Image Restoration and Metadata Extraction of Ancient Documents

Henrique Salvador Vieira Mata

Thesis to obtain the Master of Science Degree in

Mechanical Engineering

Supervisor: Prof. João Rogério Caldas Pinto

Examination Committee
Chairperson: Prof. Viriato Sérgio de Almeida Semião
Supervisor: Prof. João Rogério Caldas Pinto
Members of the Committee: Dr. Pedro Miguel Berardo Duarte Pina
Dra. Maria do Carmo Sequeira Nunes

November 2014
To improve is to change; to be perfect is to change often – Winston Churchill
Acknowledgements

I would like to express my profound gratitude to all of those who supported me in the accomplishment of this thesis. To Prof. João Rogério Caldas Pinto, for his insight in different areas of this work, for his dedication and thoughtful guidance. To Dr. Carmo Nunes, for her support and for creating the possibility to work with the Tropical Research Institute.

To Fundação para a Ciência e a Tecnologia (FCT) for granting a scholarship, in the ambit of the project PTDC/ATP-GEO/4645/201, entitled by “Geospatial Technologies on Demarcation and Management of Guinea-Bissau Boundaries”, which represented a great incentive during my investigation.

Last but certainly not least, to my family and friends, for their loving support and for providing the environment that encouraged me to pursue my goals and go one step further, with this thesis.
Abstract

One of the current goals of the owners of cultural heritage, in particular libraries, is the disclosure of their portfolios. This translates into the creation of a digital version of these documents. This process will enable, through the creation of on-line digital libraries, public to have contact with rare and ancient documents that would have been otherwise unavailable.

One common feature to all of the digitized images is that they reveal the deteriorations present in the original documents. Most of these deteriorations result from the aging process, especially if poor care was taken to store and preserve the documents over time. Thus, when handling the images of those documents, it is therefore desired to have these images restored in order to improve image segmentation and recognition operations. Moreover, restoration is also desirable in order to revamp their visual appearance.

Besides the disclosure of library contents, efforts should also focus on technologies aimed at reducing the human effort required for the annotation of the raw images with informative content. This motivates the creation of an automatic and efficient recognition system.

This thesis encompasses the subject of image processing with the goal of researching approaches to effectively restore the digital version of ancient typed documents, in order to improve OCR conditions, and their visual appearance, and the subject of metadata extraction.

The developed methods are flexible in order to handle the varying conditions found in institutions like libraries, thus providing a good restoration quality in the majority of the cases of, and retrieving most of the artifacts present in them.

Keywords:

Ancient Documents, OCR, Feature Extraction, Descriptive Metadata
Resumo

Um dos actuais objectivos das instituições responsáveis pela preservação de património cultural, em particular das bibliotecas, consiste na divulgação dos conteúdos dos seus arquivos. Esta divulgação passa pela criação de uma versão digital dos documentos que constituem estes arquivos. Este processo permitirá, através da criação de bibliotecas on line, o contacto do público com documentos antigos, que, de outra forma, seriam inacessíveis.

Uma particularidade comum a todas as imagens digitalizadas consiste na preservação da degradação presente nos documentos originais. A maioria destes sinais resulta do processo de envelhecimento, sobretudo nos casos em que poucos cuidados foram tomados no armazenamento e na preservação ao longo do tempo. Assim, ao lidar com estes documentos, é desejado que estes sejam restaurados por forma a melhorar as operações de segmentação e reconhecimento. O restauro é simultaneamente necessário para melhorar a aparência visual dos documentos.

Esta tese aborda a temática do processamento de imagem com o objectivo de pesquisar abordagens que, efectivamente, restaurem a versão digitalizada de documentos dactilografados, por forma a melhorar as condições de OCR e aparência visual, e a temática do reconhecimento de artefactos. Os métodos desenvolvidos são flexíveis por forma a lidarem com as várias condições encontradas nas instituições, oferecendo um restauro de boa qualidade na maioria dos documentos e na detecção da maioria dos artefactos presentes nestes.

Palavras-chave:

Documentos Antigos, OCR, Extracção de Características, Metadados
# List of Contents

Acknowledgements ........................................................................................................................................... i  
Abstract ......................................................................................................................................................... ii  
Resumo ............................................................................................................................................................. iii  
List of Figures .................................................................................................................................................. vi  
List of Tables ................................................................................................................................................ vii  
Abbreviations ................................................................................................................................................ viii  

1 Introduction .................................................................................................................................................... 1  
  1.1 Studied Documents .................................................................................................................................. 2  
    1.1.1 Types of documents degradation ...................................................................................................... 5  
  1.2 Objectives ................................................................................................................................................ 7  
  1.3 Main contributions .................................................................................................................................... 7  
  1.4 Organization ............................................................................................................................................ 7  

2 Pre-processing Image Techniques .................................................................................................................. 8  
  2.1 Resizing ................................................................................................................................................... 8  
  2.2 Thresholding .......................................................................................................................................... 9  
    2.2.1 Methods ............................................................................................................................................ 10  
  2.3 Deblurring ............................................................................................................................................. 13  
    2.3.1 Methods ............................................................................................................................................ 15  

3 Improving Optical Character Recognition .................................................................................................... 20  
  3.1 Evaluation Methodology ............................................................................................................................ 20  
  3.2 Statistical Measures .................................................................................................................................. 21  
  3.3 Results ...................................................................................................................................................... 23  
    3.3.1 Resizing ............................................................................................................................................. 23  
    3.3.2 Thresholding ....................................................................................................................................... 25  
    3.3.3 Deblurring .......................................................................................................................................... 32  
  3.4 Discussion of Results – OCR preprocessing improvement ........................................................................ 37  

4 Metadata Extraction ......................................................................................................................................... 38  
  4.1 Artifact Recognition Algorithm ................................................................................................................ 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1.1 Color Channel Conversion</td>
<td>41</td>
</tr>
<tr>
<td>4.1.2 Feature extraction</td>
<td>44</td>
</tr>
<tr>
<td>4.1.3 Morphological Reconstruction (MR)</td>
<td>49</td>
</tr>
<tr>
<td>4.1.4 Residual Area removing</td>
<td>53</td>
</tr>
<tr>
<td>4.1.5 Area Thresholding</td>
<td>53</td>
</tr>
<tr>
<td>4.2 Discussion of Results – Artifact Recognition</td>
<td>54</td>
</tr>
<tr>
<td>5 Conclusions and Future Work</td>
<td>56</td>
</tr>
<tr>
<td>Bibliography</td>
<td>58</td>
</tr>
<tr>
<td>6 Appendix</td>
<td>62</td>
</tr>
</tbody>
</table>
List of Figures

Fig. 1-1 - a) example of image from group 1; b) example of image from group 2; c) zoomed view of image from group 1 example; d) zoomed view of image from group 2 example ............................................. 3

Fig. 1-2 - a) example of image from group 3; b) example of image containing artifacts; c) zoomed view of image from group 3 example; d) zoomed view of image containing artifacts ........................................... 4

Fig. 1-3 - a), b) foreground degradation – blurring; c) background degradation ........................................... 6

Fig. 2-1 - Cubic kernel. From Digital Image Processing using Matlab (p. 300), by Gonzalez, 2009, Gatesmark Publishing ............................................................................................................................................. 9

Fig. 2-2 - Canny+thresholding method scheme ................................................................................................. 12

Fig. 2-3 - Deblurring Model. From Digital Image Processing using Matlab (p. 300), by Gonzalez, 2009, Gatesmark Publishing ............................................................................................................................................. 13

Fig. 3-1 - a) identification of information; b) OCR result .................................................................................. 21

Fig. 3-2 - a) Niblack – Identification of information b) Niblack - OCR result ...................................................... 27

Fig. 3-3 - a) Canny+Niblack - Identification of information b) Canny+Niblack - OCR result ......................... 29

Fig. 4-1 - Metadata Extraction algorithm ........................................................................................................ 40

Fig. 4-2 - a) original image; b) black threshold; c) blue threshold; d) red threshold ......................................... 43

Fig. 4-3 - a) image containing table; b) Sauvola threshold; c) table key feature - vertical lines; d) table key feature - horizontal lines ......................................................................................................... 45

Fig. 4-4 - Image containing stamps, underlines and annotations ................................................................. 46

Fig. 4-5 - a) black (Sauvola) threshold; b) red threshold .................................................................................. 47

Fig. 4-6 - a) key features of stamps; b) key feature of underline ................................................................. 48

Fig. 4-7 - a) black artifacts removed; b) stamp recognized; c) underlines recognized; d) annotations recognized ............................................................................................................................................. 51

Fig. 4-8 - a) original image; b) image without artifacts .................................................................................. 52
List of Tables

Table 3.1 - Finereader’s performance evaluation – resizing: modern documents (group 1) ............ 23
Table 3.2 - Performance differences between the RF analysed and the 1,00 factor for modern documents .................................................................................................................. 24
Table 3.3 - Finereader’s performance evaluation – ancient typed documents (group 2) ............... 24
Table 3.4 - Performance differences between the RF analysed and the 1,00 factor for ancient typed documents (group 2) .................................................................................................................. 24
Table 3.5 - Finereader’s performance evaluation – modern documents (group 2) reduced by 0,50.... 30
Table 3.6 - Finereader’s performance evaluation – thresholding methods ........................................ 30
Table 3.7 - Performance differences between the method analysed and ancient typed documents (group 2) .................................................................................................................. 30
Table 3.8 – Candidate filters to PSF .................................................................................................. 34
Table 3.9 - Finereader’s performance evaluation – ancient typed documents (group 3) reduced by 0,50 .................................................................................................................. 36
Table 3.10 – Finereader’s performance evaluation – deblurring methods ...................................... 36
Table 3.11 - Performance differences between ancient typed documents and deblurred images ...... 36
Table 4.1 - Structuring elements – feature extraction .......................................................................... 48
Table 4.2 - Structuring elements – artifact removal .......................................................................... 50
Table 4.3 – Artifact recognition results ............................................................................................. 55
Table 6.1 - Characters classified as FP in more than 10% of the cases in modern documents .......... 62
Table 6.2 - Characters classified as TP in less than 10% of the cases in modern documents ............ 62
Table 6.3 - Characters classified as TN in more than 10% of the cases in modern documents .......... 63
Table 6.4 - Characters classified as FN in more than 25% of the cases in modern documents .......... 64
Table 6.5 - Characters classified as FP in more than 10% of the cases in ancient documents .......... 65
Table 6.6 - Characters classified as TN in more than 10% of the cases in ancient documents .......... 66
Table 6.7 - Characters classified as FN in more than 25% of the cases in ancient documents .......... 67
Table 6.8 - Characters classified as TP in less than 10% of the cases in ancient documents .......... 68
### Abbreviations

<table>
<thead>
<tr>
<th>AR</th>
<th>Artifact Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS</td>
<td>Constrained Least Squares</td>
</tr>
<tr>
<td>CR</td>
<td>Correctly recognized</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>IICT</td>
<td>Tropical Research Institute</td>
</tr>
<tr>
<td>MR</td>
<td>Morphological Reconstruction</td>
</tr>
<tr>
<td>NP</td>
<td>Noise Power</td>
</tr>
<tr>
<td>NSR</td>
<td>Noise to Signal Ratio</td>
</tr>
<tr>
<td>OCR</td>
<td>Optical Character Recognition</td>
</tr>
<tr>
<td>OTF</td>
<td>Optical Transfer Function</td>
</tr>
<tr>
<td>PSF</td>
<td>Point Spread Function</td>
</tr>
<tr>
<td>RF</td>
<td>Reduction Factor</td>
</tr>
<tr>
<td>RP</td>
<td>Recognition Process</td>
</tr>
<tr>
<td>SM</td>
<td>Statistical Measures</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
</tbody>
</table>
1 Introduction

Extensive amounts of legacy documents are being published by on-line digital libraries worldwide. However, for these raw digital images to be really useful, they need to be transcribed into a textual electronic format that would allow unrestricted indexing, browsing and querying. Thus, an undeniable goal of these institutions is to provide ASCII versions of the documents. This can be achieved by carrying out Optical Character Recognition operations (OCR).

OCR is a practical application of state-of-the-art image processing and pattern recognition developments that allows recognition of printed or written text characters by a computer. This process involves scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, such as ASCII, commonly used in data processing[1].

The OCR process has 5 main steps: pre-processing to improve the quality of the input image, thresholding, text segmentation, character recognition, and post-processing based on knowledge of the target language of the text. These steps are most effective when applied to document text, which when collected by a scanner is generally aligned and has clear contrast between text and its uniform background.

OCR of machine-print document images has matured considerably during the last decade. OCR systems have been under development in research and industry and recognition rates as high as 99.5% have been reported on 300 dpi omnifont documents[2].

Nevertheless, while modern printed text can be recognized very accurately, with commercially available software, performing OCR on more exotic material (such as gothic fonts, ancient typesets and handwriting) is currently and noticeably less successful[3].

One of the reasons for the performance losses observed when dealing with ancient typed documents is the presence of multiple signs of degradation.

Degradation can be described as “every sort of less-than-ideal properties of real document images, e.g. coarsening due to low digitizing resolution, ink/toner drop-outs and smears, thinning and thickening, geometric deformations, etc”[4]. For text recognition work it is fundamental to have recourse to well pre-processed digital images. Therefore, restoration is necessary not only to enhance the visual appearance of a document, but also to improve the results of further segmentation and recognition operations. Hence, restoration can be thought of as a transformation process that gives the original aspect to ancient document images that show a certain state of degradation.

This work aims to improve the pre-processing stage of Finereader using techniques such as resizing, thresholding and deblurring.

After OCR pre-processing improvement, the next problem addressed in this work is the extraction of descriptive metadata from handwritten and typed documents regarding the presence of artifacts. Besides the disclosure of library contents, efforts should also focus on technologies aimed at reducing the human effort required for the annotation of the raw images with informative content. Marks left by previous readers as stamps, underlines and annotations are considered relevant, for indicating the presence of noteworthy information. Stamps allow to associate documents to institutions, while
underlines and annotations highlight information may reveal the relevant content found by a previous reader. In this sense, the extraction of metadata from this type of information constitutes an ultimate goal for institutions like libraries, motivating the building of an automatic and efficient recognition system. Such system would be able to retrieve images that contain objects recognized as tables, stamps, underlines or annotations, saving the effort from an user of browsing through large sets of images stored on digital media. In this work, 3 artifacts will be addressed: (1) stamps; (2) underlines and (3) annotations. In addition, recognition of tables was also performed.

Nevertheless, the presence of artifacts may also be considered unwanted. Due to the handling of documents throughout the time, these may suffer multiple transformations and may become almost imperceptible. Stamps, underlines and annotations are features responsible for OCR inaccuracy. By clearing artifacts from the images there will be less room in the future for misinterpretation.

Feature extraction serves not only the purpose of metadata extraction but will also allow the removal of artifacts, revamping the visual appearance of ancient documents.

1.1 Studied Documents

The presented ideas will be applied to a representative set of documents from the IICT Library on themes concerning the borders of Guinea Bissau. These documents are manuscripts and typed scripts from the late nineteenth and early twentieth century. The images were digitalized using a resolution of 300 dpi, saved in tiff format, with no compression.

These images constitute the database used for the OCR pre-processing improvement and for the artifact recognition studies. Regarding the OCR pre-processing improvement study, although the main object of study are the set of images referred, in order to have a term of comparison, a survey of Finereader’s performance for modern documents was also realized.

The dataset used consists of four groups: group 1 contains 5 A4 pages of modern documents, with font of 10pt, containing about 3700 words and 18500 characters. Group 2 and 3 consist of 5 typed written text images from XIX and early XX century each containing 1440 words with 7000 characters, 900 words with 4300 characters, respectively. Group 4 consists of 122 ancient images with artifacts, also from XIX and early XX century. In the resize study, documents from groups 1 and 2 were used in the survey of Finereader’s performance for modern documents and ancient documents. Documents from group 2 represent what is considered in this work as well preserved documents, due to the definition of their foreground and their state of conservation. The thresholding study was performed using images from this group. Regarding documents from group 3, these are characterized for exhibiting signs of foreground degradation. For this reason, these images will the object of study for deblurring techniques.

To what concerns the artifact recognition study, 122 images of ancient documents, handwritten and typewritten were used. Fig. 1-1 illustrates examples of documents from group 1 and 2, while Fig 1-2 illustrates examples of documents from group 3 and 4.
Fig. 1-1 - a) example of image from group 1; b) example of image from group 2; c) zoomed view of image from group 1 example; d) zoomed view of image from group 2 example
Fig. 1-2 - a) example of image from group 3; b) example of image containing artifacts; c) zoomed view of image from group 3 example; d) zoomed view of image containing artifacts
1.1.1 Types of documents degradation

Degradation can be seen as “every sort of less-than-ideal properties of real document images, e.g. coarsening due to low digitizing resolution, ink/toner drop-outs and smears, thinning and thickening, geometric deformations, etc”[5]. Restoration, on the other hand, can be thought as a transformation process that gives the original aspect to images that show a certain state of degradation. Restoration is needed not only to improve the appearance of a document but also to improve the results of further segmentation and recognition operations. By clearing artifacts from the images there is less room, in the future, for misinterpretation. A document can appear degraded in multiple ways. The most common sources of image degradation are related with the acquisition process and with the storage environment. Degradation is unquestionably one of the main reasons for image processing to fail. Most degradation types in document images affect both physical and semantic understandability in the document analysis tasks, such as page segmentation, classification and optical character recognition. A typology for different types of degradation on old document images has been proposed by Drira[6]. This typology was preceded by a research of all degradations, which was done by consulting various images of degraded documents. The proposed classification was made according to the further treatment that will be applied in the context of virtual document image restoration. It is decomposed into three classes:

1) foreground degradation; 2) background degradation; and 3) global degradation.

These will be described next, along with examples of typical degradations that appear on ancient documents. All images are courtesy of the Tropical Research Institute (IICT), Portugal.

Foreground Degradation

Degradation on the foreground generally leads to broken or touching foreground objects, for instance, characters. Age effects can affect the ink components of a document. Many chemical effects can occur leading to ink disappearance and some gaps can even appear in the document image causing significant loss of data and therefore affecting the document’s content.

Background Degradation

The most common degradation is characterized by the presence of artifacts in the background of documents. It includes stamps, underlines, strokes of pen, annotations, marks resulting from the scanning process, blotches due to humidity. Although water blotches may appear, the most common artifacts are the ones intentionally introduced.
Global Degradation
This type of degradation affects documents in their entirety. It refers to a transformation that can be observed in a document as a whole, i.e., without affecting uniquely the foreground or the background. This transform can act either on the localization of the pixel (skew, degraded curve) or on its value (transformation of the color). Fig. 1-3 presents examples of degradation present in ancient documents.

Fig. 1-3 - a) foreground degradation – blurring; b) background degradation
d) artifact interference – stamp; e) artifact interference – underline; d) artifact interference – annotation
1.2 Objectives

The work in this thesis was carried out with documents provided by the Tropical Research Institute (IICT), which supplied the necessary resources and served as the final target for the application of the techniques here presented. The objectives within this thesis are to provide a contribution to the field of image restoration, in particular to increase of the OCR conditions in ancient documents, to the extraction of metadata regarding the presence of artifacts and to generate a digital version of documents free of artifacts.

To this purpose, the literature was reviewed with regard to the applicability of approaches to the context of ancient documents and methods are proposed and analysed in order to successfully solve the problem. These methods are those determined to perform the best under the conducted study and experimentation.

Characteristic of these objectives is naturally the notion of their feasibility. The intent is to provide a means for restoration that works most of the times, with the realistic consciousness that performing restoration that works every time, with every image and on any condition is far from possible.

1.3 Main contributions

These are, in summary, the main contributions of this thesis:

(1) Survey of problematic characters in OCR processes, for modern and typed documents;
(2) Analysis of the impact of restoration techniques applied to ancient documents on the OCR software, through measures of precision and accuracy of the recognized characters;
(3) Extraction of a set of descriptive metadata from content present in the document;
(4) Development of an artifact removal application

1.4 Organization

Presented in Chap. 2 is an overview of the main image processing techniques used in this dissertation. Chapter 3 discusses general methods for ancient documents restoration and optical character recognition of ancient typed documents, while Chap. 4 discusses methods for metadata extraction. Lastly, Chap. 5 concludes this dissertation, indicating its main contributions and future work directions.
2 Pre-processing Image Techniques

2.1 Resizing

Mathematically, image resize is the process of transforming a sampled image from one coordinate system to another. The two coordinate systems are related to each other by the mapping function of the spatial transformation. The inverse mapping function is applied to the output sampling grid, projecting it onto the input. The result is a resize grid, specifying the locations at which the input is to be resampled. The input image is sampled at these points and the values are assigned to their respective output pixels. The resize grid does not generally coincide with the input sampling grid, taken to be the integer lattice. This is due to the fact that the range of the continuous mapping function is the set of real numbers. The solution requires a match between the domain of the input and the range of the mapping function. This is achieved by converting the discrete image samples into a continuous surface, i.e. by image reconstruction. Once the input is reconstructed, it can be resampled at any position. Conceptually, image resize is comprised of two stages: image reconstruction followed by sampling. Although resize takes its name from the sampling stage, image reconstruction is the implicit component in this procedure, which is achieved through an interpolation procedure [1].

Thus, interpolation can understood as the process of determining the values of a function at positions lying between its samples. This process is performed by convolving the discrete input signal with a continuous interpolating function [2]. The classic solution consists of approximating the continuous function as a sum of interpolation functions centered at the nodes of a regular grid, multiplied by the image value at those.

Therefore, for equally spaced data, interpolation can be expressed as:

\[
f(x) = \sum_{k=0}^{K-1} c_k h(x - x_k) \tag{2.1}
\]

where \( h \) is the interpolation kernel and \( c_k \) are the data samples themselves, applied to \( K \) data samples.

The computation of one interpolated point is realized by centring at \( x \) the interpolating function, so the value of that point is equal to the sum of the values of the discrete input scaled by the corresponding values of the interpolation kernel. The numerical accuracy and computational cost of interpolation algorithms are directly tied to the interpolation kernel. Regarding two-dimensional interpolation, the most common approach used in image processing is to decompose the problem into a sequence of several one-dimensional interpolation tasks. The key idea is to perform linear interpolation first in one
direction, and then again in the other direction. The process of interpolating in two dimensions using a sequence of one-dimensional linear interpolations is called *bilinear* interpolation. Similarly, *bicubic* interpolation is a two-dimensional interpolation performed using a sequence of one-dimensional *cubic* interpolations [2].

**Interpolation kernel**

The interpolation kernel applied during the resize process was the *cubic* kernel. This is considered to be the interpolation kernel that offers the best relation precision/computational effort and is the most common kernel used in image processing [3]. Its graphical representation can be seen in Fig. 2-1.

The *cubic* kernel is defined by:

\[
h_c(x) = \begin{cases} 
1.5|x|^3 - 2.5|x|^2 + 1, & |x| \leq 1 \\
-0.5|x|^3 + 2.5|x|^2 - 4|x| + 2, & 1 < |x| \leq 2 \\
0, & \text{otherwise}
\end{cases}
\]  

Fig. 2-1 - Cubic kernel. From Digital Image Processing using Matlab (p. 300), by Gonzalez, 2009, Gatesmark Publishing

It is important to note that in the general case, when using cubic interpolation, it can give rise to values outside the range of the input data. Consequently, when using this method in image processing it is necessary to properly clip or rescale the results into the appropriate range for display [1].

**2.2 Thesholding**

A binary image is a digital image that has only two possible values for each pixel. Typically the two colors used for a binary image are black and white, though any two colors can be used. The value corresponding to the black is 0, and for the white is 1. In order to binarize an image, assuming that it is in the RGB color space, it is necessary to convert it first into greyscale. Only after this operation, the image is able to be subjected to thresholding. Having the image in a greyscale allows the construction of a graphic with the distribution of the tonality. This type of graphs is called histograms. The horizontal axis of the graph represents the tonal variations, whereas the vertical represents the number of pixels in that particular tone. The left side of the horizontal axis represents the black and dark areas, the middle represents medium grey and the right hand side represents light and pure white areas. This representation allows an understanding of how dark, or light, an image actually is. Considering an intensity histogram corresponding to an image, \( f(x, y) \), composed of light objects on a dark background, in such a way that object and background pixels have intensity levels, these can be grouped into two dominant modes. It is possible to extract the objects from the background, selecting
a threshold $T$ that separates these modes. This threshold determines which pixels get the value 0 and which get the value 1.

The binary image $g(x,y)$ is defined as:

$$g(x,y) = \begin{cases} 1, & f(x,y) > T \\ 0, & f(x,y) \leq T \end{cases}$$

(2.3)

when $T$ is a constant applicable over an entire image, the preceding equation is referred to as *global* thresholding. When the value of $T$ changes over an image, it is used the term *adaptive thresholding* [2].

### 2.2.1 Methods

**Niblack**

Niblack’s thresholding method[10] performs adaptive thresholding and was selected because it is frequently cited and has been thoroughly reviewed with other types of documents. Comparisons have revealed that it outperforms other thresholding algorithms under different application domains[11]. The main idea of Niblack’s method is to build a threshold surface, based on the local mean, $m$, and standard deviation, $s$, of grey values computed over a small neighbourhood around each pixel in the form of:

$$t(x,y) = m(x,y) + K \times s(x,y)$$

(2.4)

Where $K$ determines the weight of the standard deviation in the calculation of the threshold.

**Sauvola**

Sauvola’s thresholding method[12] performs adaptive thresholding and was chosen because it is a modification of Niblack’s method aimed at dealing better with the cases in which the background contains light texture, big intensity variations and uneven illumination. These properties are characteristic of images of ancient documents. Sauvola’s method improve Niblack’s method by imposing a hypothesis on the grey values of text and non-text pixels - text pixels have near 0 grey values and non-text pixels have near 255 grey values (considering 8 bites) - and compute the local threshold value as:

$$t(x,y) = m(x,y) \times \left[ 1 + K \left( \frac{s(x,y)}{R} - 1 \right) \right]$$

(2.5)

Where $R$ is the dynamic range of standard deviation.

Sauvola’s method can outperform Niblack’s for well-scanned document images, but it faces difficulties
when dealing with images that do not agree with the hypothesis on which it relies, e.g., documents images captured at insufficient illumination, especially when the grey values of text and non-text pixels are close to each other. In Savoula’s method, as in Niblack’s method, the value of standard deviation is the responsible factor for the change in threshold’s values. Considering the greatest contrast possible, \( s \to 128 = R \) (considering 8 bites), thus:

\[
t(x, y) = m(x, y) \times \left[ 1 + K \left( \frac{R}{R} - 1 \right) \right] = m(x, y)
\]

which is the same result obtained in Niblack’s method, for the same situation. The difference between these methods becomes clearer when dealing with a low contrast. If the contrast tends to zero, i.e., if \( s \to 0 \), thus:

\[
t(x, y) = m(x, y) \times \left[ 1 + K \left( \frac{0}{R} - 1 \right) \right] = m(x, y)[1 - K]
\]

**Otsu**

Otsu’s method[13] was included in this study for completeness, to perform global thresholding. It has been thoroughly evaluated [14,15] and used before. This method selects a threshold that maximizes between-class variance after creating the histogram of an intensity image. This threshold is then applied to all image pixels. It has the advantage of not requiring the input of parameters, for assuming that histograms are bimodal and illumination is uniform. This method selects a threshold that maximizes between-class variance after creating the histogram of the intensity image. The aim is to find the threshold value where the sum of foreground and background spreads is at its minimum.

**Fuzzy C-Means Clustering**

The Fuzzy C-Means method[16,17] is well known and revealed to present good results in a previous work by Pinto et al[18] Fuzzy C-Means is used to partition an image into N clusters, where each pixel has a membership grade to each cluster. Groups, or clusters, are created based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Objects within a cluster should have high similarity between each other, but should also be very dissimilar to objects in other clusters. Because clustering is performed in the color space, cluster centers can be interpreted as colors that describe the clusters themselves. Consequently, cluster centers are used to determine which cluster relates to the darkest color. All pixels that share a membership grade to that cluster that is greater than an \( M \) value are selected as valid foreground pixels. In this case, the grey-scale values are clustered into two fuzzy classes corresponding to background and foreground pixels. All pixels that share a membership grade that is greater than an \( M \) value are selected as valid foreground pixels[19].
Canny Edge Detection with Niblack Post-processing

This method is a composition of algorithms, similar in part to what was used by Tan et al[20], with some variations. As outlined in Fig. 2.2, a Canny edge detector[21] is used initially, followed by foreground recovery and an adaptive thresholding algorithm, in this case Niblack’s.

![Canny+thresholding method scheme](image)

Canny edge detection is used to detect the edges within the image based on the observation that the edges of valid foreground are usually sharper than those of the background interference. This method finds edges by searching for the maximum of the gradient in an image, being the gradient calculated through the derivative of the Gaussian filter. It uses two different thresholds with the purpose of distinguishing the weak edges from the strong edges. The weak edges will be included afterwards in case of being linked to strong edges [2].

Following the detection of the edges, recovery is needed in order to restore the original foreground. The dilation is used with a structuring element that corresponds to a square of size $W \times W$, with the purpose of filling the regions around the edges.

The dilation operator retains not only valid parts of the foreground but also background and degradation surrounding the edges. Therefore, adaptive thresholding is used to remove the remaining interference. Niblack’s method was once again selected for this task, due to the reasons mentioned while describing this adaptive thresholding method.

Canny Edge Detection with Sauvola Post-processing

This method is similar to the previous one, with the difference that Sauvola method is used in the postprocessing phase, instead of Niblack.
2.3 Deblurring

Images are produced to record or display useful information. However, due to imperfections in the imaging and capturing process, the recorded image, invariably, represents a degraded version of the original scene[22]. The use of deblurring techniques aims to remove or minimize known degradations in an image, using a priori knowledge of the degradation phenomenon.

Recovering an unblurred image from a single, motion-blurred photograph has long been a fundamental research problem in digital imaging. If one assumes that the blur kernel or Point Spread Function (PSF) is shift-invariant, the problem reduces to that of image deconvolution. Thus, image deblurring can be categorized into two types: blind deconvolution and non-blind deconvolution.

In non-blind deconvolution, the motion blur PSF is assumed to be known or computed elsewhere. The only task remaining is to estimate the unblurred latent image. Traditional methods such as Weiner[23] and Richardson-Lucy deconvolution[24] were proposed decades ago, but are still widely used in many image restoration tasks nowadays because they are simple and efficient.

**Deblurring Model**

A blurred or degraded image, in the spatial domain, can be approximately described by the following equation:

\[
g(x, y) = h(x, y) * f(x, y) + \eta(x, y)
\]

where \(g(x, y)\) is the blurred image, \(h(x, y)\) the PSF, \(f(x, y)\) the original image and \(\eta(x, y)\) the noise affecting the system.

![Deblurring Model](image-url)

Fig. 2.3 - Deblurring Model. From Digital Image Processing using Matlab (p. 300), by Gonzalez, 2009, Gatesmark Publishing
**Point Spread Function – PSF**

The distortion operator, $h(x, y)$, known by PSF models the blurring of the image, and is considered to be linear and shift-invariant.

By linearity:

$$H[af_1(x, y) + bf_2(x, y)] = aH[f_1(x, y)] + bH[f_2(x, y)]$$  \[ (2.9) \]

and by shift-Invariance:

$$H[f(x - \alpha, y - \beta)] = g(x - \alpha, y - \beta)$$  \[ (2.10) \]

The PSF can be seen as the irradiance distribution that results from a single point source. Although the source may be a point, as a consequence of the diffraction of light and the presence of defects in the image device, the image of a point source occupies an area of finite dimensions, rather than a point. In this sense, the image of a complex object can be seen as a convolution of the true object and the PSF. In a systems perspective, the PSF represents the impulse response of a focused optical system, i.e., the response of an imaging system to a point object [2].

**Frequency domain**

Convolution in the spatial domain and multiplication in the frequency domain constitute a Fourier transform pair, which allows writing the preceding model in an equivalent frequency domain representation:

$$G(u, v) = H(u, v) \times F(u, v) + N(u, v)$$  \[ (2.11) \]

The terms present in the equation are the Fourier transforms of the corresponding terms in the spatial domain. The degradation function $H(u, v)$ is also known as the optical transfer function (OTF) [4].
2.3.1 Methods

Direct Inverse Filtering

The simplest approach in what concerns image restoration is to ignore the noise affecting the image system and inverse the blurring process. Dividing both sides of equation 2.11 by $H(u, v)$, it is possible to produce an estimate of the form:

$$\hat{F}(u, v) = \frac{G(u, v)}{H(u, v)} = Y(u, v) \times G(u, v)$$  \hspace{1cm} (2.12)

Where

$$Y(u, v) = \frac{1}{H(u, v)}$$  \hspace{1cm} (2.13)

is the frequency-domain filter. The advantage of this filter is that it requires only the blur PSF as a priori knowledge. Thus, the inverse filter produces perfect reconstruction in the absence of noise. However, noise is always present, corrupting any recorded image. It is introduced by the medium through which the image is created (random absorption or scatter effects), by the recording medium (sensor noise), by measurement errors due to the limited accuracy of the recording system, and by quantization of the data for digital storage. Thus, taking the noise into account, dividing again both sides of equation 2.11 by $H(u, v)$ we will obtain an estimate of the form:

$$\hat{F}(u, v) = \frac{G(u, v)}{H(u, v)} = F(u, v) + \frac{N(u, v)}{H(u, v)}$$  \hspace{1cm} (2.14)

From this equation, it is wanted to have the second term as small as possible, so the estimate can better approaches the Fourier transform of the true image.

Nevertheless, the inverse filter may not exist due to the presence of zeros in $H(u, v)$ at the selected frequencies $(u, v)$. Even if the blurring function spectral representation $H(u, v)$ does not actually go to zero but becomes small, the second term in equation 2.14, known as the inverse filtered noise, will become very large, leading inverse filtered images to be dominated by amplified noise.
The typical approach when attempting inverse filtering is to form the ratio:

\[ F(u, v) = \frac{G(u, v)}{H(u, v)} \]  \hspace{1cm} (2.15)

and then limit the frequency range for obtaining the inverse, to frequencies "near" the origin. The idea is that zeros in \( H(u, v) \) are less likely to occur near the origin because the magnitude of the transform typically is at its highest values in that region [4].

**Wiener deconvolution**

A Wiener filter seeks an estimate \( \hat{f} \) that minimizes the statistical error function between the ideal and the restored image:

\[ e^2 = E \left\{ (f - \hat{f})^2 \right\} \]  \hspace{1cm} (2.16)

where \( e \) is the error, \( E \) the is expectancy operator, \( f \) the original image and \( \hat{f} \) the reconstructed image. In essence, the Wiener filter works by accepting low-noise frequency components and rejecting high-noise frequency components, that is, it amplifies the amplitude of the spectral components whose signal energy dominates the noise energy, shrinking the spectral components in which the noise starts to dominate the signal.

The solution to eq. 2.16 in the frequency domain is:

\[ \hat{F}(u, v) = \left[ \frac{1}{H(u, v)} \times \frac{|H(u, v)|^2}{|H(u, v)|^2 + S_n(u, v)/S_f(u, v)} \right] G(u, v) \]  \hspace{1cm} (2.17)

where the term \( S_n(u, v) \) stands for the power spectrum of the noise, and where \( S_f(u, v) \) stands for the power spectrum of the undegraded image. The ratio \( S_n(u, v)/S_f(u, v) \) is called the noise-to-signal power ratio. If the noise power spectrum is zero for all relevant values of \( u \) and \( v \), this ratio becomes zero and the Wiener filter reduces to the inverse filter [2].

**Constrained Least Squares Regularized Filter**

The constrained least squares (CLS) regularized filter approaches the process of obtaining a meaningful restoration by biasing the solution toward the minimizer of some specified constraint function. In this sense, it is desired to find the minimum of a criterion function, such as the second derivative of an image (Laplacian):

\[ C = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} [V^2 f(x, y)]^2 \]  \hspace{1cm} (2.18)
Subject to the constraint:

\[ \| g - H \hat{f} \|^2 = \| \eta \|^2 \quad (2.19) \]

The frequency domain solution to this optimization problem is given by the expression:

\[ P(u, v) = \left( \frac{H^*(u, v)}{|H(u, v)|^2 + \gamma |P(u, v)|^2} \right) G(u, v) \quad (2.20) \]

where \( H^*(u, v) \) is the complex conjugate of the Fourier transform of the PSF, \( P(u, v) \) is the Fourier transform of the Laplacian filter and \( \gamma \) is a constant, which is responsible to control the balance between noise and observed data. The filter \( P(u, v) \) has a large amplitude at high frequencies, where the noise tends to be dominant, reducing the noise effects at high frequencies. Choosing a proper \( P(u, v) \) and \( \gamma \) can, therefore, minimize higher-order derivatives. The only unknowns in the preceding formulation are \( \gamma \) and \( \| \eta \|^2 \). Nevertheless, it can be shown that \( \gamma \) can be found iteratively if \( \| \eta \|^2 \) (which is proportional to the noise power) is known. Likewise the Wiener filter, the CLS filter also has a term in its denominator adjudicated to noise. However, this term operates differently in the CLS method than in the Wiener filter. The main difference between these two methods relies on the determination of the noise term. The Wiener filter has the Noise to Signal Ratio (NSR) term, which can be determined deterministically if the power spectrum of the noise and the undegraded image are known, while the CLS method has the Noise Power (NP), which is a scalar and can be only determined after an iterative process.

**Lucy-Richardson algorithm**

The Lucy-Richardson (LR) method is an iterative algorithm that attempts to find the maximum-likelihood solution given knowledge of the PSF and the assumption of Poisson noise distribution. The linear imaging equation states that:

\[ g(x, y) = h(x, y) * f(x, y) + \eta(x, y) \quad (2.21) \]

so that the noise \( \eta(x, y) \) is the difference between the output distribution \( g(x, y) \) - the image actually measured - and the unknown input distribution \( f(x, y) \) convolved with the PSF - \( h(x, y) \):

\[ \eta(x, y) = g(x, y) - h(x, y) * f(x, y) \quad (2.22) \]

Substitution of an appropriate estimate of the input distribution \( f(x, y) \), in the limit of negligible noise, would satisfy the following equation:

\[ g(x, y) - f(x, y) * h(x, y) = 0 \quad (2.23) \]
Adding the input distribution \( f(x, y) \) to both sides of Equation 2.23, leads to:

\[
f(x, y) = f(x, y) + [g(x, y) - f(x, y) \ast h(x, y)]
\]  
(2.24)

This equation can be seen as an iterative procedure in which a new estimate of the input is given as the sum of the previous estimate with a correction term. The correction term is the difference between the measured image and the prediction of it using the current estimate of the input. Writing this as an iterative procedure, the \((k + 1)th\) estimate of the input is thus given by:

\[
f_{k+1}(x, y) = f_k(x, y) + [g(x, y) - f_k(x, y) \ast h(x, y)]
\]  
(2.25)

The procedure described by equation 2.25 is started by setting \( f_0(x, y) = g(x, y) \)

The basic assumptions of the Lucy-Richardson method are twofold:

- PSF is known
- Noise in the output is governed by the Poisson density function

Considering a discrete form of Equation 2.25, pixels in the observed image can be represented in terms of the point spread function and the original image as:

\[
g_i = \sum_j h_{ij} f_j
\]  
(2.26)

where the summation over index \( j \) provides the contribution of each input pixel, affected by the PSF, \( h(x, y) \), to the observed output pixel.

It is customary to normalize the discrete PSF so that:

\[
\sum_i \sum_j h_{ij} = 1
\]  
(2.27)

The iterative Lucy-Richardson formula is given by:

\[
f_j^{(k+1)} = f_j^{(k)} \times \frac{g_i h_{ij}}{c_i},
\]  
(2.28)

where \( c_i = \sum_j h_{ij} f_j^{(k)} \)

If this iteration converges, it converges to the maximum likelihood solution for \( f_j \) [2].
Blind Deconvolution

Blind deconvolution problems appear in image analysis when both the blur and the true image are unknown. This method can be performed iteratively, having each iteration improving the estimation of the PSF and the scene, or non-iteratively, where one application of the algorithm, based on exterior information, extracts the PSF.

A fundamental approach to blind deconvolution is based on Maximum-Likelihood Estimation (MLE), an optimization strategy used for obtaining estimates of quantities corrupted by noise [2]. The optimization problem is solved iteratively with specified constraints and, assuming convergence, the specific $f(x, y)$ and $h(x, y)$ that result in a maximum are the restored image and the PSF. In this process, the image is assumed to be positive, and comprise of an object with known finite support against a uniformly black, grey or white background. The support refers to the smallest rectangle within which the object is completely encompassed. If the background is black, which corresponds to a pixel value of zero, the support is the smallest rectangle within the true image pixels are non-zero.

After a random initial guess for the image, the algorithm alternates between the image and Fourier domains, enforcing known constraints in each domain. The constraints are based on information available about the image and PSF. The undegraded image, $f(x, y)$, and the PSF, $h(x, y)$, are assumed to be positive with finite known support. The image-domain constraints are imposed by replacing negative-valued pixels within the region of support and non-zero pixels outside the region of support with zero-valued pixels. The Fourier domain constraint involves estimating the PSF using the Fast Fourier Transform (FFT) of the degraded image and image estimate. At the $n^{th}$ iteration:

$$
\hat{H}_n(u, v) = \frac{G(u, v) \hat{F}_{n-1}^*(u, v)}{|\hat{F}_{n-1}(u, v)|^2 + \frac{\alpha}{|\hat{H}_{n-1}(u, v)|^2}}
$$

where the subscript $n$ represents the iteration of the algorithm, and where $\hat{F}(u, v)$ and $\hat{H}(u, v)$ represent the Fourier transform of the image estimation and PSF estimation, respectively. The constant $\alpha$ represents the energy of the additive noise, which is the responsible for a reliable restoration [4].
3 Improving Optical Character Recognition

The OCR software used to serve as reference and to measure the improvements made during the restoration process is the ABBYY Finereader 11 Professional. ABBYY FineReader is a leading commercial OCR software package which allows highly successful recognition of most modern documents[1], considered as one of the most advanced commercial packages for OCR [5]. It performs extensive layout retention, which, although it is not of direct concern of this project, implies detection and optional extraction of pictures, graphs and other non-text content. The selection of FineReader among other OCR software was based on previous work for the Biblioteca Nacional de Lisboa where several options were considered[27]. This investigation analysed a number of professional reviews and conducted realistic tests with the trial versions of the various applications when available. It was concluded that FineReader performed best with ancient material and that its corresponding development kit should be used.

3.1 Evaluation Methodology

In order to evaluate the outcome of a technique, the image that results from a studied algorithm is saved in the tiff format, with no compression, and inserted into Finereader. After, the OCR result is compared with the original text (ground truth). The OCR process starts segmenting a given image in text, background and pictures. Subsequently, the process turns to the recognition of characters in the detected text components of the document image. The result is the set of characters recognized, in which the potentially incorrect characters are highlighted in blue. The red underlines indicate the words not recognized by Finereader’s dictionary. An example of these results is presented in Fig. 3-1.

The result is saved in a txt file, and equally in a docx, in order to keep the information about the underlined characters. This docx file is then saved in a html file. Through this conversion it is possible to gain access to the characters underlined. In html code, the underlining is materialized by a code that precedes the underlined characters. In this sense, every time this code is detected, the application will consider that the following characters are underlined. The underlining stops when a code which gives the indication of the ending of the underlining is detected. It is critical that the number of words between the OCR result file and the original text file is the same to ensure that the same words are being compared. Whenever a recognized word present a different number of letters than the corresponding original word, the counting application skipped this word and moved to the next one, ignoring any of the letters present in that word.

Summarizing, in order to perform statistical measures of the results, three files are required: the original text, the ocr result and the html version of this.
3.2 Statistical Measures

Taking into account that Finereader gives information about potentially inaccurate recognized characters, by underlying these, it is possible to categorize the recognition of a character in four ways:

- **True Negative (TN):** underlined characters, that are incorrectly recognized (appropriately underlined)
- **False Negatives (FN):** underlined characters, correctly recognized (underlined unnecessarily)
- **False Positives (FP):** characters not underlined, that were incorrectly recognized (lack of underlining)
- **True Positives (TP):** characters not underlined, that were correctly recognized

According with this classification, the information obtained was organized in a confusion matrix. The statistical measures of performance precision, recall and specificity will be described next:

**Precision**

It is the most important measure of performance, for providing the ratio between the characters not underlined, that are actually correctly recognized (TP) and all the characters that are not underlined (TP+FP). In other words, it is an indicator of reliability, in the sense that it indicates the trust that can be put into a character not underlined as a correctly recognized character. Precision is defined by:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.1)$$
Recall
Due to the uncertainty degree associated in the classification process, Finereader may underline characters that are correctly recognized. Thus, recall intents to measure the ratio between the characters correctly recognized not underlined and all the characters correctly recognized (TP+FN) and can expressed by:

\[
Recall = \frac{TP}{TP + FN}
\]  (3.2)

Specificity
It indicates the odds of a character being actually inaccurate. This measure gains importance in case of high precision, for indicating to a potential corrector all the characters needing correction, allowing saving time in the verification of its classification. It can be seen as a measure of efficiency of inaccurate characters classification. Specificity is defined by:

\[
Specificity = \frac{TN}{TN + FP}
\]  (3.3)

Besides these indicators, it is also considered in the analysis the number of correctly recognized words - words with the right numbers of characters, and the number of correctly recognized characters. These two parameters are presented in the form of percentage:

**Percentage of correctly recognized words**

\[
\% \text{ c.r.}^1 \text{ words} = \frac{\text{Total number of words in the ocr result file}}{\text{Total number of words in the original file}}
\]  (3.4)

**Percentage of correctly recognized characters**

\[
\% \text{ c.r.}^1 \text{ ecognized characters} = \frac{TP}{TN + FN + FP + TP}
\]  (3.5)

---

1 c.r. – correctly recognized
3.3 Results

3.3.1 Resizing

A factor responsible for Finereader performance losses is the presence of noise, acquired during the scanning process[28]. Thus, due to the loss of information that occurs in the resizing process, this technique arises as a possible solution to minimize the noise effect, increasing Finereader’s hit rate.

Resizing is used to create a new version of the image with a different width and/or height in pixels, changing the image file size as well as the image resolution. When the number of pixels is increased, the process takes the designation of upsampling, and for the opposite, the process takes the designation of downsampling. When images are upsampled, the number of pixels increases, but, with reference to the original subject, new image detail cannot be created that was not already present in the original image. When images are downsampled, information in the original image has to be discarded to make the image smaller[29].

Resize will be applied to typed modern and ancient documents. Every image will be reduced using three reduction factors (RF), which are 0,75, 0,50 and 0,25. The study stops in the 0,25 RF for being noticed that, from this RF on, OCR operations are fruitless.

The Matlab function imresize is the function used to realize the resize operation. This function takes as inputs the image to be reduced, the RF and the interpolation method.

<table>
<thead>
<tr>
<th>Reduction Factors</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,00</td>
<td>100,00</td>
<td>96,04</td>
<td>100,00</td>
<td>95,79</td>
<td>99,67</td>
</tr>
<tr>
<td>0,75</td>
<td>99,97</td>
<td>96,43</td>
<td>89,58</td>
<td>96,18</td>
<td>99,78</td>
</tr>
<tr>
<td>0,50</td>
<td>99,99</td>
<td>95,83</td>
<td>95,65</td>
<td>95,59</td>
<td>99,59</td>
</tr>
<tr>
<td>0,25</td>
<td>98,24</td>
<td>91,93</td>
<td>50,51</td>
<td>88,96</td>
<td>99,02</td>
</tr>
</tbody>
</table>
Table 3.2 - Performance differences between the RF analysed and the 1,00 factor for modern documents

<table>
<thead>
<tr>
<th>Reduction Factors</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0,75</td>
<td>-0,03</td>
<td>+0,40</td>
<td>-10,42</td>
<td>+0,39</td>
<td>+0,11</td>
</tr>
<tr>
<td>0,50</td>
<td>-0,01</td>
<td>-0,21</td>
<td>-4,35</td>
<td>-0,21</td>
<td>-0,08</td>
</tr>
<tr>
<td>0,25</td>
<td>-1,76</td>
<td>-4,11</td>
<td>-49,49</td>
<td>-6,83</td>
<td>-0,65</td>
</tr>
</tbody>
</table>

Table 3.3 - Finereader’s performance evaluation – ancient typed documents (group 2)

<table>
<thead>
<tr>
<th>Reduction Factors</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,00</td>
<td>99,15</td>
<td>95,48</td>
<td>52,94</td>
<td>93,85</td>
<td>99,44</td>
</tr>
<tr>
<td>0,75</td>
<td>99,01</td>
<td>95,21</td>
<td>66,10</td>
<td>92,62</td>
<td>93,82</td>
</tr>
<tr>
<td>0,50</td>
<td>99,08</td>
<td>94,12</td>
<td>54,62</td>
<td>92,35</td>
<td>98,96</td>
</tr>
<tr>
<td>0,25</td>
<td>98,16</td>
<td>90,02</td>
<td>66,96</td>
<td>85,65</td>
<td>98,61</td>
</tr>
</tbody>
</table>

Table 3.4 - Performance differences between the RF analysed and the 1,00 factor for ancient typed documents (group 2)

<table>
<thead>
<tr>
<th>Reduction Factors</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,00</td>
<td>-0,14</td>
<td>-0,27</td>
<td>+13,16</td>
<td>-1,23</td>
<td>-5,62</td>
</tr>
<tr>
<td>0,75</td>
<td>-0,07</td>
<td>-1,36</td>
<td>+1,68</td>
<td>-1,5</td>
<td>-0,48</td>
</tr>
<tr>
<td>0,50</td>
<td>-0,99</td>
<td>-5,46</td>
<td>+14,02</td>
<td>-8,2</td>
<td>-0,83</td>
</tr>
<tr>
<td>0,25</td>
<td>-0,14</td>
<td>-0,27</td>
<td>+13,16</td>
<td>-1,23</td>
<td>-5,62</td>
</tr>
</tbody>
</table>

Problematic characters survey

2 The positive values represent the improvements achieved by the method, while the negative represent performance losses
Observing the OCR results for modern and typed documents it is possible to notice that some characters present a pattern in their classification. This group gives focus to the classification expected for each character. This survey considers a character to be common in FP and TN classifications when classified with these classifications in more than 10% of the times they appear. In FN cases, given their abundance in the 0,25 RF for both types of documents, characters are considered to be common in this classification when these are classified as FN in more than 25% of the cases. Finally, in order to summarize the problematic characters survey, it is considered as a character of problematic recognition, any character whose classifications as TP is lower than 10%. The number of occurrences of some characters, between RF, may vary due to the difference of the percentage of words correctly recognized in each RF. The characters distribution through the four classifications can be consulted in the appendix.

Summary
Analysing the results of ancient typed documents, it can be seen that some characters are not represented enough to generate a reliable distribution about their type of classification. Therefore, a higher number of pages should be considered in the study in order to include a significant amount of characters for every character. However, some conclusions can be drawn. Finereader, as expected, revealed a lower performance for ancient documents than for modern. Still, the specificity, for the 0,25 RF, is higher in ancient documents than for modern documents. When comparing the SM of the ancient documents with the modern documents between RF, it was noticed that, apart the specificity, none of the statistical measures differences between this two types of images is higher than 5%, regardless the RF considered. Therefore, the specificity is the only statistical measure that is considerably affected by the type of image used. Concerning the number of words analysed, although variable, it is superior to 99,5% for modern documents and superior to 98,5% for typed documents, for every RF. From the results obtained for the two types of documents and for the RF analysed, the 0,50 RF presented the best relation size reduction/SM losses. For this reason, the following studies were realized using images reduced by this factor. Looking at the tables for both types of documents, it is possible to notice that, for every classification, the RF affects a different group of characters. Yet, part of these characters is transversal to all of the RF. Difficulties were noticed in the recognition of punctual signs, characters with accents, particularly in the vowels “a” and “o”, and in characters in its majuscule form, particularly “K” “Q” “W” and “Â”. The characters presented in the table showing the characters whose classification as TP is less than 10% can be considered as characters of difficult recognition, requiring for this reason special attention during the correction process. To guarantee the quality of the correction of the OCR result, each character must be revised. Nevertheless, from the analysis made to the distribution of the characters by the four classifications it is possible to predict each character classification, using this to find more easily the incorrectly recognized unmarked characters.

3.3.2 Thresholding
Many segmentation approaches that aim to extract clear text from either noisy or textured backgrounds have been reported and compared in the literature [27-35].
Negishi et al. [30] described an automatic thresholding algorithm to extract character bodies from noisy backgrounds. Their algorithm removes black frames that surround page areas and equalizes the background to eliminate the overlapping between the grey levels of character parts and those of the background. An automatic thresholding method is then applied to convert the original image into a reasonably noise-free binary image. Liu and Srihari [31] used a thresholding algorithm based on texture features to extract characters from run-length featured texture backgrounds. Their algorithm utilizes two fundamental attributes of document images: 1) the characters normally occupy a separable grey-level range in the greyscale histogram; and 2) the text images contain highly structured stroke units.

Liang et al. [32] presented a morphological approach to extract text strings from regular periodic overlapping text-background images, useful for removing backgrounds with repetitive patterns. Don [33] used noise attribute features based on a simple noise model to overcome the difficulty that some objects do not form prominent peaks in the histogram encountered by many conventional global thresholding methods. Leedham et al. [34] compared several thresholding techniques for separating text and background in degraded historical documents. Tan et al. [35] consider the extraction of a clean foreground from historical handwritten documents. With the observation that edges of the valid foreground are sharper than those of the interference, edge detection algorithms can be used to detect the foreground edges and discard the degradation. Castro et al. [5] discuss general approaches to the restoration of ancient music documents, giving focus to the problem of background homogenization and bleed-through removal.

Image thresholding is a common first step to document image analysis, usually performed in the pre-processing stage of image processing applications such as optical character recognition (OCR). The pre-processing work, that offers the basic binary image of the text, start of the whole process, has the responsibility to present clear and homogeneous results, to allow an exploration of this “virtual” text, free from the defects.

There are two major classes of thresholding algorithms: global, which decides on one partition for the entire image based the properties of the entire image, and local (or adaptive), which considers the properties within smaller regions to produce possibly differing partitions for each region. Otsu’s method for greyscale images [6], a global thresholding technique prevalent in document recognition literature, uses a histogram to search for the partition that minimizes the variance in grey levels within each set in the partition. Where local thresholding is more appropriate, it is common to use a technique similar to Niblack’s method for greyscale images [7], which sets each region’s threshold to be a fixed fraction of one standard deviation above or below the average grey level within the region.

Segmenting text and background is a difficult classification problem, and the accuracy of this segmentation is of utmost importance to the output of an OCR system. Although document image thresholding has been studied for many years, the thresholding of degraded document images is still an unsolved problem. This can be explained by the fact that the modeling of the document foreground/background is very difficult due to various types of document degradation such as uneven
illumination, image contrast variation, bleeding-through, and smear.

**Algorithms parameters settings**

**Niblack / Sauvola**

Both methods have one parameter that determines the size of a square mask, $N$, and another that determines the value of the coefficient that affects the standard deviation, $K$.

In both methods an increase in the size of the mask leads to an increase of the thickness of the foreground, this is, of the characters. Regarding the remaining parameter, although affecting the threshold calculation differently in each method, it works upon the standard deviation in both cases. As a result, when increasing the value of $K$, the threshold lowers, providing a “lighter” image.

Regarding the Niblack method, the major drawback noticed was the noise present in the end result. This method tends to produce a large amount of noises in non-text image regions [12], which makes the analysis realized by Finereader, apart from time consuming also inefficient. The best parameters were found at $N=51$, $K=0.8$. An example of the information segmentation and the OCR result obtained, using Niblack’s method, is presented in Fig. 3.2:

![Fig. 3.2 – a) Niblack – Identification of information b) Niblack - OCR result](image)

Concerning the Sauvola method, this method showed a better performance compared with the performance of the preceding method, dealing better with the noise and presenting a good preservation of the foreground. The best parameters for Sauvola method were found at $N=5$, $W=0.20$.
**Fuzzy C-Means Clustering**

The parameters for this method are the number of clusters in which the pixels can be grouped, \( N \), and the membership grade that determines which pixels are selected as valid foreground pixels, \( M \).

When the numbers of clusters increase, a reduction in the thickness of the foreground occurs. The reduction of the foreground also occurs when the minimum membership grade for a pixel to belong to a cluster, decreases. Although the reduction of thickness may retrieve some characters, other characters that were already recognizable may lose this characteristic. Hence, taking into consideration that it is not possible to recover at the same time all characters, the goal is to find a set of parameters that returns the higher number of characters recognized. Nevertheless, this method has the drawback of requiring considerable computational effort. The best parameters were found at \( N=3, M=0.75 \).

**Otsu**

This method, for having as a single parameter the histogram of the image, does not allow any adaptation.
Canny Edge Detection with Niblack/Sauvola method

This composed method requires the same parameters referred for the thresholding methods and an additional parameter, which is responsible for the reconstruction of the foreground. The reconstruction is made through a morphological operation – dilation. Therefore, the parameter to define is the size of the mask to be used in this operation. Taking into consideration that this method consists in an adaptation of the thresholding methods used before, it would be expected that this adaptation would conduct to an improvement of the results. However, dilation has revealed itself insufficient to increase significantly the character definition, although some improvements were achieved. In this sense, this method consists, essentially, in an alternative to the original threshold methods. The best parameters were found at $D=9$, $N=51$, $K=-0.80$. The II and the OCR result for the typed image 1, using Canny+Niblack’s method, are presented in Fig. 3-3.

Although Canny+Niblack method showed some improvements in the II and OCR, this method still reveals a week performance, when compared with the previous method. On the other hand, the Canny+Sauvola method has shown a considerable potential regarding the increase of the hit rate. The best parameters were found at $D=9$, $N=25$, $K=0.18$. 

Table 3.5 - Finereader’s performance evaluation – modern documents (group 2) reduced by 0,50

<table>
<thead>
<tr>
<th>Reduction Factors</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,50</td>
<td>99,08</td>
<td>94,12</td>
<td>54,62</td>
<td>92,35</td>
<td>98,96</td>
</tr>
</tbody>
</table>

Table 3.6 - Finereader’s performance evaluation – thresholding methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sauvola</td>
<td>98,60</td>
<td>93,67</td>
<td>54,97</td>
<td>90,99</td>
<td>95,21</td>
</tr>
<tr>
<td>Canny+Sauvola</td>
<td>98,85</td>
<td>93,37</td>
<td>60,99</td>
<td>90,83</td>
<td>95,70</td>
</tr>
<tr>
<td>Fuzzy</td>
<td>98,42</td>
<td>94,27</td>
<td>49,47</td>
<td>91,52</td>
<td>93,20</td>
</tr>
<tr>
<td>Otsu</td>
<td>97,38</td>
<td>88,50</td>
<td>60,21</td>
<td>83,56</td>
<td>93,48</td>
</tr>
</tbody>
</table>

Table 3.7 - Performance differences between the method analysed and ancient typed documents (group 2)³

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sauvola</td>
<td>-0,48</td>
<td>-0,45</td>
<td>+0,35</td>
<td>-1,36</td>
<td>-3,75</td>
</tr>
<tr>
<td>Canny+Sauvola</td>
<td>-0,23</td>
<td>-0,75</td>
<td>+6,37</td>
<td>-1,52</td>
<td>-3,26</td>
</tr>
<tr>
<td>Fuzzy</td>
<td>-0,66</td>
<td>+0,15</td>
<td>-5,15</td>
<td>-0,83</td>
<td>-5,76</td>
</tr>
<tr>
<td>Otsu</td>
<td>-1,70</td>
<td>-5,62</td>
<td>+5,59</td>
<td>-8,79</td>
<td>-5,48</td>
</tr>
</tbody>
</table>

³ The positive values represent the improvements achieved by the method, while the negative represent performance losses
Summary

Thresholding is applied in order to remove noise and every type of information present in the image that can possibly affect the OCR result. The final result of this method isolates the foreground, easing this way the OCR process. However, during this process, information related to characters may be lost, worsen the OCR conditions.

Analysing the table with the differences between the original image and the image resulting from the thresholding methods, it is possible to observe that 3 of the 4 methods studied improved the specificity. However, these improvements also came with losses for the remaining SM. The losses, in 2 of the 3 methods that improved specificity, were higher than any improvement made. The only method that did not improve the specificity was the Fuzzy-clustering method, which increased negligibly the recall, and presented losses for the remaining SM.

However, the main achievement obtained from the thresholding methods is related with the reduction of the image file size. The initial images - which are the original images reduced by 0,50 – presented a file size of 5,8 MB, while the images obtained from the thresholded methods presented sizes of around 250 KB. In another words, through the use of thresholding techniques, the image size file is reduced more than 20 times, preserving all the OCR potential of the image.
3.3.3 Deblurring

During the digitization process, the original image is inevitably affected by a blurring effect and by the insertion of noise. These phenomena materialize through the creation of a “shadow” around the characters of ancient typed documents. This “shadow” may group together two, or more, consecutive characters, affecting considerably the performance of the OCR software[36].

To reduce this effect, deblurring methods will be applied with the purpose of sharpening the foreground, in order to achieve a higher definition of the characters, making the process of recognition to become more efficient.

Blurring is a form of bandwidth reduction of an ideal image owing to the imperfect image formation process [8]. In addition to blurring effects, noise always corrupts any recorded image. Noise may be introduced by the medium through which the image is created, by the recording medium (sensor noise or by the measurement errors due to the limited accuracy of the recording system.

A comprehensive literature review can be found in Kundur and Hatzinakos [9]. As demonstrated in Fergus et al. [10] the real PSF caused by camera shake is complex, beyond a simple parametric form (e.g., single one-direction motion or a gaussian) assumed in previous approaches[43-45]. In Fergus et al. [10] natural image statistics together with a sophisticated variational Bayes inference algorithm are used to estimate the PSF. The image is then reconstructed using a standard non-blind deconvolution algorithm.

Even with a known PSF, non-blind deconvolution [46-48] is still under-constrained. Reconstruction artifacts, e.g., “ringing” effects or color speckles, are inevitable because of high frequency loss in the blurred image. The errors due to sensor noise and quantizations of the image/PSF are also amplified in the deconvolution process. For example, more iterations in the Richardson-Lucy (RL) algorithm[40] will result in more “ringing” artifacts.

Recently, spatially variant PSF estimation has also been proposed in Bardsley et al[41]. In Levin[42], the image is segmented into several layers with different PSFs. The PSF in each layer is unidirectional and the layer motion velocity is constant.

Image denoising is a classic problem extensively studied. The challenge of image denoising is how to compromise between removing noise and preserving edge or texture. Bilateral filtering also been a simple and effective method widely used in computer graphics[42,43]. Other approaches include anisotropic diffusion[43] PDE-based methods[44], fields of experts[45] and nonlocal methods[46].

Algorithms parameters settings

The procedure used to determine a parameter for deblurring methods consisted in scanning values of different orders of magnitude, in order to guarantee that the best value was found. After determining the order of magnitude in which the ideal value is, the domain of search is centered in this same value and the limits of the domain of search assume the value selected plus or minus half of the same value, according it’s the upper or lower limit. The domain of search starts with $10^{-5}$ as the lower limit and with $10^{5}$ as the upper limit.

During this search, though some images did not show obvious improvements, these were also
analysed, leaving the decision of considering the set of parameters for further use to Finereader’s performance. This methodology was carried out using images resampled with a 0.50 RF.

**Modeling the degradation function**

The system PSF is typically a fixed or deterministic quantity that is determined by the physical hardware which constitutes the overall imaging system. Since the PSF is the result of a design and engineering process, it is reasonable to expect some useful knowledge from it, which could be used later on to assist in the restoration problem.

However, in this case, instead of having the PSF associated with a physical hardware, the PSF is related to an aging process. This aging process, rather than producing a blur, originates fading of the foreground. Even though the deblurring methods are meant to be used in different situations, they proved to be an interesting solution for these types of problems.

Thus, to achieve a successful image restoration it is required a good knowledge of the PSF and noise. Yet, when no information is available about the PSF, it is possible to resort to blind deconvolution for inferring the PSF, which is obtained simultaneously with the restored image. However, when additional information of designated point-sources is added much greater robustness and reliability is achieved.

In order to avoid the use of a blind deconvolution method, an iterative procedure was used to find a suitable PSF. This procedure consists in choosing a random non-blind deconvolution method to serve as a stand for searching a suitable PSF. In this sense, it was assumed that the PSF that originates the best result is the PSF that better approaches the true PSF.

**PSF – Search**

The search of a suitable PSF was made through the use of the Wiener deconvolution method, using 0.3 as test value for the NSR parameter. The tuning of the NSR parameter will be performed after determining which type of PSF to use. The filters studied to represent the blurring function are the filters belonging to the Matlab function *fspecial*. These filters are described in table 3.9.
Table 3.8 – Candidate filters to PSF

<table>
<thead>
<tr>
<th>Filter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>Averaging filter</td>
</tr>
<tr>
<td>Disk</td>
<td>Circular averaging filter (pillbox)</td>
</tr>
<tr>
<td>Gaussian</td>
<td>Gaussian lowpass filter</td>
</tr>
<tr>
<td>Laplacian</td>
<td>Approximates the two-dimensional Laplacian operator</td>
</tr>
<tr>
<td>LOG</td>
<td>Laplacian of Gaussian filter</td>
</tr>
<tr>
<td>Motion</td>
<td>Approximates the linear motion of a camera</td>
</tr>
<tr>
<td>Prewitt</td>
<td>Prewitt horizontal edge-emphasizing filter</td>
</tr>
<tr>
<td>Sobel</td>
<td>Sobel horizontal edge-emphasizing filter</td>
</tr>
</tbody>
</table>

The parameters of each filter were analysed using the referred methodology, in order to guarantee that every set of parameters was considered. Among all the filters studied, the Gaussian filter revealed the better results and was chosen to represent the PSF.

Methods Parameters

Taking into consideration that every non-blind deconvolution method uses the observed image and the PSF as parameters, the following analysis will focus on the specific parameters of each method.

Wiener deconvolution

This method distinguishes itself from the inverse filtering for integrating a term in the denominator that reflects the noise-to-signal ratio (NSR) present in the image. After determining the PSF to use, the next step is determining the best value for the NSR parameter. Once again, the iterative procedure used before is carried out in order to find the NSR value that would provide the best result. The analysis realized led to find 0.06 as the best value for the NSR parameter.

Constrained Least Squares (Regularized) Filtering

Likewise the Wiener method, this method integrates in the denominator a term associated with noise. In this case, the term responsible for the noise reduction is composed by two parameters: the Fourier transform of the Laplacian operator and a constant that affects it. This constant, designated by \( \gamma \), is what determines the success of the restoration and it can be found iteratively, if \( ||\eta||^2 \) is known. On the other hand, \( ||\eta||^2 \) is proportional to the noise power, making the noise power the parameter to search for.

A good starting estimate for the noise power is \( MN(\sigma^2 + \sigma_n^2) \) where \( M \) and \( N \) are the dimensions of the image, and the parameters inside the brackets are the noise variance and noise squared mean. Due to the lack of information concerning the noise distribution, the iterative procedure referred was used to determine this parameter. The final value considered for the noise power was \( 4 \times 10^3 \).
Lucy-Richardson Algorithm

In this method an iterative procedure is used to obtain an estimate of the undegraded image. The two parameters that this method requires are the PSF and the number of iterations that the algorithm may perform until obtaining the final estimation.

In theory, increasing the number of iterations would lead to an improvement of the estimation obtained, this is, it would bring the estimation closer to the undegraded image. However, increasing the number of iterations is not a guarantee of improvement. In fact, it was noticed that increasing the number of iterations beyond five iterations would have an adverse effect in the OCR. This number represents, therefore, the maximum number of iterations that the Lucy-Richardson method may complete before ringing starts to appear. It’s important to have in mind that this method relies on the assumption that the PSF is known, and that the noise distribution is governed by the Poisson density function.

Blind Deconvolution

The blind deconvolution method has as parameters an initial guess of the PSF and the maximum number of iterations that the algorithm is allowed to complete. The PSF restoration is affected strongly by the size of the initial guess and less by the values it contains [2]. For this reason, it is specified an array of 1’s as initial PSF. Regarding the size of the initial PSF, this was chosen to be equal to the size of the PSF determined previously for the non-blind deconvolution methods. Thus, the initial PSF is a matrix of ones, whose size is 5x5. As in the Lucy-Richardson, this method did not converge when the number of iterations increased beyond 5. The explanation to this phenomenon can be explained in the same way as explained before for the Lucy-Richardson algorithm, that is, the assumptions made for the algorithm are not met.
Table 3.9 - Finereader’s performance evaluation – ancient typed documents (group 3) reduced by 0,50

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,50</td>
<td>92,65</td>
<td>74,68</td>
<td>68,95</td>
<td>62,71</td>
<td>87,96</td>
</tr>
</tbody>
</table>

Table 3.10 – Finereader’s performance evaluation – deblurring methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiener</td>
<td>77,91</td>
<td>70,25</td>
<td>53,67</td>
<td>49,13</td>
<td>89,88</td>
</tr>
<tr>
<td>CLS</td>
<td>84,52</td>
<td>71,66</td>
<td>84,52</td>
<td>54,73</td>
<td>79,42</td>
</tr>
<tr>
<td>Lucy</td>
<td>85,83</td>
<td>75,67</td>
<td>58,28</td>
<td>58,24</td>
<td>91,68</td>
</tr>
<tr>
<td>Blind</td>
<td>68,96</td>
<td>62,95</td>
<td>48,76</td>
<td>39,78</td>
<td>84,70</td>
</tr>
</tbody>
</table>

Table 3.11 - Performance differences between ancient typed documents and deblurred images

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>Specificity [%]</th>
<th>Characters c.r. [%]</th>
<th>Words c.r. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiener</td>
<td>-14,74</td>
<td>-4,43</td>
<td>-15,28</td>
<td>-13,58</td>
<td>+1,91</td>
</tr>
<tr>
<td>CLS</td>
<td>-8,13</td>
<td>-3,03</td>
<td>+15,57</td>
<td>-7,99</td>
<td>-8,55</td>
</tr>
<tr>
<td>Lucy</td>
<td>-6,82</td>
<td>+0,99</td>
<td>-10,67</td>
<td>-4,47</td>
<td>+3,71</td>
</tr>
<tr>
<td>Blind</td>
<td>-23,68</td>
<td>-11,73</td>
<td>-20,19</td>
<td>-22,93</td>
<td>-3,26</td>
</tr>
</tbody>
</table>

Summary

The purpose behind the use of deblurring methods is sharpening the foreground, in order to achieve a higher definition of the characters, contributing to make the OCR process to become more efficient.

Observing the results it is possible to observe that the major improvement was made for the specificity, using the CLS method. On the other hand, this method has as drawback the decrease of the others SM. With the exception of the CLS method, the other methods revealed higher losses than any improvement made. Concerning the blind deconvolution method, this did not present any improvement, showing a worse performance than any non-blind method, in opposition to tresholding, deblurring did not reduce the size of the images.

Nevertheless, the results obtained using deblurring techniques may be the result of an inadequate modelation of the PSF, or the use of an inappropriate set of parameters. Thus, before discarding deblurring as pre-processing technique, further investigation should be carried out regarding PSF modelation and choice of method’s parameters.

---

The positive values represent the improvements achieved by the method, while the negative represent negative performances.
3.4 Discussion of Results – OCR preprocessing improvement

The goal of increasing Finereader’s performance using resizing was achieved with the increase in the Recall, percentage of characters correctly recognized and percentage of words correctly recognized, for the 0,75 RF, for modern documents.

As expected, Finereader’s performance decreases when submitted to typed documents. However, the only SM substantially affected is the specificity. However, for a 0,25 RF, the specificity is actually higher than in modern documents.

Regarding typed documents, to guarantee the quality of the correction of the OCR result, each character must be revised. Nevertheless, from the analysis made to the distribution of the characters by the four classifications it is possible to predict each character classification, using this to find more easily the FP cases.

The study realized for modern and typed documents was built upon on documents with the same type of characteristics, and with the same signs of degradation (for typed documents). For this reason, in order to reinforce the conclusions obtained, documents of different types should be included in the study.

Concerning thresholding, although improving specificity, the main achievement obtained from this technique was the drastically reduction of image file size, while preserving all the OCR potential. Despite the improvements made, none of the methods could improve Finereader’s precision or the percentage of characters correctly recognized. Thus, this technique is essentially useful for reducing the size file of an image.

Regarding deblurring, none of the methods analysed showed any significant improvement. In addition significant losses in the SM occurred. Concerning the size of the image, this method did not affect it significantly.

During the process of recognition, some words were splitted, turning into two or more words. Therefore, to allow the counting of the results, every splitted word was reunited. It should be taken into account that the number of words correctly recognized is different in every method, affecting the number of characters recognized, and, therefore, the results.
4 Metadata Extraction

Different work has been done in regard to the area of document image analysis. Muge et al.[47] described a methodology for automatic extraction and recognition of features from books of the Renaissance. Using grey level images, the approach consists in a page segmentation followed by a mathematical morphology methodology.

Pinto et al.[18] used a method that combines fuzzy clustering with mathematical morphology in order to extract interference from images of old text documents. Using a Fuzzy C-Means algorithm with 3 clusters, an image is partitioned into background, foreground and interference. Mathematical morphology is afterwards used to remove underlines and small artifacts that may still be present within the detected foreground. Edge-based segmentation has also been reported to work.

D. Frejlichowski et al.[48] proposed a method of detection, localization and segmentation of stamps (imprints) in digitized documents based on color segmentation techniques and shape analysis. Based on a color space conversion and through the projection of the image in the horizontal and vertical directions, the algorithm is able to seek for areas of high intensity. Candidates for stamps are segmented and passed to the stage where width to height proportion and standard deviation of pixel intensities are then calculated.

Nakai et al.[49] described a method for annotations extraction that considers annotations as the difference between the compensated original and annotated document images. Extracted annotations are evaluated by comparing each pixel of the ground truth and difference images. Before applying the procedures of the proposed method, degradations such as dithering, pixel value distortions, and local displacements are compensated one by one.

G. Zhu et al.[50] present a stamp detection framework based on parameter estimation of connected edge features, that introduces a template-based junction removal technique, addressing the problem of stamps overlaying with spectrum of background contents.

Algorithm

In this work is presented a methodology that integrates color segmentation and mathematical morphology techniques to automatically extract and recognise artifacts from ancient documents. This algorithm takes advantage of the fact that, in the set of documents analysed, most of the times, artifacts usually appear using featured colors, justifying the use of color segmentation techniques in the retrieval of these artifacts.

The color segmentation technique used is thresholding, and will be used to identify the artifact regions, leaving their recognition up to the mathematical morphology operations, responsible for retrieving the distinctive morphological features that characterize each artifact.

In order to simplify the colour thresholding expression, the conversion from RGB to HSV is executed. The reason behind the selection of the HSV color space is related with the hue component. This channel gives the dominant wavelength of a color, allowing to simplify the color thresholding expression and making the color segmentation to become more efficient. Using the binary images obtained with the thresholding, a mathematical morphology based algorithm follows next in order to retrieve the regions with the morphological features that characterize an artifact. Fundamentally, these
operations seek for the straight lines that characterize the artifact sought. To ensure that only areas belonging to artifacts are considered, an area opening operation is executed to remove residual areas. The final criterion to infer about the presence of an artifact is an area threshold, which considers the presence of an artifact if the number of pixels is equal or higher than a fixed threshold. For stamps, in addition to this criterion, the number of vertical lines has to equal to be two. Regarding tables, the number of both vertical and horizontal lines has to be equal or higher than two.

Analysing the set of documents analysed, it was assumed that tables and stamps are the only artifacts showing vertical lines, while underlines, besides tables and stamps, are the only artifacts exhibiting horizontal lines. Annotations, due to absence of a morphological feature to look for, are considered to be the remaining parts of the colored region after the removal of all the others artifacts.

Considering that some artifacts share common morphological key features, artifacts may be incorrectly classified. To avoid inaccurate classifications, after the recognition of an artifact, the feature extraction step is carried out again, using this time a more tolerant structuring element, and subjecting the result to a morphologic reconstruction operation in order to retrieve the whole artifact, removing it afterwards from the binary image that is going to be used to recognize the next artifact.

The sequence of artifacts recognized is hereafter described. Tables, for having longer lines than stamps, are the first artifacts to be identified. After being recognized, these are removed from the image, leaving stamps as the only artifacts exhibiting vertical lines. Stamps follow next in the recognition process. Once again, after being identified, these are also removed, due the presence of horizontal underlines present in the top and bottom of the stamps that could be understood as underlines in the underline recognition process. Finally, after the recognition of the tables, stamps and underlines, the only artifacts remaining should be the annotations.

Due to the presence of documents with blue foreground, the recognition of blue annotations was not performed, for the reason that this work assumes that artifacts have a different color from the background and foreground. In this sense, only red annotations were considered. Nevertheless, not all the red areas remaining after the removal of all the other artifacts can be considered as annotations. Some documents present artifacts such as signatures, dates, serial numbers, among others inscriptions, that were introduced with the red color. In addition, red typed text was also noticed in the set of documents analysed. Although the referred artifacts and the red typed text cannot be considered as regular annotations, considering that the use of the red color has the purpose of highlighting relevant content, these were perceived as such. The recognition process of an artifact, after the color thresholding stage, is repeated for the different colors in that artifact may appear. The full algorithm is presented next in Fig. 4-1.
Fig. 4-1 - Metadata Extraction algorithm
4.1 Artifacts Recognition Algorithm

4.1.1 Color Channel Conversion

Arts such as stamps, underlines and annotations are often introduced with the intention of excel their contents. To achieve this goal, a particular set of colors is usually used: red and blue. The color segmentation stage will be responsible for the retrieval of these colors, in order to find possible areas where artifacts are expected to exist. The color segmentation approach applied is based on thresholding. The technique used consists in establishing a set of values for the components of the color sought and to compare them with color components of each pixel, in order to evaluate if it can be classified as belonging to the color in analysis. The thresholding expression, for reds and blues, is defined in terms of hue according with the typical values found on the artifacts. However, blacks and whites, for being considered as the entirety and the absence of all colors, respectively, do not have a dominant wavelength, which allows them to satisfy any threshold expression based exclusively in a restriction of the hue channel. Thus, to avoid blacks and whites to be recognized as blues or reds, restrictions in the saturation and value were included. The thresholding expression for reds requires that the saturation of these to be higher than 0.23. Due to the absence of cases in which black is classified as red, restrictions regarding the value channel were not included. According with the values observed for red, the hue and saturation values to classify a color as red are:

\[
0.00 < H < 0.07 \text{ or } 0.95 < H < 1.00 \\
S > 0.23
\]  

(4.1)

Considering that only three colors (blacks, reds and blues) are present in the foreground of the documents analysed, and that blacks can be avoided with a restriction in the value channel, the exclusion of reds is supposed to preserve only the blues. In this sense, blues can also be perceived as non-reds. To ensure that any red would be considered as blue, a gap between the two sets was established. The saturation restriction for the identification of blues establishes that saturation has to be higher than 0.02 and lower than 0.15. To avoid the presence of blacks, the set of values for the value channel establishes that these have to be higher than 0.50 and lower than 0.90. Thus the total thresholding expression for blues is:

\[
0.12 < H < 0.90 \\
0.02 < S < 0.15 \\
0.50 < V < 0.90
\]

(4.2)

The recognition of blacks will be done using a different thresholding technique - the Sauvola method. In order to apply this method, a conversion to grey levels has to be executed.

The color segmentation stage is not applied in the same way for all the artifacts recognition processes.
Tables, for instance, usually appear in black, which makes their search in red or blue regions to be senseless. In this sense, the color thresholding stage focus only on the colors that usually features the artifacts. In the set analyzed, the colors associated to stamps are black and blue. Underlines appear most of the times in red and blue, although some cases of black underlines exist. Finally, annotations appear in the color of red and blue. Nevertheless, as referred, only red annotations will be considered. An example of the results obtained in this stage, for the three colors referred, is presented in Fig. 4.2
Fig. 4-2 - a) original image; b) black threshold; c) blue threshold; d) red threshold
4.1.2 Feature extraction

The morphological features identification stage is responsible for the retrieval of the key morphological features that characterizes each artifact. This stage consists in applying an erosion process to the binary image obtained in the color thresholding stage, through the use of an appropriate structuring element, with the purpose of removing all the areas that are not related with the key morphological features of the artifact searched.

Tables
The features sought in the recognition of tables are vertical and horizontal lines. Although others artifacts exhibit these in their morphology, the lines present in tables are by far greater than the lines present in any other artifact. For this reason, the image is eroded once using a vertical line for the detection of the vertical lines of the tables, having the same process repeated for horizontal lines, using an horizontal line as structuring element. For being considered that a table is composed for at least 2 vertical lines and 2 horizontal lines, in case a table is present in the image, the result of each erosion should present at least 2 lines in both directions. The chance of misleading a horizontal straight line from a table with an underline, although possible, is minored by the use of a greater structuring element that should only detect lines from tables. An example of table recognition is presented in figure 4-3.
Fig. 4-3 – a) image containing table; b) Sauvola threshold; c) table key feature - vertical lines; d) table key feature - horizontal lines
In order to demonstrate the algorithm performance in the retrieval of stamps, underlines and annotations another image was used. This image, is presented in Fig. 4-4, while the result of the color thresholding stage, that is, from the black and red threshold, is presented in fig. 4-5. The blue threshold is not presented, since there is not any blue region present in the image.

Fig. 4-4 - Image containing stamps, underlines and annotations
Fig. 4-5 - a) black (Sauvola) threshold; b) red threshold
Stamps
After the recognition and removal of tables, the remaining vertical lines present in an image should come from stamps, which motivate the use of vertical line as structuring element. However, due to the smaller size that the vertical lines of stamps exhibit, the mask needed to retrieve this morphological features is also smaller. The result obtained in this stage for stamps is presented in fig. 4-6.

Underlines
Apart from tables and stamps, underlines are the only artifacts that present straight horizontal lines in their morphology. This motivates the use of an horizontal line as structuring element in the erosion process. The result obtain in this stage for underlines is presented in fig. 4-6

Annotations
To what concerns annotations, due to their arbitrary shape, and considering the absence of a morphological feature to look for, instead of an erosion process, a subtraction process takes place. With all the artifacts detected so far are removed from the binary image obtained, any significant area persisting will be, therefore, considered as annotations.

![Fig. 4-6 - a) key features of stamps; b) key feature of underline](image)

Table 4.1 - Structuring elements – feature extraction

<table>
<thead>
<tr>
<th>Direction</th>
<th>Size [pixels]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tables</td>
<td>Horizontal and Vertical 180 and 180</td>
</tr>
<tr>
<td>Stamps</td>
<td>Vertical 65</td>
</tr>
<tr>
<td>Underlines</td>
<td>Horizontal 75</td>
</tr>
</tbody>
</table>
4.1.3 Morphological Reconstruction (MR)

After the color thresholding stage, the artifacts previously marked are removed in order to avoid interferences in the recognition of the artifact to be recognized. To remove an artifact, this needs to be entirely detected. To achieve this, a morphological reconstruction operation is applied to the result of an erosion process used to identify the morphological features that characterizes the artifact to be removed.

MR can be seen as a set of repeated dilations of an image, called the marker image, until the contour of the marker image fits under a second image, called the mask image[51]. In this case, the marker is the image resulting from the erosion process and the mask is the result of the color thresholding stage. The MR resulting image - a binary image in which the artifacts to be removed appear in white, in a black background – is added to the initial image, obtained in the color thresholding stage. Adding the image with the artifact to be removed to the one obtained in the color thresholding stage will convert the regions that constitute the artifact to be removed into background, eliminating their presence of the image that is going to be used for the recognition of the next artifact. It should be noticed that the erosion process used in the feature extraction stage is different from the one used to create the mask for the morphological reconstruction operation. The difference consists in the use of a greater structuring element. The size of the structuring element used in the erosion process is directly related with restrictiveness used in the process, that is, with the severity with which the areas are eroded. In the feature extraction stage, the only areas expected to remain are the ones related with the morphological features sought. Thus, the use of a greater structuring element in the erosion process is expected to produce more reliable result. Regarding the erosion process used to obtain the mask for the morphological reconstruction process, a smaller structuring element should be considered, in view of finding, at least, part of the morphological feature that characterizes an artifact. Oversizing the structuring element can conduct to a severe erosion that would remove all the traces of the morphological features that characterize the artifact. The removal of these features would drive the algorithm to assume the non-existence of the artifact to be reconstructed. Thus, the use of a smaller, less restrictive structuring element is expected to be more appropriate in the retrieval of morphological features parts’ of the artifact to be reconstructed, allowing possible its further removal.

Tables, for being the first artifact to be detected, are the only artifact skipping the artifact removal stage. In the morphological reconstruction of stamps, the morphological features used to build the mask for the morphological reconstruction process are, as referred, the vertical straight lines present in the border of the stamp. Hopefully, the result of the MR should present the entire border of the stamp. Nevertheless, stamps often contain in their interior other vertical and horizontal lines that may interfere in the recognition process of underlines and annotations. In this sense, to avoid these to interfere in the recognition of underlines and annotations, in addition to the borders, the interior of the stamp should also be removed.

The removal of the entire stamp is then achieved through the use of a bounding box, which will after convert all of its pixels into white. The bounding box has its limits defined by the areas retrieved as the
sides of the stamp and will assume as its left corner the higher left point of the left vertical line that defines the stamp’s border and use as edges the distance between this point and the remaining corners, which were defined in the same ways as the left upper corner. With the bounding box defined, the pixels contained in the inside of the bounding box will assume the value of the background pixels, preventing any interference from these underline and annotation recognition process. The following table presents the values used in the recognition of the morphological features that characterizes the artifacts

Table 4.2 - Structuring elements – artifact removal

<table>
<thead>
<tr>
<th>Direction</th>
<th>Size [pixels]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tables</td>
<td>Horizontal and Vertical 150 and 150</td>
</tr>
<tr>
<td>Stamps</td>
<td>Vertical 40</td>
</tr>
<tr>
<td>Underlines</td>
<td>Horizontal 25</td>
</tr>
</tbody>
</table>

The removal of all the artifacts, for appearance purposes, is achieved by extracting all the recognized artifacts in the three colors segmented, from the original image. The process starts with the application of the Sauvola method to the original image in order to retrieve and remove all black artifacts. Next, the original image is multiplied by the resulting binary obtained with all the black artifacts removed, resulting in the image presented in fig. 4-7. In the next stage this image will go through the color thresholding stage to retrieve all the red areas, in order to make their subsequent removal possible. The process stops in the red artifact removal due to the non-existence of blue regions. The artifacts retrieved, and removed, are displayed in figure 4-7. The complete removal of all artifacts present in the original image can be seen in fig. 4-8.
Fig. 4-7 - a) black artifacts removed; b) stamp recognized; c) underlines recognized; d) annotations recognized
Fig. 4-8 – a) original image; b) image without artifacts
4.1.4 Residual Area removing

The feature extraction stage is expected to remove all areas that are non-related with the morphological features that characterize the artifacts. The exclusion of these areas will allow inferring about the presence of artifacts in case significant areas still persist. However, some areas that are non-related with these features may pass through the erosion process. Despite appearing in smaller sizes, the sum of their areas may overcome the size of the areas related with the morphological features sought, affecting the conclusion regarding the presence of artifact. Thus, in order to ensure that all the areas present in the image resulting from the morphological feature identification stage are related with the morphological features that characterize artifacts, an area opening takes place.

The filter that removes from a binary image its connected components with area smaller than a parameter is called area opening[52].

The threshold value that separates significant areas from residual areas was fixed at 100 for tables, underlines and annotations, while for stamps this was fixed at 50.

4.1.5 Area Thresholding

The decision regarding the presence of artifacts occurs in this stage. The criterion used to infer about the presence of an artifact is based on the size of the areas related with morphological features that characterize an artifact. Thus, an image is classified as containing an artifact, if the area of the morphological features that characterizes the artifacts is superior than a fixed threshold. This criterion aims to ensure the presence of a significant area related with the morphological features that characterizes artifact. The area threshold for the presence of tables, underlines and stamps was fixed in 50 pixels and for annotations in 500.

Regarding the recognition of tables and stamps, an additional criterion is used. To classify an image as containing tables, the number of vertical and the number of horizontal areas has to be larger than two, for being considered that tables are composed, at least, by two vertical and horizontal lines. Regarding stamps, the criterion only requires the presence of two areas, related with the sides of the stamp.
4.2 Discussion of Results – Artifact Recognition

In order to test the proposed techniques, the new approach was applied to a representative set of documents from the IICT Library. The results of table recognition were conditioned by the low number of tables present in the set analysed. The lines that define tables are usually drawn with the same color as text, which makes their recognition to rely exclusively on the morphologic stage. Moreover, a considerable part of these had their borders defined with dashed lines, leading the algorithm to incorrect classifications. The mask used in the erosion process spared lines from margins and from annotations with vertical strokes, which were after considered as being part of tables, leading to inaccurate classifications. To avoid this, a criterion regarding the number of vertical and horizontal lines was implemented in decision stage to classify an image as containing tables.

Stamp recognition, for using the same process of table removal, with the difference of using a smaller vertical line as mask, faces the same problems of table recognition. However, due to the feature color with which stamps often appear, their recognition is favored by the color thresholding stage, which excludes some of the agents responsible for inaccuracy.

Still, the presence of annotations with vertical straight lines was found to be one of the main causes of inaccuracy in the recognition of stamps. However, most of the inaccurate classifications were due to fading.

Fading is a critical problem when occurring in the interior of the morphological feature that characterizes an artifact, if considered that the less pronounced parts of the morphological features may be despised during the color thresholding stage. Despising these parts may lead to the disintegration of the morphological feature, which will cause their removal when applying area opening, in the removal of residual areas stage, driving the algorithm to inaccurate classifications. A foreseen drawback in the morphological stage was related with the rotation that stamps might present. However, the mask used was able to tolerate considerable rotations.

To what concerns the recognition of underlines, this was mostly conditioned by their consistency. The pressure variations that occur in the underlining process conduct to a fading of some parts of the underline. In addition, underlines that presented considerable undulations, or that are angled, may be severely eroded in the erosion process, in the feature extraction stage, leaving behind only a smaller part of their extension, conducting to a situation similar to the one caused by fading. A possible solution to identify the entirety of an underline may rely on morphological reconstruction. Nevertheless, in cases where the underlines overlap the text, the MR will recognize text as being part of the underlines. In the documents analysed, the annotations usually appear in red. Considering that red appears in the extremes regions of the hue channel, its identification is easily processed, allowing an accurate thresholding expression and therefore an accurate recognition of all the red coloured artifacts. However, not all red areas are related to these artifacts. Some documents present red typed characters in headings, headlines, titles and others that are part of the original document. Although they were not introduced after the creation of the document, as a regular annotation, they still point relevant content, which led to consider those as annotations. The results obtained for artifact recognition are presented in Table 4.3.
Table 4.3 – Artifact recognition results

<table>
<thead>
<tr>
<th></th>
<th>Quantity</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tables</td>
<td>10</td>
<td>87,50</td>
<td>70,00</td>
<td>99,10</td>
</tr>
<tr>
<td>Stamps</td>
<td>37</td>
<td>68,18</td>
<td>81,08</td>
<td>83,52</td>
</tr>
<tr>
<td>Underlines</td>
<td>71</td>
<td>82,10</td>
<td>90,70</td>
<td>52,78</td>
</tr>
<tr>
<td>Annotations</td>
<td>58</td>
<td>84,61</td>
<td>94,82</td>
<td>84,37</td>
</tr>
</tbody>
</table>

Total: 122
5 Conclusions and Future Work

Ancient documents often present multiple types of degradation. Naturally, this degradation is visible in the images that result upon the digitization process. This makes it harder to work with those images, not only from a human point of view – the readers and, therefore, the users of those images – but also from a machine point of view – processes that perform recognition and extract information from images. Set out with well-defined goals, the intent of this thesis was to research whether it was possible to automatically restore images of ancient documents in order to improve OCR operations and extracting metadata regarding the presence of artifacts. Research was then conducted in order to determine which of the available methods could be used for OCR improvement, and to study ways of metadata extraction regarding the presence of artifacts.

It was shown that resizing the image to a lower dimension, till a reduction factor of about 0.5 has all the advantages of manipulating shorter files without degrading the OCR performance even improving some measures. Other techniques were experimented to improve the OCR performance such us several adaptive binarization and deblurring algorithms. Although some improvements were registered results are not yet satisfactory when the documents ink is too faded.

Another contribution of this work was the survey about recognition of problematic characters. From the analysis made to the distribution of the characters by the four classifications, for modern and typed documents, it is possible to predict each character classification, using this to find more easily inaccurate recognized characters that are not marked by the software.

A considerable contribution of this part of the work concerns the results obtained from thresholding. The images resulting from this technique, whose file sizes were nearly 20 times smaller than the original images files, presented approximately the same OCR results than the original ones, even improving some statistical measures.

Regarding metadata extraction, an automatic and efficient recognition approach of artifacts such as tables, stamps, underlines and annotations was developed. This method provides a combination of techniques as well as some new ideas that were developed.

It is worth noticing that generalized Hough transform are not used to detect lines or other regular shapes, since it is a very time-consuming. In this sense, the approach is much more flexible and less complex than the one presented in [50].

The algorithm presented a high degree of generalisation as the current experiments indicate, even though the differences present in the set of images analysed. Furthermore, the feature extraction method developed allowed to generate an artifact free version of the documents.

Future Work

Far from being perfect systems, the results of this work can and should be improved upon and complemented before being integrated with a large scale infrastructure, as they hopefully will be, namely in digital media projects of the IICT.

Thus, the work that should follow this thesis should include a bigger set of ancient documents images,
in order to reinforce the conclusions obtained regarding the survey of problematic characters recognition.

Considering that the efficiency of deblurring methods are directly linked to the accuracy with which the imaging degradation process is modelled, the unsatisfactory results obtained with the deblurring methods may be improved by performing research in the PSF area.

Due to the variety of types of documents present in institutions, the thresholding and deblurring techniques could benefit from the development of an automatic parameter adjustment system based on measurable properties of the documents being processed.

Regarding metadata extraction, the developed method may be subject of further improvement by including a stage where the documents are segmented in text and non-text areas, restricting the search of underlines and annotations to where they are expected to appear. This restriction would allow looking for annotations with the same color of the text. Further developed fuzzy partnership relations combined with neural networks decision rules are expected to improve results in the proposed task of text and non-text area separation. The recognition of stamps was restricted to rectangular forms. Thus, further work should be developed in order to extend the recognition of stamps to other forms.

Finally, some research is still necessary in order to fully automate the application of these methodologies to the mass production of ancient documents.


6 Appendix

Modern

Table 6.1 - Characters classified as FP in more than 10% of the cases in modern documents

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>24</td>
<td>37.5</td>
</tr>
<tr>
<td>â</td>
<td>17</td>
<td>11.8</td>
</tr>
<tr>
<td>ð</td>
<td>46</td>
<td>10.9</td>
</tr>
<tr>
<td>,</td>
<td>348</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Table 6.2 - Characters classified as TP in less than 10% of the cases in modern documents

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ð</td>
<td>46</td>
<td>2.2</td>
</tr>
<tr>
<td>I</td>
<td>24</td>
<td>8.3</td>
</tr>
</tbody>
</table>

5 The others RF don’t have any character classified as FP in more than 10% of the cases
Table 6.3 - Characters classified as TN in more than 10% of the cases in modern documents

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>۴</td>
<td>۲۶</td>
<td>۵۰,۰</td>
</tr>
<tr>
<td>۴</td>
<td>۴۷</td>
<td>۲۷,۷</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>۴</td>
<td>۲۶</td>
<td>۵۰,۰</td>
</tr>
<tr>
<td>۴</td>
<td>۴۵</td>
<td>۱۳,۳</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>۴</td>
<td>۲۷</td>
<td>۵۹,۳</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>۴</td>
<td>۴۵</td>
<td>۴۶,۷</td>
</tr>
<tr>
<td>۴</td>
<td>۱۷</td>
<td>۴۱,۲</td>
</tr>
<tr>
<td>۴</td>
<td>۴۶</td>
<td>۳۷,۵</td>
</tr>
<tr>
<td>۴</td>
<td>۳۶</td>
<td>۱۱,۱</td>
</tr>
<tr>
<td>۴</td>
<td>۵۳</td>
<td>۱۳,۲</td>
</tr>
<tr>
<td>۴</td>
<td>۴۶</td>
<td>۱۹,۶</td>
</tr>
</tbody>
</table>
Table 6.4 - Characters classified as FN in more than 25% of the cases in modern documents

<table>
<thead>
<tr>
<th></th>
<th>1,00</th>
<th></th>
<th>0,75</th>
<th></th>
<th>0,50</th>
<th></th>
<th>0,25</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>char</td>
<td></td>
<td>events</td>
<td>rate</td>
<td>char</td>
<td></td>
<td>events</td>
<td>rate</td>
<td>char</td>
</tr>
<tr>
<td>!</td>
<td>20</td>
<td>55,0</td>
<td></td>
<td>¨</td>
<td>49</td>
<td>46,9</td>
<td></td>
<td>ò</td>
</tr>
<tr>
<td>ò</td>
<td>47</td>
<td>42,6</td>
<td></td>
<td>'</td>
<td>45</td>
<td>42,2</td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>O</td>
<td>22</td>
<td>36,4</td>
<td></td>
<td>I</td>
<td>20</td>
<td>40,0</td>
<td></td>
<td>'</td>
</tr>
<tr>
<td>'</td>
<td>46</td>
<td>34,8</td>
<td></td>
<td>O</td>
<td>22</td>
<td>36,4</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>I</td>
<td>26</td>
<td>27,0</td>
<td></td>
<td>I</td>
<td>26</td>
<td>62,2</td>
<td></td>
<td>!</td>
</tr>
<tr>
<td>:</td>
<td>51</td>
<td>46,9</td>
<td></td>
<td>:</td>
<td>51</td>
<td>46,9</td>
<td></td>
<td>ó</td>
</tr>
<tr>
<td>U</td>
<td>17</td>
<td>29,4</td>
<td></td>
<td>U</td>
<td>17</td>
<td>29,4</td>
<td></td>
<td>Ü</td>
</tr>
<tr>
<td>ä</td>
<td>17</td>
<td>35,3</td>
<td></td>
<td>H</td>
<td>12</td>
<td>33,3</td>
<td></td>
<td>ç</td>
</tr>
<tr>
<td>!</td>
<td>20</td>
<td>30,0</td>
<td></td>
<td>!</td>
<td>20</td>
<td>30,0</td>
<td></td>
<td>í</td>
</tr>
<tr>
<td>y</td>
<td>46</td>
<td>28,3</td>
<td></td>
<td>y</td>
<td>46</td>
<td>28,3</td>
<td></td>
<td>J</td>
</tr>
</tbody>
</table>
Ancient

<table>
<thead>
<tr>
<th></th>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,00</td>
<td>A</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>à</td>
<td>56</td>
<td>62,5</td>
</tr>
<tr>
<td></td>
<td>ª</td>
<td>8</td>
<td>25,0</td>
</tr>
<tr>
<td></td>
<td>ò</td>
<td>9</td>
<td>22,2</td>
</tr>
<tr>
<td>0,75</td>
<td>o</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>à</td>
<td>51</td>
<td>56,9</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>2</td>
<td>50,0</td>
</tr>
<tr>
<td></td>
<td>à</td>
<td>6</td>
<td>16,7</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>14</td>
<td>14,3</td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>58</td>
<td>10,3</td>
</tr>
<tr>
<td>0,50</td>
<td>:</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>à</td>
<td>54</td>
<td>74,1</td>
</tr>
<tr>
<td>0,25</td>
<td>á</td>
<td>56</td>
<td>57,1</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>12</td>
<td>25,0</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>12</td>
<td>16,7</td>
</tr>
<tr>
<td></td>
<td>®</td>
<td>8</td>
<td>12,5</td>
</tr>
<tr>
<td></td>
<td>í</td>
<td>8</td>
<td>12,5</td>
</tr>
<tr>
<td></td>
<td>é</td>
<td>34</td>
<td>11,8</td>
</tr>
</tbody>
</table>

Table 6.5 - Characters classified as FP in more than 10% of the cases in ancient documents
Table 6.6 - Characters classified as TN in more than 10% of the cases in ancient documents

<table>
<thead>
<tr>
<th></th>
<th>1.00</th>
<th></th>
<th>0.75</th>
<th></th>
<th>0.50</th>
<th></th>
<th>0.25</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>char</td>
<td>events</td>
<td>rate</td>
<td></td>
<td>char</td>
<td>events</td>
<td>rate</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>75.0</td>
<td></td>
<td></td>
<td>7</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>50.0</td>
<td></td>
<td></td>
<td>7</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>50.0</td>
<td></td>
<td></td>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>8</td>
<td>37.5</td>
<td></td>
<td></td>
<td>6</td>
<td>50.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>à</td>
<td>3</td>
<td>33.3</td>
<td></td>
<td></td>
<td>1</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>'</td>
<td>9</td>
<td>33.3</td>
<td></td>
<td></td>
<td>58</td>
<td>48.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ò</td>
<td>13</td>
<td>23.1</td>
<td></td>
<td></td>
<td>13</td>
<td>46.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>9</td>
<td>22.2</td>
<td></td>
<td></td>
<td>67</td>
<td>37.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>6</td>
<td>16.7</td>
<td></td>
<td></td>
<td>16.7</td>
<td>16.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>62</td>
<td>14.5</td>
<td></td>
<td></td>
<td>14.3</td>
<td>14.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>25.0</td>
<td></td>
<td></td>
<td>33</td>
<td>33.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>j</td>
<td>23</td>
<td>17.4</td>
<td></td>
<td></td>
<td>60</td>
<td>60.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>)</td>
<td>5</td>
<td>60.0</td>
<td></td>
<td></td>
<td>55</td>
<td>55.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>44.4</td>
<td></td>
<td></td>
<td>35</td>
<td>35.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ç</td>
<td>40</td>
<td>35.0</td>
<td></td>
<td></td>
<td>33</td>
<td>33.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>240</td>
<td>33.3</td>
<td></td>
<td></td>
<td>32</td>
<td>32.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.7 - Characters classified as FN in more than 25% of the cases in ancient documents

<table>
<thead>
<tr>
<th></th>
<th>1,00 char events rate</th>
<th>0,75 char events rate</th>
<th>0,50 char events rate</th>
<th>0,25 char events rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>;</td>
<td>1 100</td>
<td>y 3 66,7</td>
<td>K 1 100</td>
<td>K 1 100</td>
</tr>
<tr>
<td>_</td>
<td>1 100</td>
<td>) 5 60,0</td>
<td>y 3 100</td>
<td>Q 1 100</td>
</tr>
<tr>
<td>k</td>
<td>1 100</td>
<td>j 23 52,2</td>
<td>) 5 80,0</td>
<td>y 3 100</td>
</tr>
<tr>
<td>°</td>
<td>1 100</td>
<td>W 2 50,0</td>
<td>ó 9 66,7</td>
<td>` 3 66,7</td>
</tr>
<tr>
<td>ô</td>
<td>13 69,2</td>
<td>ó 8 50,0</td>
<td>j 22 59,1</td>
<td>j 23 65,2</td>
</tr>
<tr>
<td>'</td>
<td>3 66,7</td>
<td>D 9 44,4</td>
<td>&quot; 8 50,0</td>
<td>) 5 60,0</td>
</tr>
<tr>
<td>y</td>
<td>3 66,7</td>
<td>B 7 42,9</td>
<td>ô 9 55,6</td>
<td>' 9 55,6</td>
</tr>
<tr>
<td>)</td>
<td>5 60,0</td>
<td>é 28 39,3</td>
<td>à 24 50,0</td>
<td>W 2 50,0</td>
</tr>
<tr>
<td>ô</td>
<td>2 50,0</td>
<td>ó 13 38,5</td>
<td>ŭ 22 59,1</td>
<td>á 24 50,0</td>
</tr>
<tr>
<td>&quot;</td>
<td>9 44,4</td>
<td>&quot; 6 33,3</td>
<td>é 34 35,3</td>
<td>à 20 50,0</td>
</tr>
<tr>
<td>j</td>
<td>23 43,5</td>
<td>' 3 33,3</td>
<td>' 34 32,1</td>
<td>W 2 50,0</td>
</tr>
<tr>
<td>ô</td>
<td>9 33,3</td>
<td>à 6 33,3</td>
<td>à 34 32,4</td>
<td>â 7 28,6</td>
</tr>
<tr>
<td>D</td>
<td>11 27,3</td>
<td>é 10 30,0</td>
<td>- 31 29,0</td>
<td>à 7 28,6</td>
</tr>
</tbody>
</table>
Table 6.8 - Characters classified as TP in less than 10% of the cases in ancient documents

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>'</td>
<td>3</td>
<td>0,0</td>
</tr>
<tr>
<td>:</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>;</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>_</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>k</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>o</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>ã</td>
<td>8</td>
<td>0,0</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0,0</td>
</tr>
<tr>
<td>ô</td>
<td>2</td>
<td>0,0</td>
</tr>
<tr>
<td>ô</td>
<td>13</td>
<td>7,7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>'</td>
<td>3</td>
<td>0,0</td>
</tr>
<tr>
<td>:</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>;</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>_</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>K</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>0,0</td>
</tr>
<tr>
<td>õ</td>
<td>6</td>
<td>0,0</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0,0</td>
</tr>
<tr>
<td>à</td>
<td>7</td>
<td>0,0</td>
</tr>
<tr>
<td>ô</td>
<td>2</td>
<td>0,0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>char</th>
<th>events</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>'</td>
<td>3</td>
<td>0,0</td>
</tr>
<tr>
<td>:</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>;</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>_</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>K</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>Q</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>W</td>
<td>2</td>
<td>0,0</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>0,0</td>
</tr>
<tr>
<td>õ</td>
<td>1</td>
<td>0,0</td>
</tr>
<tr>
<td>õ</td>
<td>8</td>
<td>0,0</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0,0</td>
</tr>
<tr>
<td>à</td>
<td>7</td>
<td>0,0</td>
</tr>
<tr>
<td>à</td>
<td>56</td>
<td>0,0</td>
</tr>
<tr>
<td>ô</td>
<td>9</td>
<td>0,0</td>
</tr>
<tr>
<td>ô</td>
<td>2</td>
<td>0,0</td>
</tr>
</tbody>
</table>