held an annual series of tests to verify the accuracy of commercial OCR tools and research prototypes available at the time. In many experiments, the accurate recognition of characters was about 99%. Although it is a high success rate, it still can greatly affect the retrieval of text information. For illustration, if a page contains 3000 characters, words within an average size of 6 characters and all errors occurring in different words, this rate implies 30 misspelled words in a single page.

Another study [7] evaluated how defects applied to image, such as skew and downsampling, can influence error rates in OCR. One of the programs evaluated kept high rates of accuracy when the documents were skewed for up to 12 degrees. However, in another program evaluated, the precision was more than 99% to a document when skewed by 1 degree and dropped to 33% when skewed 3 degrees. For different resolutions ranging from 75 dpi to 250 dpi, six tools were evaluated for accuracy. For three of these tools, the variation in accuracy, considering the correct recognition of words, when the resolution was decreased from 250 to 150 dpi, was small (about 2%). However, it fell dramatically to resolutions of 100 dpi and 75 dpi, reaching up to 0% detection.

Such behavior could also be observed in another study [9]. In the experiments, there was a substantial increase in the number of errors when the resolution has dropped from 300 dpi to 200 dpi. On the other hand, the increased resolution of 400 dpi showed practically no change in errors rates.

2.2. OCR and Computer Forensics

The use of OCR in the computer forensics is very recent and was first observed on a large scale when the Access Data released in the first half of 2010 the version 3.1 of Access Data FTK [4]. This version brought the functionality of optical character recognition built into the tool. According to [11], the feature works as follows *"Scans graphics files for text and converts graphics-text into actual text. That text can then be indexed, searched and treated as any other text in the case"*.

One of the factors that influence the error rates in optical character recognition is the OCR software used. The FTK has embedded to him the Tesseract OCR [11]. It is a program developed by Hewlett-Packard (HP) between 1984 and

1994 but never exploited commercially [12]. In 1995, it was very well qualified in the tests conducted by the University of Nevada, USA [10]. Currently the software is in version 3 and supports several languages (including English and Portuguese).

3. Experiments

In order to assess the accuracy of FTK OCR keyword search functionality, we first made experiments to verify how FTK can deal with different languages, such as Portuguese and English. The experiments aim to verify how the keyword search in images documents behaves using the OCR functionality of the FTK software for both texts in Portuguese and in English. The intention is to unveil how many of the searched words are returned by the tool and compare it with ideal, which would be 100% match.

To do this, searches were carried out by known keywords in several sets of images, each containing a different degradation. The degradations selected to be applied to the images in this experiment were: downsampling, skews and salt-and-pepper noise. The version 3.2 of FTK was used in the experiments.

This experiment ranges from the generation of documents to be scanned until verification of the results of the keyword search in FTK. An overview of this process is illustrated in Figure 2. In the following subsections each of these steps will be detailed.

3.1. Document Generation

The documents were created from simple text files, obtained from the corpus "Summ-it" [13]. This corpus consists of 50 newspaper articles in Portuguese from the Science supplement of "Folha de São Paulo" newspaper. Ten files were chosen from this corpus to perform the experiments. English texts were obtained from these ten text files after automatic translation done by Google Translate [14]. Table 1 displays the file name from the original corpus Summ-it, the number of characters in each one of them in Portuguese and English and the total number of characters for each language.

The set was formatted in 12 pts "Times New Roman" using LaTeX [15]. By running the program *pdflatex*, these files were converted to PDF format, resulting in twenty documents of just one page each one.

All pdf documents were printed in a Hewlett-Packard Laserjet 3800 DTN printer with resolution of 300 dpi and printed in A4 white paper with 75 g/m2 and 210x297mm. Then these printed texts were carefully scanned in order to avoid the inclusion of noise and skew, using a Hewlett-Packard Scanjet 8270 with a resolution of 300 dpi, generating TIFF images.

Table 1. Selected files from the Corpus Summit-it.

File Name from Corpus	Number of Characters	
	Portuguese	English (translated)
CIENCIA_2001_19858.txt	2921	2822
CIENCIA_2002_22005.txt	2841	2724
CIENCIA_2002_22023.txt	2406	2197
CIENCIA_2002_22027.txt	2952	2720
CIENCIA_2003_24212.txt	2896	2733
CIENCIA_2004_26417.txt	2748	2591
CIENCIA_2005_28743.txt	2632	2468
CIENCIA_2005_28752.txt	2778	2630
CIENCIA_2005_28754.txt	2838	2665
CIENCIA_2005_28756.txt	2790	2620
Total Number of Characters	**27802**	**26170**

3.2. Image Degradation

The images created in the previous step are considered high-quality images for use in optical character recognition, because there are no page layout variations, only with known font set, not skewed, and with a resolution considered optimal for use in OCRs.

Because most of the images in computer forensic examinations do not have such good quality, a series of degradation has been applied to them, to simulate a behavior closer to the real world. The degradations were applied separately in the images. One of the advantages of the synthetic images degradation is the ability to keep them under control [16], changing according to the goals of the experiment.

An image processing software called *imagemagick* [17] was used to apply resolution and skew degrade to the images. To apply salt-and-pepper degradation was used the GNU Octave [18] *imnoise* command. Table 2 shows the types of degradations applied, the *imagemagick* and Octave commands used and the name of each set created, for further reference.

Fig. 2. Overview of the experiment.

Table 2. Types of degradation applied to images.

Degradation	Command-line	Group Name
Resolution 200 dpi	convert -resample 200 -depth 8 -type Grayscale [pdf_file] [tif_file]	Res200
Resolution 150 dpi	convert -resample 150 -depth 8 -type Grayscale [pdf_file] [tif_file]	Res150
Skew +3	convert –rotate 3 -depth 8 -type Grayscale [pdf_file] [tif_file]	Skw3
Skew +6	convert –rotate 6 -depth 8 -type Grayscale [pdf_file] [tif_file]	Skw6
Salt-and-pepper 1%	imnoise([tif_file],'salt & pepper', 0.01)	S&P1
Salt-and-pepper 3%	imnoise([tif_file],'salt & pepper', 0.03)	S&P3

3.3. Keyword Selection

The texts in Portuguese and English created in Section 3.1 (Document Generation) were processed to identify all the words that occur in them. In order to do this separation, were considered tokens the following characters: "[space] . , ; : - " ( ) [ ]". Thereafter, the words that appear only once in each set of texts were identified, which were separated into two groups: words with 5 or more characters and less or equal 7, and words with 8 or more characters.

For Portuguese texts, 451 unique words were identified between 5 and 7 characters and 561 words greater than 7 characters. For English texts were identified 398 and 368 words, respectively. These words were then saved in two text files, encoded in UTF-8, with one word per line.

In preliminary experiments, it was found that the OCR functionality used by FTK does not work well with accented words typically found in the texts of the Portuguese language. Thus was created one more set of tests for the category of texts in Portuguese. This set consists of all words in Portuguese already created, except those with at least one of the following characters in its formation: "á ã â é ê í ó ô õ ú ü ç". Table 3 shows the keywords files created, along with the number of words contained in each one.

Table 3. Keywords files created.

File Description	Statistics (number of words)	File ID
Portuguese with accents (words 5 and 7 chars)	451	pt1
Portuguese with accents (words 8 chars)	561	pt2
Portuguese without accents (words 5 and 7 chars)	375	pt3
Portuguese without accents (words 8 chars)	424	pt4
English (words 5 and 7 chars)	398	en1
English (words 8 chars)	368	en2

3.4. FTK OCR Search

The stage of Image Degradation generated 12 sets of images (6 for each language), each one containing 10 documents images. For each of these sets was created a folder. Then a new FTK case was created for each folder, using the default settings of the FTK and enabling the conversion of images to text using OCR. All limitations on the minimum and maximum size of image files to be converted were removed.

After processing the case by FTK, was performed an indexed search using the words from the files listed in Table 3. To perform the search, the keywords file was imported in the appropriate language and then searches were conducted in all OCRed files.

All searches were carried out using exact search, i.e., strings that had any difference with the searched keyword were not considered a match.

3.5. Results Validation

In order to validate the results, a comparison between the number of hits returned by the indexed search against the number of words in the imported file was made. If the OCR result was 100% accurate, these numbers should be equal. However, with the occurrence of OCR errors, the exact search returns fewer hits. Let be the number of keywords to search and the number of hits obtained, the hit rate is calculated as follows:

$$h/k \qquad (I)$$

Table 4 identifies groups of documents images that were used in the experiment, with their main characteristics. For the group labeled "ptDOC" in Table 4 were both carried out keywords searches using accented words and non-accented words (see Table 3). Figure 3 illustrates the entire proposed method, identifying each of the steps described above and their interactions.

4. Results and Discussions

The results obtained after applying the proposed method was summarized in four charts. Figures 4 and 5 are related to documents in Portuguese and the other two to the documents in English. The charts show the percentage of

Table 4. Identification of the groups used in the experiment.

Group	Subgroup	Subgroup Name	Keywords Files Used (from Table 3)
Documents in Portuguese (ptDOC)	Resolution 200 dpi	Res200	pt1, pt2, pt3, pt4
	Resolution 150 dpi	Res150	pt1, pt2, pt3, pt4
	Skew +3	Skw3	pt1, pt2, pt3, pt4
	Skew +6	Skw6	pt1, pt2, pt3, pt4
	Salt-and-pepper 1%	S&P1	pt1, pt2, pt3, pt4
	Salt-and-pepper 3%	S&P3	pt1, pt2, pt3, pt4
Documents in English (enDOC)	Resolution 200 dpi	Res200	en1, en2
	Resolution 150 dpi	Res150	en1, en2
	Skew +3	Skw3	en1, en2
	Skew +6	Skw6	en1, en2
	Salt-and-pepper 1%	S&P1	en1, en2
	Salt-and-pepper 3%	S&P3	en1, en2

searched keywords that were found in the OCRed texts, based on equation (I). In all charts the images noises were presented in increasing order of success.

Through analysis of Figures 4 and 5 it is possible to identify that the success rate from non-accented words was considerably higher. The analysis of the OCRed texts generated by FTK led to the conclusion that in most cases the accent was ignored or caused confusion in the recognition process, resulting in errors, as can be seen, for example, in exchange for the letter "ç" for "e" several times.

For the group of documents in English, the variation in the word length influenced the success rate for different types of image degradation, as seen in Figures 6 and 7. In English documents the salt-and-pepper 3% degradation significantly decreased the success of searches compared to others.

Although it was expected that applying a higher skew would result in worse results for searches, it did not happen. The causes of this behavior were not identified, but may be related to the OCR software's ability to skew detection and correction.

In general the search results for texts in English were considerably more accurate than those applied to Portuguese texts. While the average number of hits for the texts in Portuguese was below 80%, in English texts the rate leaped to about 95%.

The developed method may lead to a small degree of error since some words can be recognized by the OCR in the wrong way, creating another word that is also on the list of keywords. While this can happen, the chance is very small and does not invalidate the achieved results.

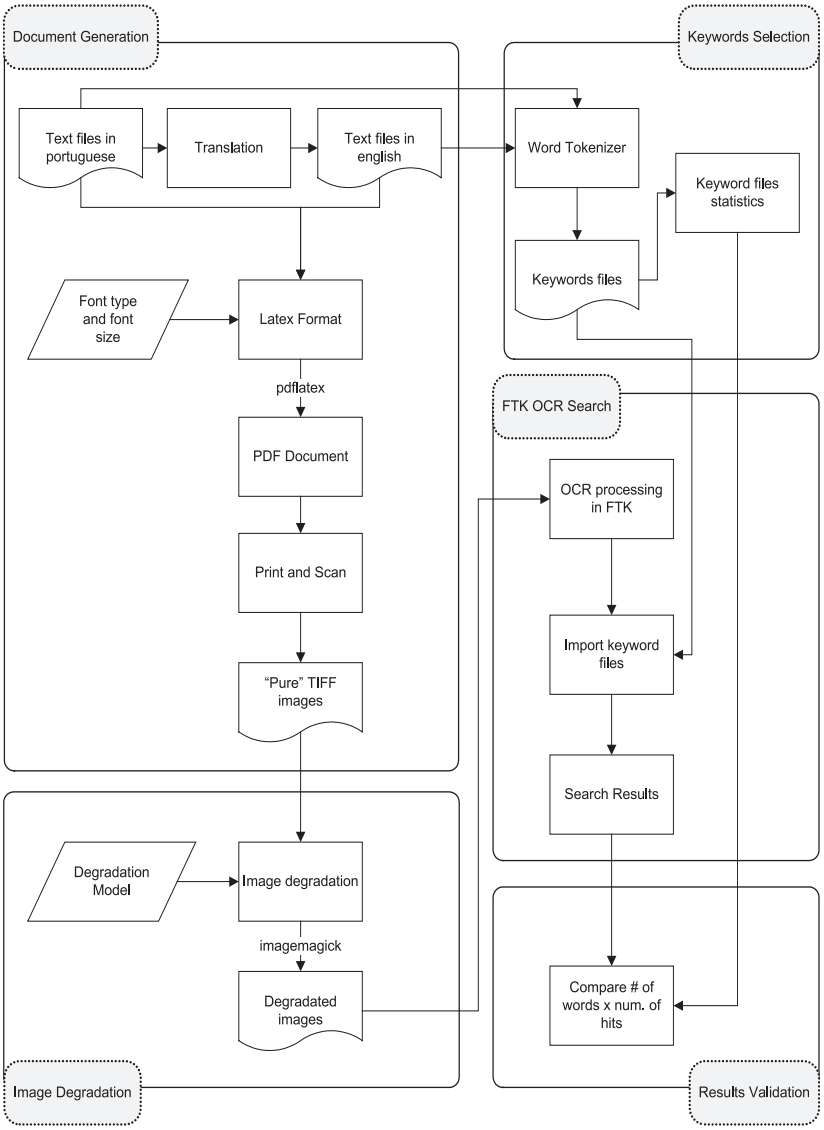

Fig. 3. Proposed method for OCR errors evaluation on keywords searches in images.

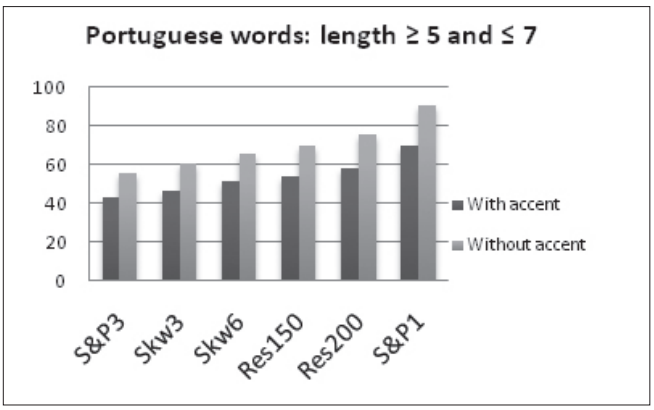

Fig. 4. Search results for keywords in Portuguese (with and without accent) with length ≥ 5 and ≤ 7.

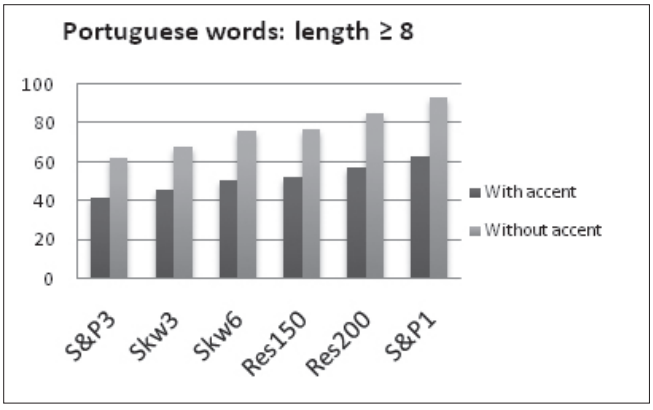

Fig. 5. Search results for keywords in Portuguese (with and without accent) with length ≥ 8.

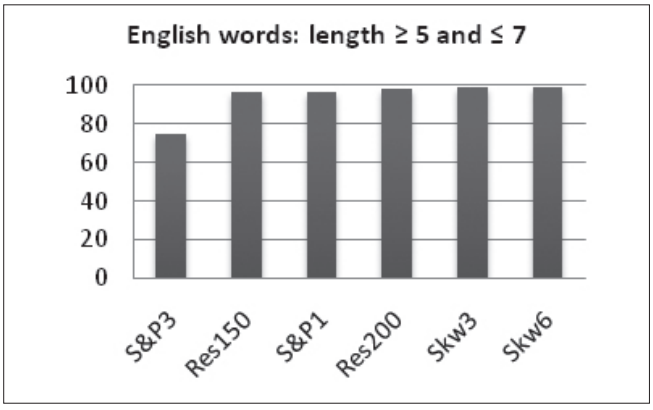

Fig. 6. Search results for keywords in English with length ≥ 5 and ≤ 7.

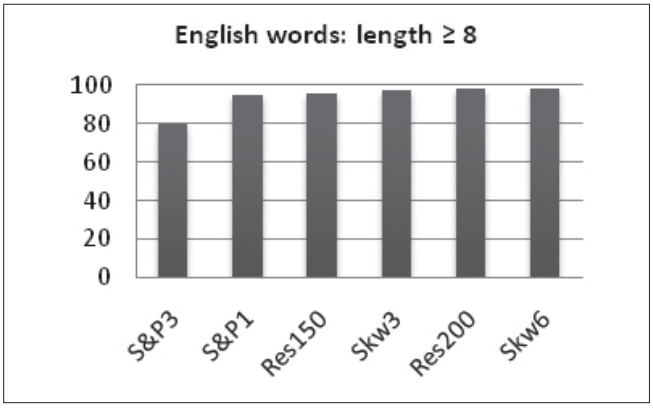

Fig. 7. Search results for keywords in English with length ≥ 8.

5. Conclusion and Future Work

The need for tools to automate part of the computer forensics experts work is increasingly and OCR functionality built into one of the most used tools in the world is certainly a breakthrough.

In this paper, we proposed a method to evaluate the effects of OCR errors on information retrieval in the context of computer forensics. The method was described in detail and tools and public data were used, allowing the experiment to be reapplied to other versions of OCR, with another set of images, another language or even another software.

The experiments showed that the search results for English texts achieved high success rates. However, the same was not observed for the texts in Portuguese. Possibly the reason for this large difference in success between the two languages is the fact that the FTK is not prepared to use Tesseract OCR on Portuguese texts, although Tesseract already supports Portuguese texts.

Image degradation, in particular the application of salt-and-pepper 3% noise, had a great influence on the results obtained. Since it is impossible to guarantee the quality of images that computer forensics experts receive during the digital storage media exams, the image degradation can be a great villain in the use of OCR technology. Thus, it is necessary to create mechanisms to improve search results. One way would be to implement approximate search algorithms that consider the most common OCR errors in order to ensure that important documents can be retrieved. Although the FTK has the fuzzy search feature, this is not the ideal solution, because the common OCR errors are not taken into account, which can lead to many false positives.

Identify the most common errors that occur using the OCR functionality of FTK is a theme that can be examined in depth in future work.

References

[1] Business Week. (1975, June) www.businessweek.com. Available at http://www.businessweek.com/print/technology/content/may2008/tc2080526_547942.htm. Accessed in July, 2011.

[2] A. J. Sellen and R. H.R. Harper, The Myth of the Paperless Office, MIT Press Cambridge, MA, USA, 2003. ISBN:026269283X.

[3] V. Hardy and H. Wu and Y. Zhang and L. E. Dyson. Paper Usage Management and Information Technology: An Environmental Case Study at an Australian University. Proceedings of The 5th International Business Information Management Association Conference. Cairo, Egypt, 2005.

[4] Access Data Forensic ToolKit (FTK), avaiable http://www.accessdata.com. Accessed in June, 2011.

[5] H. Fujisawa. A view on the past and future of character and document recognition. In Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 01, pages 3–7, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2822-8

[6] G. Nagy, T. A. Nartker, S. V. Rice. Optical character recognition: An illustrated guide to the frontier. In Storage and Retrieval for Image and Video Databases, 1999.

[7] C. A. B. Mello and R. D. Lins. A Comparative Study on OCR Tools. Proceedings of Vision Interface 99, pp. 700--704, Québec, Canada, May, 1999.

[8] L. G. Shapiro, G. C. Stockman. Computer Vision. Prentice Hall, january 2001. ISBN 0130307963.
[9] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fifth annual test of ocr accuracy. 1996.
[10] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fourth annual test of ocr accuracy. 1995.
[11] Access Data. FTK 3.2 User Guide. John Wiley & Sons, 2010.
[12] R. Smith. An overview of the tesseract ocr engine. In International Conference on Document Analysis and Recognition, pages 629-633, 2007.
[13] Corpus Summit-it. Avaiable at http://www.nilc.icmc.usp.br:8180/portal/news.jsp?id=6. Accessed in July, 2011.
[14] Google Translate. Avaiable at http://www.translate.google.com. Accessed in July, 2011.
[15] LaTex. Avaiable at http://www.latex-project.org. Accessed in June, 2011.
[16] T. Kanungo. Document degradation models and a methodology for degradation model validation. PhD thesis, Seattle, WA, USA, 1996. UMI Order No. GAX96-30083.
[17] ImageMagick. Avaiable at http://www.imagemagick.org. Accessed in July, 2011.
[18] GNU Octave. Avaiable at http://www.gnu.org/software/octave/. Accessed in June, 2011.

Mateus de Castro Polastro obtained his bachelor's degree in Computer Science at the University of Campinas (Unicamp), Brazil. He is currently enrolled in the M.Sc. program in Computer Forensics at the University of Brasilia (UnB), Brazil. Since 2007, he has worked as a criminal forensic expert for the Brazilian Federal Police (DPF).

Nalvo F. Almeida earned bachelor's degree in Math at Universidade Federal de Mato Grosso do Sul (1985), master's degree in Computer Science at Federal University of Rio de Janeiro (1992) and Ph.D. in Bioinformatics at University of Campinas (2002). He is Associate Professor at College of Computing, Federal University of Mato Grosso do Sul since 1987 and spent two years as visiting scholar at Virginia Bioinformatics Institute, in Blacksburg, USA. His interest areas are Bioinformatics and Theory of Computing.