UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ELETRÔNICA E SISTEMAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MARCOS ANTONIO MARTINS DE ALMEIDA

STATISTICAL ANALYSIS APPLIED TO DATA CLASSIFICATION AND IMAGE FILTERING

RECIFE
2016

MARCOS ANTONIO MARTINS DE ALMEIDA

STATISTICAL ANALYSIS APPLIED TO DATA CLASSIFICATION AND IMAGE FILTERING

Thesis presented to UFPE as a partial fulfillment of the requirements for the degree of Doctor in Electrical Engineering
Line of Research: Image Processing

Supervisor: Prof. Dr. Rafael Dueire Lins.

RECIFE
2016

Catalogação na fonte
Bibliotecária Maria Luiza de Moura Ferreira, CRB-4 / 1469

A447s Almeida, Marcos Antonio Martins de.
Statistical analysis applied to data classification and image filtering / Marcos Antonio Martins de Almeida. - 2016.
158 folhas, il., gráfs., tabs.

Supervisor: Prof. Dr. Rafael Dueire Lins.
Thesis (Doutorado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Elétrica, 2016.
Include References and Appendix.

1. Electrical Engineering. 2. Data processing. 3. Data classification. 4. Image filtering. I. Lins, Rafael Dueire (Orientador). II. Título.

UFPE

621.3 CDD (22. ed.) BCTG/2017-217

Universidade Federal de Pernambuco
Pós-Graduação em Engenharia Elétrica

PARECER DA COMISSÃO EXAMINADORA DE DEFESA DE TESE DE DOUTORADO DE

MARCOS ANTONIO MARTINS DE ALMEIDA

TÍTULO

"STATISTICAL ANALYSIS APPLIED TO DATA CLASSIFICATION AND IMAGE FILTERING"

A comissão examinadora composta pelos professores: RAFAEL DUEIRE LINS, DEINFO/UFRPE; VALDEMAR CARDOSO DA ROCHA JÚNIOR, DES/UFPE; GABRIEL DE FRANÇA PEREIRA E SILVA, UAG/UFRPE; SÍLVIO DE BARROS MELO, CIN/UFPE e MARIA LENCASTRE PINHEIRO DE MENEZES CRUZ , POLI/UPE, sob a presidência do primeiro, consideram o candidato **MARCOS ANTONIO MARTINS DE ALMEIDA APROVADO.**

Recife, 21 de dezembro de 2016.

MARCELO CABRAL CAVALCANTI
Coordenador do PPGEE

RAFAEL DUEIRE LINS
Orientador e Membro Titular Interno

GABRIEL DE FRANÇA PEREIRA E SILVA
Membro Titular Externo

VALDEMAR CARDOSO DA ROCHA JÚNIOR
Membro Titular Interno

SÍLVIO DE BARROS MELO
Membro Titular Externo

MARIA LENCASTRE PINHEIRO DE MENEZES CRUZ
Membro Titular Externo

I dedicate this thesis to my family

AGRADECIMENTOS

Agradeço a Deus, fonte de vida, amor e justiça em minha vida.

À minha família, minha mãe Dalva (in memoriam) e minhas irmãs Marilene e Maridalva e cunhados Gesildo e Freitas (in memoriam), pelo carinho e atenção, como também a minha esposa Suely Almeida, pela paciência, dedicação e apoio, juntamente com meus filhos Débora Almeida e Victor Almeida, fontes de motivação na minha vida.

Ao amigo e orientador prof. Dr. Rafael Dueire Lins, pela oportunidade de participar do seu grupo de alunos orientandos e poder desfrutar do seu grande conhecimento e de sua vasta cultura e que contribuiu sobremaneira para o meu aprendizado em Processamento de Imagens e áreas afins. O seu exemplo como Professor e orientador me contagiou e motivou para a elaboração dessa tese.

Ao professor Fernando Campello pelos valiosos conselhos e orientações ainda na época mestrado. Ao amigo e irmão na fé professor Frederico Dias Nunes pelo grande incentivo e apoio. Aos professores do Programa de Pós-Graduação em Engenharia Elétrica da UFPE, pela dedicação e atenção dispensados durante o curso, e que muito contribuiram para o meu aprendizado. Aos funcionários do PPGEE, em especial a Andréa Tenório, secretária do Programa, pela atenção e orientação nos assuntos formais e administrativos do curso. À Profa. Sidney Ann Pratt pela verificação e correção dos textos iniciais da tese.

Muito Grato a todos.

Marcos Antonio Martins de Almeida

1. Now faith is the substance of things hoped for, the evidence of things not seen.

2. For by it the elders obtained a good report.

3. Through faith we understand that the worlds were framed by the word of God, so that things which are seen were not made of things which do appear.

[Hebrews 11]

ABSTRACT

Statistical analysis is a tool of wide applicability in several areas of scientific knowledge. This thesis makes use of statistical analysis in two different applications: data classification and image processing targeted at document image binarization. In the first case, this thesis presents an analysis of several aspects of the consistency of the classification of the senior researchers in computer science of the Brazilian research council, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico. The second application of statistical analysis developed in this thesis addresses filtering-out the back to front interference which appears whenever a document is written or typed on both sides of translucent paper. In this topic, an assessment of the most important algorithms found in the literature is made, taking into account a large quantity of parameters such as the strength of the back to front interference, the diffusion of the ink in the paper, and the texture and hue of the paper due to aging. A new binarization algorithm is proposed, which is capable of removing the back-to-front noise in a wide range of documents. Additionally, this thesis proposes a new concept of “intelligent” binarization for complex documents, which besides text encompass several graphical elements such as figures, photos, diagrams, etc.

Keywords: Data processing. Data classification. Image filtering.

RESUMO

Análise estatística é uma ferramenta de grande aplicabilidade em diversas áreas do conhecimento científico. Esta tese faz uso de análise estatística em duas aplicações distintas: classificação de dados e processamento de imagens de documentos visando a binarização. No primeiro caso, é aqui feita uma análise de diversos aspectos da consistência da classificação de pesquisadores sêniores do CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, na área de Ciência da Computação. A segunda aplicação de análise estatística aqui desenvolvida trata da filtragem da interferência frente-verso que surge quando um documento é escrito ou impresso em ambos os lados da folha de um papel translúcido. Neste tópico é inicialmente feita uma análise da qualidade dos mais importantes algoritmos de binarização levando em consideração parâmetros tais como a intensidade da interferência frente-verso, a difusão da tinta no papel e a textura e escurecimento do papel pelo envelhecimento. Um novo algoritmo para a binarização eficiente de documentos com interferência frente-verso é aqui apresentado, tendo se mostrado capaz de remover tal ruído em uma grande gama de documentos. Adicionalmente, é aqui proposta a binarização “inteligente” de documentos complexos que envolvem diversos elementos gráficos (figuras, diagramas, etc).

Palavras-chave: Processamento de dados. Classificação de dados. Filtragem de imagens.

List of Figures

1.1 Historical Documents from Nabuco's bequest. 19

1.2 Binarized Document from Nabuco's bequest by Proposed Method. 20

1.3 Example of complex document encompassing text and several graphic elements. 21

2.1 Number of citations in each database for the universe of researchers studied. . 34

2.2 Number of publications in each database for the universe of researchers under study. 34

2.3 Distribution of Senior researchers by CNPq level and regions of Brazil. Source: Lattes database. 35

2.4 Number of researchers by institution in Brazilian regions. Source: Lattes database. 35

2.5 Number of doctoral and master supervisions and the total number of articles in periodicals as function of CNPq research level. Source: Lattes database. . . 36

2.6 Per capita number of completed supervisions of Ph.D. and M.Sc. and the number of articles published in periodicals as function the research of CNPq. Source: Lattes database. 36

2.7 CNPq Researcher level versus lapsed-time after PhD degree. Source: Lattes database. 37

2.8 *h*-index per capita. Source: ArnetMiner (AM), Web of Science (WS), Scopus (Sc) and Google Scholar (GS) databases. 37

2.9 Number of published articles in the different databases for the population under study. 39

2.10 The number of citations in the different databases. 40

2.11 Number of articles by the CNPq senior researchers in Computer Science by region. 41

2.12 The *h*-index in the different databases. 42

2.13 Number of citations made to the seniors researchers in Computer Science from CNPq by region. 45

2.14 h-index of the researchers by Region. 46

2.15 Discriminant analysis approach. Source: ArnetMiner database. 48

2.16 Discriminant analysis approach. Source: Microsoft Academic database. 48

2.17 Discriminant analysis approach. Source: ArnetMiner database. 50

2.18 Discriminant analysis approach. Source: Microsoft Academic database. 50

2.19 Discriminant analysis approach. Source: ArnetMiner database. 52

2.20 Discriminant analysis approach. Source: Microsoft Academic database. 52

2.21 Actual CNPq researcher level and regrouping, using discriminant analysis and k-mean methods. 54

3.1 Historical Document from Nabuco Bequest with back-to-front interference. . . 56

3.2 RGB Colour Histogram of Figure 3.1. 56

3.3 Gray-scale version of Figure 3.1. 57

3.4 Binarized document of Figure 3.1. 58

3.5 Historical Documents from Nabuco's bequest. 60

3.6 Historical Documents from Nabuco's bequest. 61

3.7 Historical Documents from Nabuco's bequest. 62

3.8 Historical Documents from Nabuco's bequest. 63

3.9 Printed Documents from the SBrT file. 63

3.10 Block Diagram of the synthetic image generator. 64

3.11 Original images without back-to-front interference. 65

3.12 Image of a synthetic document with back-to-front interference. 66

3.13 Sample texture of historical document. 66

3.14 Aging mask generated by image synthesizer. 67

3.15 Synthesized image with back-to-front interference generated by the scheme in Figure 3.10. 67

3.16 Synthetic images generated. 69

3.17 Co-occurrence matrix. 70

3.18 Procedures for the analysis of filters. 71
3.19 ISODATA filtering of the images in Figure 3.16. 74
3.20 Pun filtering of the images in Figure 3.16. 76
3.21 Kapur-Sahoo-Wong filtering of the images in Figure 3.16. 78
3.22 Johannsen-Bille filtering of the images in Figure 3.16. 80
3.23 Yen-Chang-Chang filtering of the images in Figure 3.16. 82
3.24 Otsu filtering of the images in Figure 3.16. 84
3.25 Mello-Lins filtering of the images in Figure 3.16. 86
3.26 Roe-Mello filtering of the images in Figure 3.16. 88
3.27 Silva-Lins-Rocha filtering of the images in Figure 3.16. 90
3.28 Wu-Lu filtering of the images in Figure 3.16. 92

4.1 Block Diagram of the proposed method. 98
4.2 Input and Output images of the Red channel using the proposed filter. 101
4.3 Gray-level Images Histogram Red channel from proposed filter. 102
4.4 Input and Output images of the Green channel using the proposed filter. . . . 103
4.5 Output Images Histogram Green channel from proposed filter. 105
4.6 Input and Output images of the Blue channel using the proposed filter. 106
4.7 Output Images Histogram for the Blue channel using the proposed filter. . . . 108
4.8 Result of the Proposed Filter with Historical Documents. 113
4.9 Result of the Proposed Filter with Historical Documents. 114

5.1 Example of complex document encompassing text and several graphic elements.117
5.2 Result of the direct binarization of the document presented in Figure 5.2 using Otsu global binarization algorithm. 118
5.3 LiveMemory processing scheme used in the document shown in Figure 5.1. . . 119
5.4 Binarization scheme proposed. 120
5.5 Binarization scheme proposed. 123
5.6 Compound complex block encompassing different graphical elements and its binarization using Otsu algorithm. 123
5.7 Photo of a group of people presented in the document page in Figure 5.1. . . . 125
5.8 Photo of a group of people presented in the document page in Figure 5.1. . . . 125

5.9 Five different logos and their binarization using Otsu algorithm. 126

5.10 Final Block binarized image. 128

6.1 Critical bound. 148

6.2 Hypothesis test (Dawoud & Kamel, 2004). 149

6.3 Behavior of the binary entropy function h(p). 158

List of Tables

2.1 Eigenanalysis of the correlation matrix of all databases. 43

2.2 Comparison of databases using the discriminant analysis method. 47

2.3 Regrouping in three groups using k-mean method. 49

2.4 Regrouping in four groups using k-mean method. 51

3.1 Texture information of historical document. 67

3.2 IsoData Filter Result. 75

3.3 Pun Filter Result. 77

3.4 Kapur-Sahoo-Wong Filter Result. 79

3.5 Johannsen-Bille Filter Result. 81

3.6 Yen-Chang-Chang Filter Result. 83

3.7 Otsu Filter Result. 85

3.8 Mello-Lins Filter Result. 87

3.9 Roe-Mello Filter Result. 89

3.10 Silva-Lins-Rocha Filter Result. 91

3.11 Wu-Lu Filter Result. 93

3.12 Overall algorithm classification for 1,600 synthetic images with $0 \leq \alpha \leq 1$ in steps of 0.1. 94

4.1 Output proposed filter for the Red channel. 100

4.2 Output of the proposed filter for the Green channel. 104

4.3 Output proposed filter for the Blue channel. 107

4.4 Texture information of historical document. 109

4.5 Analysis of Variance of Threshold. 109

4.6 Model Summary. 109

4.7 Result of the Simulation with Synthetic Documents using the Proposed Filter. 110

4.8 Result of the Proposed Filter with Synthetic Documents. 111

4.9 Continuation of Table 4.8. 112

5.1 The size in Kbytes of the image in Figure 1. 118

5.2 Confusion Matrix for the random forest cascaded classifier. 122

6.1 Summary of Statistical Decisions. 147

Contents

1 INTRODUCTION 18

1.1 Motivations 18

1.2 Methodology 21

1.3 The Structure of This Thesis 21

2 CONSISTENCY ANALYSIS OF A RESEARCH ASSESSMENT 24

2.1 Introduction 24

2.2 Material and Methods: Statistical Tools 29

2.2.1 The Principal Component Analysis Method 30

2.2.2 The Discriminant Analysis Method 31

2.2.3 The *k*-Mean Method 33

2.3 Results 33

2.3.1 General Statistics 34

2.3.2 The Principal Component Analysis Method 42

2.3.3 The Discriminant Analysis Method 46

2.3.4 The *k*-Mean Method 48

2.4 Conclusions 53

3 ASSESSING BINARIZATION TECHNIQUES FOR DOCUMENT IMAGES WITH BACK-TO-FRONT INTERFERENCE 55

3.1 Introduction 55

3.2 Generation and Filtering of Images with Synthetic Degradation 58

3.2.1 Removing the Back-to-Front Interference 68

3.2.2 Co-Occurrence Matrices as a Measure of Quality 70

3.3 Thresholding Algorithms 72

3.3.1 The IsoData Method 73

3.3.2 Pun Method 75
3.3.3 Kapur-Sahoo-Wong Filter 77
3.3.4 Johannsen-Bille Method 79
3.3.5 Yen-Chang-Chang Method 81
3.3.6 Otsu Threshold Method 83
3.3.7 Mello-Lins Algorithm 85
3.3.8 Roe-Mello Algorithm 87
3.3.9 Silva-Lins-Rocha Algorithm 89
3.3.10 Wu-Lu Algorithm 91
3.4 Classification of Binarized Images of the Filters Studied 93
3.5 Conclusions 95
4 A NEW BINARIZATION ALGORITHM FOR IMAGES WITH BACK-TO-FRONT INTERFERENCE 97
4.1 The New Algorithm 97
4.2 The Bilateral Filter 98
4.3 The Decision Making Block 98
4.4 Image Classification 99
4.5 Results of the Proposed Method 100
4.6 Threshold Calculated using Regression Model 108
4.7 Classification of Binarized Images Including the Proposed Filter 110
4.8 Results of the Proposed Method 113
4.9 Conclusions 114
5 BINARIZING COMPLEX SCANNED DOCUMENTS 116
5.1 Introduction 116
5.2 The New Binarization Scheme 119
5.3 The Block Classifier 121
5.4 Binarizing Photos 124
5.5 Binarizing Logos 125
5.6 Binarizing Documents 126
5.7 Binarizing Pie Diagrams and Histograms 126
5.8 Conclusions 127

6 CONCLUSIONS AND LINES FOR FURTHER WORK 129

REFERENCES 133

APPENDIX A 139

1 INTRODUCTION

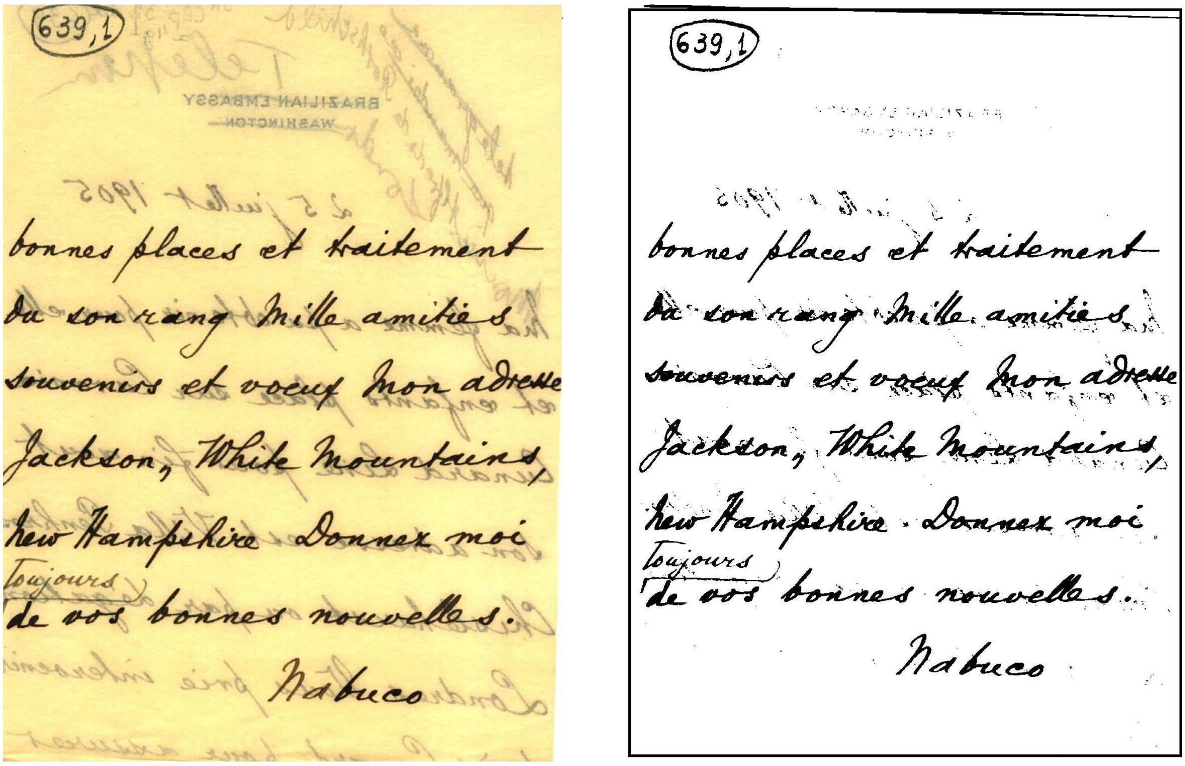
This Chapter presents the global motivation for the work developed in this thesis that shows applications of statistical analysis to different areas of knowledge. The structure of the document is also outlined.

1.1 Motivations

Data analysis is a process of inspecting, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. The data can be worked, manipulated, interpreted, and transformed into meaningful conclusions drawn from empirical research studies. This thesis, presents two different applications of statistical analysis.

The motivation for the first application was to investigate and verify whether the scientific production of senior researchers in computer science that hold scholarships from the Brazilian research council CNPq is consistent with their classification. In addition, it aims to verify if there is any kind of institutional or regional bias in it. The researchers were re-assessed using data structure analysis and clustering methods such as the *principal component analysis*, the *discriminant analysis*, and the *k-means* having as input data from ArnetMiner (ArnetMiner, 2016), Google Scholar (Scholar, 2016), Microsoft Academic (Academic, 2016), Scopus (Scopus, 2016), the Web of Science (of Science, 2016), and data in the CNPq Lattes (Lattes, 2016) Curriculum Vitae of each of the researchers. The different re-classifications were compared with the one by CNPq and the mismatches and incongruences analysed. Minitab 17 was used as the statistical tool for the tests performed.

The second application of statistical analysis made here is assessing the quality of the algorithms used for binarizing documents with back to front interference, a phenomenon that appears whenever a document is handwritten or typed on both sides of translucent paper. Figure 1.1 presents an example of one those letters from Nabuco bequest that was binarized by Otsu's Algorithm.



639,1

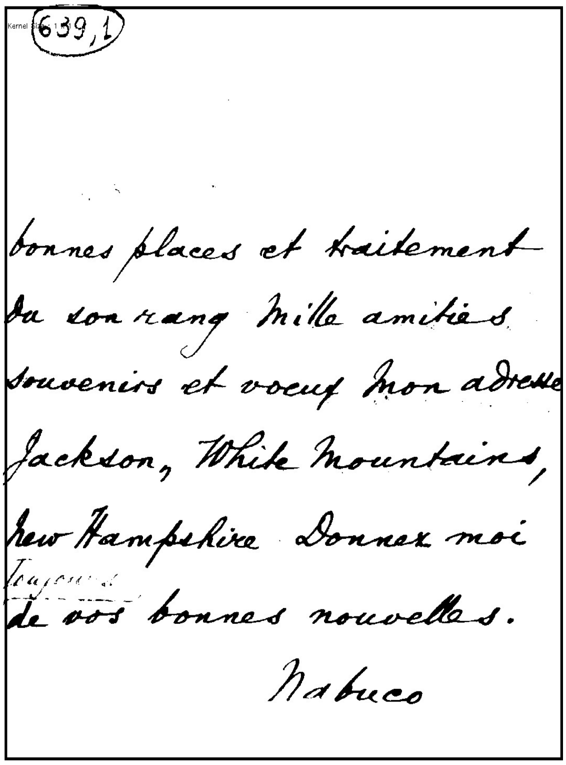
bonnes places et traitement
de son rang. Mille amitiés,
souvenirs et voeux. Mon adresse
Jackson, White Mountains,
New Hampshire. Donnez moi
toujours
de vos bonnes nouvelles.
Nabuco

(a) Historical Document. (b) Historical Document Binarized by Otsu's Algorithm.

Figure 1.1: Historical Documents from Nabuco's bequest.

Such a noise is often found in historical documents, in which the paper ages and becomes darker, increasing the degree of difficulty in the noise filtering process. One of the important contributions of this thesis is to show that no binarization algorithm is able to perform such a filtering efficiently to all kinds of documents as the degree of the noise intensity, the hue of the paper background and the opacity of the paper vary drastically from document to document.

The experience gained with the assessment of the algorithms to binarize documents with back-to-front interference provided the motivation in attempting to develop a new algorithm for such a purpose using a *bilateral filter*, an image filtering technique that smoothens the image while preserving its edges. Statistical techniques such as linear prediction were applied in the automatic calculation of the threshold value of the filter. Figure 1.2 shows the binarized image from Nabuco's bequest by proposed method development in the thesis.



Kernel Size : 1 x 1
639,1

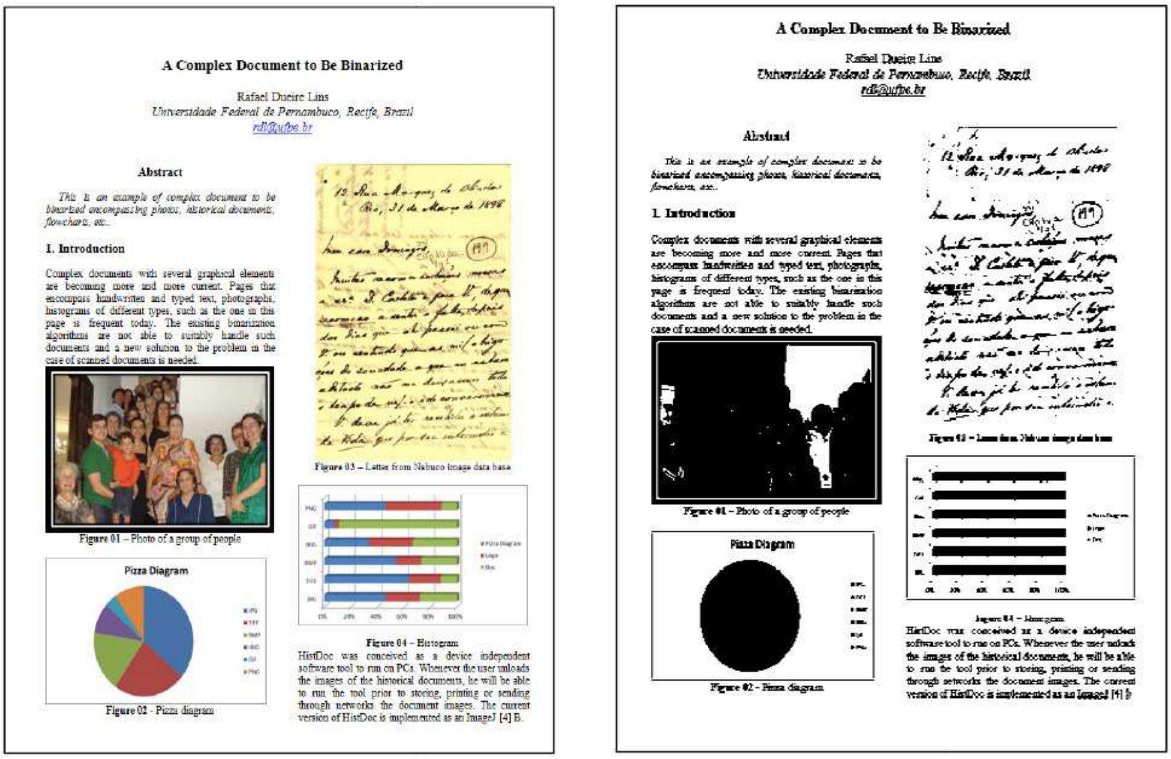
bonnes places et traitement
de son rang. Mille amitiés
souvenirs et voeux. Mon adresse
Jackson, White Mountains,
New Hampshire. Donnez moi
toujours
de vos bonnes nouvelles.
Nabuco

(a) Binarized Historical Document.

Figure 1.2: Binarized Document from Nabuco's bequest by Proposed Method.

The widespread use of document and image editing tools today has drastically changed the complexity of documents, which often encompass not only text but also all sorts of graphical elements such as photos, histograms, pie (or pizza) diagrams, etc, according Figure 1.3a. The direct binarization of such pages yielded the complete loss of the information of the graphical elements, in a similar situation to the one shown in Figure 1.3b. No global binarization algorithm can be capable of being a *all-case-winner*.

This thesis proposes a new scheme for the binarization of complex documents that decomposes the image in blocks, classifies each of them, and introduces the concept of *semantic binarization*. The idea here is to preserve image information. For instance, the binarization of a pie diagram would cause a complete loss of the original information as the color information is lost. Colors are replaced by different textures to preserve the original information.



A Complex Document to Be Binarized

Rafael Dueire Lins
Universidade Federal de Pernambuco, Recife, Brazil
rdl@ufpe.br

Abstract

This is an example of complex document to be binarized encompassing photos, historical documents, flowcharts, etc..

1. Introduction

Complex documents with several graphical elements are becoming more and more current. Pages that encompass handwritten and typed text, photographs, histograms of different types, such as the one in this page is frequent today. The existing binarization algorithms are not able to suitably handle such documents and a new solution to the problem in the case of scanned documents is needed.

Figure 01 – Photo of a group of people

Figure 02 - Pizza diagram

Figure 03 – Letter from Nabuco image data base

Figure 04 – Histogram

HistDoc was conceived as a device independent software tool to run on PCs. Whenever the user unloads the images of the historical documents, he will be able to run the tool prior to storing, printing or sending through networks the document images. The current version of HistDoc is implemented as an ImageJ [4] B.

(a) Example of complex document encompassing text and several graphic elements.

(b) Result of the direct binarization of the document presented in Figure 1.3a using Otsu global binarization algorithm.

Figure 1.3: Example of complex document encompassing text and several graphic elements.

1.2 Methodology

The two applications developed on in this thesis have one thing in common: both can be represented by a vector of characteristics that qualifies the object to be studied. In the first application, the characteristic vector represents the level of scientific production of each researcher, formed by a set of parameters defined by the different databases. In the second application, the characteristic vector stands for the texture of the historical document. Tools for multivariate statistical analysis were used in this thesis for both applications.

1.3 The Structure of This Thesis

This Thesis has six chapters including this introduction, which presents the motivation for the work developed.

Chapter 2 makes use of cluster analysis techniques to analyze the consistency of the classification of senior researchers in Computer Science of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), the Brazilian Research Council. The use of multivariate statistical techniques applied to data from international public databases were compared with the CNPq classification. General statistics were presented as: the number of researchers by region, by institution, by level of classification, by scientific production, etc.

Chapter 3 assesses the quality of the images generated by several binarization algorithms found in the technical literature. The images analyzed in this study belong to the bequest of documents of Joaquim Nabuco, held by the Joaquim Nabuco Foundation, a social science research centre in Recife, Brazil, and images from the SBrT (Sociedade Brasileira de Telecomunicações) dataset developed in the LiveMemory Project (Lins, 2010a).

Chapter 4 presents a new algorithm for filtering out the back-to-front interference in documents. The new filtering proposed uses a global technique performed in four steps:

1. Filtering the image using the Bilateral filter.
2. Splitting image in the RGB components.
3. Decision-making block for each RGB channel based an adaptive binarization method inspired by Otsu's method with a choice of the threshold level, and.
4. Classification of the binarized images to decide which of the RGB components best preserved the text information in the foreground.

The quality of the binary images was assessed *quantitative* and *qualitatively* both in chapters 3 and 4 of this thesis. The quantitative analysis was made by calculating matrices of the co-occurence probability of the pixels. Visual inspection was used in the qualitative analysis. Both assessment methods have led to similar conclusions in terms of the efficiency of filtering techniques.

Chapter 5 presents the analysis of the complex documents encompassing several graphic elements such as photos, histograms, pie (or pizza) diagrams besides text. The scheme automatically decomposes a document image and identifies each of the graphical elements to provide a suitable binarization for each of them independently. The main principle of the scheme proposed here was the recognition that each of the blocks in the complex document has a different nature and no binarization algorithm that does not take such information into account has the slightest chance of succeeding in keeping the fundamental information of the original document. Most binarization algorithms are suitable for scanned text documents and do not work adequately with complex documents that encompass various graphic elements simultaneously.

This thesis presents a new binarization algorithm that works on complex documents. Each of the elements in the image is processed depending on its nature, thus their binarization takes that into account to preserve the original content. The goal is to investigate algorithms that can classify such graphic elements and to find a new "intelligent" dithering algorithm that takes into account the content of the photo, logos, pie diagrams to be binarized.

Chapter 6 draws the conclusions a presents lines for further work along the lines of this thesis.

Appendix A shows the basis of construction of the various algorithms evaluated in this thesis.

2 CONSISTENCY ANALYSIS OF A RESEARCH ASSESSMENT

This Chapter analyzes the consistency of the assessment for senior researchers in Computer Science of The Brazilian Research Council CNPq. Information from the public databases ArnetMiner (ArnetMiner, 2013), Web of Science (of Science, 2013), Scopus (Scopus, 2013), Lattes (Lattes, 2013), Google Scholar (Scholar, 2013) and Microsoft Academic (Academic, 2013) were compared with the CNPq classification. Several statistical classification strategies were considered. Regional and institutional information were also taken into account.

2.1 Introduction

Assessing research is a challenging task. A number of measures have been developed aiming at assessing the scientific production of researchers and their impact. The number of citations a given paper or author has is an evidence of its/his importance. The h-index, introduced by Hirsch (2005), is defined as the number of papers with citation number higher or equal to h, while the rest of the N papers have less than h citations each. The g-index was introduced by Leo Egghe in 2006 to refine the h-index. According to Egghe (2006, 2007), the g-index gives more weight to the highly-cited papers. Given a number of papers ranked in a decreasing order according to the citations received, the g-index is the largest number such that the top g articles received (altogether) at least g^2 (g square) citations. A comparative discussion between the g and h indices may be found in reference Costas (2008). Many people have raised questions about the validity and consistency of research assessments and such indices. A paper or an author may have high citation indices for having been proved wrong or for being controversial. The *sociability* of the researcher and of the prestige of his institution of affiliation are some of the factors that influence the indices of a researcher. Labbé (2010) shows in his article that there may be ways of manipulating the h-index and articles, showing the Scigen that is an automatic generator of amazing articles using the jargon of computer science. Scigen is based on hand-written context-free grammar and has been developed by the PDOS research group at MIT CSAIL.

Although error-prone, research assessments need to be made. All funding agencies struggle to make *qualitative* unbiased evaluations, but the difficulty of such process always drops down to a *quantitative* one. According to Glenn Hampson the executive director of the National Science Communication Institute and the Open Scholarship Initiative program in the United States. Scientific journals have undergone profound changes around the world in recent years in order to increase agility in the publishing process - still considered very slow - and to promote open access, which has not yet advanced at the desired speed, among other objectives, Hampson (2017). Hampson also says that magazine publishing is part of the large, complex ecosystem of scientific communication, which is poorly defined and has evolved through a variety of disconnected efforts and initiatives. Very easily one finds people who wrote their names in the history of science and that produced only very few works of fundamental importance. Unfortunately, that is not the usual case, and even more so since the last decades of the 20^th century in which sciences advance in very small steps pulled by hundreds of researchers in each very specialized areas of knowledge.

The whole point in checking the consistency of a research assessment is to try to verify the degree of fairness of the whole process and to check if there is some sort of intentional or unintentional bias in its results. Assessing research is even more difficult in developing countries such as Brazil, in which the Federal annual budget for funding research in all areas is much smaller than qualified demand. As a result, there is a fierce competition for research funds. The CNPq, actively participates in the formulation, implementation, monitoring, evaluation and dissemination of the National Policy on Science and Technology in Brazil. Among such activities, CNPq provides personal grants to the researchers who excel in their research activities. Although the value paid by such a grant ranges today between US$ 250.00 to US$ 300.00 per month, that grant is seen as a recognition of the quality of the research work of the beholder, being thus a symbol of status and prestige. Researchers holding a CNPq grant also qualify for special research funding calls from CNPq. Postgraduate programs and undergraduate courses are also assessed based on the number of staff holding CNPq-research grants. There are approximately 6,000 Doctors in Computer Science in the Brazil. The CNPq research grants today are classified in levels $1A$, $1B$, $1C$, $1D$, and 2, in decreasing order. The research grants have a duration of 3 or 4 years and there are about 350 research grants today. There are two grant award meetings of the board per year. The most senior group is the target of this work. There were 78 research grants in 1A, 1B and 1C levels in 2014.

As already mentioned, a qualitative research evaluation is an unfruitful task. The history of mankind has shown that several times the geniuses in art and science were far ahead of their time and were only recognized much later on, sometimes posthumously. Such exceptional situation is completely out of the scope of any attempt to measure scientific production. This work focuses on checking if the scientific production of the senior researchers in computer science holding CNPq grants is compatible with their classification. Besides that, it aims to verify if there is some sort of institutional or regional bias in it. It is important to mention that the evaluation process is made by a board (*Comitê Assessor*) of 6 to 8 senior grant holders that follow an indication of the whole community of senior grant holders. The scientific director of CNPq invites those indicated researchers to join the assessment committee for 1, 2, or 4 years attempting to bring in some regional and institutional representativity. No member of such a board may be reconducted immediately after serving on it, guaranteeing a rotativity of the researchers on the board.

This work uses distinct statistical methods to cluster the data found in different research assessment public databases such as ArnetMiner (ArnetMiner, 2016), Web of Science (of Science, 2016), Scopus (Scopus, 2016), Google Scholar (Scholar, 2016), and Microsoft Academic (Academic, 2016). The universe of senior researchers studied (levels 1A, 1B, and 1C of CNPq) encompasses 78 researchers. Besides those public databases, this assessment also took into consideration the CV-Lattes of those researchers, which is a public *curriculum vitae* used by CNPq and all Brazilian research and postgraduate programs. The data in the CV-Lattes is informed by the researcher, reporting different activities (education, affiliations, articles published in journals and conferences, grants, teaching activities, etc.) following standard fields. The researcher states the veracity of the information declared in his CV-Lattes and may suffer legal penalties in case of false declarations. The CV-Lattes platform automatically checks the DOI of publications (if available), cross-checks information of supervision activities between the CVs of the advisor and student, demands confirmation of institutions that provide grants, etc. The result of the different clustering strategies using the different databases is compared with the classification of the CNPq grants to verify the degree of agreement of the two methods.

The data from ArnetMiner were collected in the period between 11/18/2013 and 11/19/2013; from the Web of Science, Scopus, and Lattes between 11/20/2013 and 11/27/13; from Google Scholar on 12/02/2013; and from Microsoft Academic in the period between 12/03/2013 and 12/06/2013. Not all the 78 senior researchers were found in all databases. The number of CNPq senior researchers found in each database is as follows: ArnetMiner (75 researchers), Web of Science (78 researchers), Scopus (78 researchers), Lattes (78 researchers), Google Scholar (27 researchers) and Microsoft Academic (58 researchers).

The following data was collected for each of the studied researchers from the different databases for later classification/clustering, if available:

- # citation: The total number of citations to publications by a researcher.
- # publication: - The total number of publications by a researcher.
- *h*-index: A researcher has index h if h of his N papers have at least h citations each, and the other $(N - h)$ papers have at most h citations each.
- *g*-index: a variant of the h-index, which takes into account the citation evolution of the most cited papers over time.
- activity: the total number of papers published in the last n years or during the whole carreer.
- diversity: the number of different research fields a researcher has publications in. The author-conference-topic model to obtain the research fields for each expert.
- sociability: The score of sociability is basically defined based on how many coauthors an expert has.

The following indices are present and were used for classification in each of the databases:

- ArnetMiner (AM): activity, number of citations, *h*-index, papers, *g*-index, sociability and diversity.

- Web of Science (WS): number of articles, # citations, *h*-index and medium.
- Scopus (Sc): total number of articles, citations, mean and *h*-index.
- Lattes: articles in scientific journals, number of papers published in proceedings, books authored, book chapters, number of patents, and number of successful Ph.D. supervisions.
- Google Scholar (GS): number of citations, *h*-index, *i10*-index, citations(2008), *h*-index(2008) and *i10*-index(2008).
- Microsoft Academic (MA): number of publications, number of citations, cited by and number of co-authors.

It is also important to observe that besides not all source databases have different coverage of the universe studied of 78 senior researchers, their data also vary as, for instance, the number of papers or citations by a given researcher may vary from database to database, and from them to the Lattes CV, which is informed by the researcher. An analysis of the possibility of existing a bias in grant distribution between the regions is also provided here.

Brazil is geopolitically divided into five regions (also called macroregions) by the Brazilian Institute of Statistics and Geography (IBGE); each region is composed of three or more states. Officially recognized, the division in such regions is not merely an academic and geopolitical one; social and economic factors, including funding are also region-based. The five Regions of Brazil with and their states are:

1. North (N): Acre, Amapá, Amazonas, Pará, Rondonia, Roraima, Tocantins.
2. Northeast (NE): Alagoas, Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, Sergipe.
3. MidWest (CO): Goiás, Mato Grosso, Mato Grosso do Sul, Distrito Federal (Federal District).
4. Southeast (SE): Espírito Santo, Minas Gerais, Rio de Janeiro, São Paulo.
5. South (S): Paraná, Rio Grande do Sul, Santa Catarina.

The Southeast and South regions are the most densely populated and economically developed in Brazil.

To the best of the knowledge of the author of this thesis, no work reports on the analysis of consistency of the assessment of researchers with CNPq grants. The only effort that attempts to do so is the paper by Barata (2003) which analyzes the profile of the researchers' in public health with CNPq personal research grants. The analysis considered the researchers undergraduate and graduate degrees, field of expertise, scientific output, and publications. The validity of the analysis presented here goes far beyond the analysis of the fairness of the distribution of personal research grants in Brazil. It also shows how difficult it is to make a fair and unbiased comparative research assessment.

2.2 Material and Methods: Statistical Tools

A problem that arises quite often in many research areas is that, given a set of n individuals, grouping them into k homogeneous classes or subsets, heterogeneous with each other (i.e., individuals of different subgroups are dissimilar).

There are two kinds of classification procedures: supervised and unsupervised classification. The supervised classification is the essential tool used for extracting quantitative information from object data. Using this quantitative information, the analyst has enough data available to generate representative parameters for each class of interest. This step is called training. Once trained, the classifier is then used to attach labels to all the object data according to the trained parameters. According to Richards (1993), the Maximum Likehood Classification is the most widely used kind of supervised classification used. Its effectiveness depends on how reasonably accurate is the estimation of the mean vector m and the covariance matrix for each spectral class data.

On the other hand, unsupervised classification does not require *a priori*, knowledge of the classes, making use of some clustering algorithm to classify the data. These procedures can be used to determine the number and location of the unimodal spectral classes. One of the most commonly used unsupervised classification method is the *migrating means clustering*. Richards (1993) describes this method as initially assigning each object one of the n-clusters which have as central objects n-randomly chosen ones. Then, the distances of all objects to the centers of all the clusters is calculated. Objects are moved to minimize the sum of the square error. Such a procedure is iteratively repeated until stability is reached.

In attempting to choose an appropriate analytical technique, to classify the CNPq Senior researchers, three methods were used: the principal component analysis approach, (Anderson, 1984), the Discriminant Analysis, (Hair, 2009), that is used to determine which variables are the best predictors and the k-means method, (Hair, 2009).

2.2.1 The Principal Component Analysis Method

The Principal component analysis is appropriate when one has obtained measures on a number of observed variables and wishes to develop a smaller number of artificial variables (called *principal components*) that will account for most of the variance in the observed variables. The principal components may then be used as a predictor or criterion variables in the subsequent analyses.

A Principal component analysis focuses in explaining the variance-covariance structure of a set of variables through a few linear combinations of such variables. Its general objectives are: data reduction and interpretation. A Principal component can be defined as a linear combination of optimally-weighted observed variables.

The following procedure will perform a Principal Components Analysis on a set of data.

- Step 1: Get some data set; $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$.

- Step 2: Subtract the mean; The mean subtracted is the average across each dimension. This produces a data set whose mean is zero.

- Step 3: Calculate the covariance matrix; The formula for covariance is always measured between 2 dimensions. The formula for covariance is:

$$cov(X,Y) = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \tag{2.2.1}$$

- Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix; One should recall that the covariance is always measured between 2 dimensions. If one has a data set with more than 2 dimensions, there is more than one covariance measure that can be calculated.

The covariance matrix for a 2−dimensional data set, has 2 rows and 2 columns:

$$\begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix} \tag{2.2.2}$$

- Step 5: Choosing components and forming a feature vector.

In general, once the eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, the highest to the lowest. This gives the components in order of significance. Now, one can decide to ignore the components of lesser significance.

If one leaves out some components, the final data set will have less dimensions than the original one. More precisely, if one originally has n dimensions in the data, one calculates n eigenvectors and eigenvalues, and then one chooses only the first p eigenvectors, then the final data set has only p dimensions.

Algebraically, the principal components are linear combinations of the p random variables $X_1, X_2, \dots, X_p$. Geometrically, those linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \dots, X_p$ as the coordinate axes, according to Hair (2009).

2.2.2 The Discriminant Analysis Method

Discriminant analysis is a method that can be applied for data classification and selection. The methods of discriminant analysis are based on the assumption that a subdivision of the data is available, and the aim is to seek directions in space showing the separation of such subgroups or to determine a rule for future classifications. But according to Crammer (2003), often there is no classification of its type available, and the problem is to identify which (and how many) are the different classes of existing individuals in the data set available.

In the following, the linear discriminant analysis method is discussed, in which the classification problem can be solved by finding the linear functions that best divide the groups into clusters.

The aim of the Discriminant Analysis is to find an equation that will help to predict the value of a dependent variable based on the values of a set of independent variables. The dependent variable is qualitative. The dependent variable must be nonmetric, representing groups of researches that are expected to differ on the independent variables. The dependent variables are chosen as the most representative variables of the groups of interest.

The choice of a dependent variable is such of best representing the groups of interest.

Discriminant analysis is a parametric technique to determine which weights for the quantitative variables or predictors best discriminate between two or more groups of cases and do so better than chance, according to Crammer (2003). The analysis creates a discriminant function which is a linear combination of the weights and scores on such variables. The maximum number of functions is either the number of predictors or the number of groups minus one, whichever of these two values is the smallest.

$$Z_{jk} = a + w_1 X_{1k} + w_2 X_{2k} + \cdots + w_n X_{nk}, \tag{2.2.3}$$

where:

$Z_{j,k}$ = discriminant Z score of discriminant function j for object k. a = intercept. w_i = discriminant weight for independent variable i. X_{ik} = independent variable i for object k.

Whenever there are several groups, the discriminant functions offer a separation threshold, but it is unable to predict the number of members in each of the classes. Thus, in addition to improving the explanation of group membership, these additional discriminant functions add insight into the various combinations of independent variables that discriminate between groups. With three categories of the dependent variable, discriminant analysis can estimate two discriminant functions, each representing a different dimension of discrimination. Thus, one can now calculate two discriminant scores for each respondent.

Discriminant analysis can be used to classify observations into two or more groups if one has a sample with known groups. Discriminant analysis can also be used to investigate how variables contribute to group separation.

2.2.3 The k-Mean Method

The aim of the k-means algorithm is to divide m points in n dimensions into k clusters such that the sum of the squares is minimized within clusters.

The k-mean is the most popular partitioning method of clustering. It was firstly proposed by MacQueen(1967) (apud Johnson (2007)). It is an unsupervised, non-deterministic, numerical, iterative method of clustering. In k-mean each cluster is represented by the mean value of objects in the cluster. Here, a set of n objects is divided into k clusters in such a way that the intercluster similarity is low and intracluster similarity is high. The degree of similarity is measured in terms of the mean value of objects in a cluster, according to Yadav (2013).

The k-mean procedure assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of the following steps:

1. Partition the items into k initial clusters.

2. Proceed through the list of items, assigning an item to the cluster whose centroid is the nearest (the distance is usually computed using the Euclidean distance). Recalculate the centroid for the cluster that receives the new item and for the cluster that loses it.

3. Repeat step 2 until no more reassignments take place.

The k-means clustering algorithm works best when sufficient information is available to make a good starting cluster assignments. The final assignment of the items to a cluster will be, to some extent, dependent upon the initial partition or the selection of the points, which are taken as *seeds*.

2.3 Results

The following results were obtained from the databases using the statistical methods of analysis outlined. All the data either mentioned explicitly or not refer to the specific set of senior researchers of CNPq studied.

2.3.1 General Statistics

The representativeness of the databases analysed is a fundamental starting point. The volume of data (published articles of all researchers analysed) and the number of citations to those articles. The databases with the largest volume of cumulative citations for the population studied is ArnetMiner, as shown in Figure 2.1, the one with the smallest volume is the Web of Science database. The largest volume of cumulative publications is found in the Lattes database, which is informed by the researchers themselves. The second largest is in the ArnetMiner database. The Web of Science is the smallest database in terms of the volume of accumulated citations for the researchers studied, as shown in Figure 2.2.

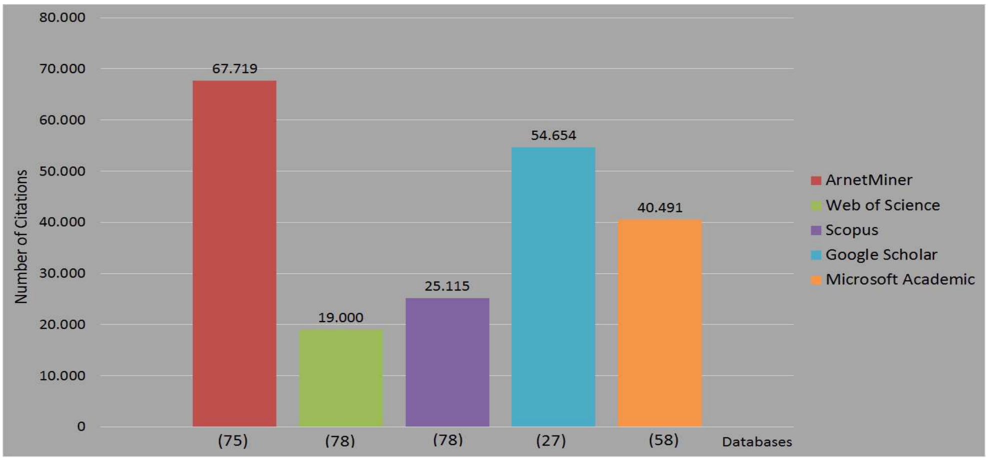


Figure 2.1: Number of citations in each database for the universe of researchers studied.

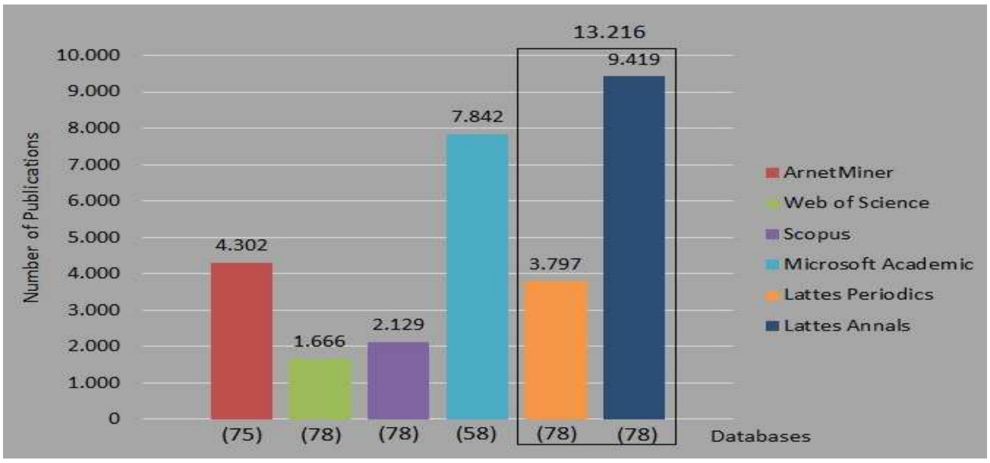


Figure 2.2: Number of publications in each database for the universe of researchers under study.

Figure 2.3 shows the distribution of the Senior researchers in Computer Science Brazil by the rank of CNPq and by region at the snapshot taken in data, in which one may observe that there is a strong concentration in the Southeast, the most economically developed region in Brazil. The Northeast region is represented by the Federal University of Pernambuco (UFPE), while the South region by Federal University of Rio Grande do Sul (UFRGS). The distribution of researchers by institution is shown in Figure 2.4. The largest number of top researchers (1A) at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) with eight researchers, followed by the University of São Paulo (USP) and the Federal University of Rio de Janeiro (UFRJ), with seven researchers 1A-level each.

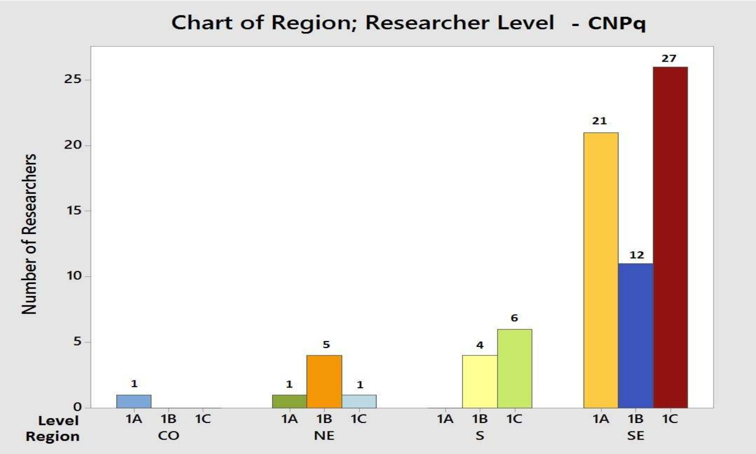


Figure 2.3: Distribution of Senior researchers by CNPq level and regions of Brazil.
Source: Lattes database.

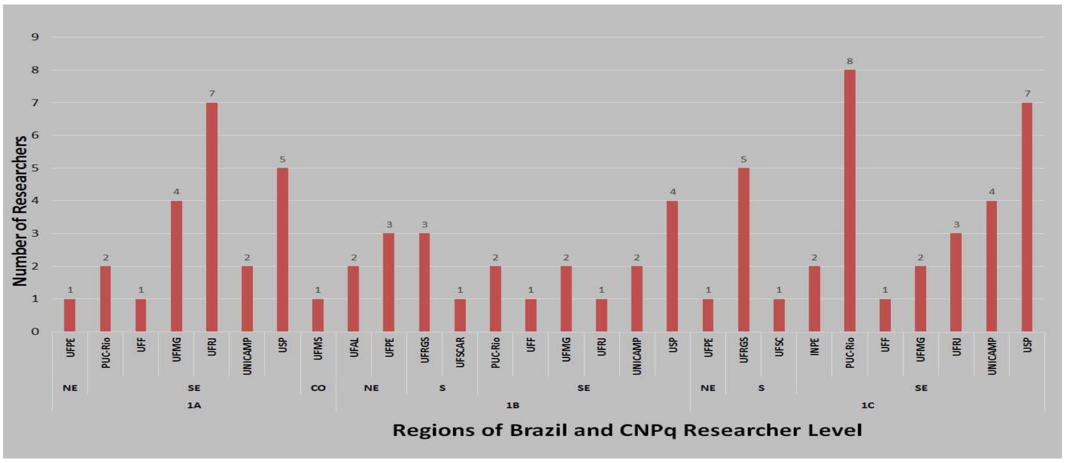


Figure 2.4: Number of researchers by institution in Brazilian regions.
Source: Lattes database.

Figure 2.5 shows the scientific production of the researchers studied by the classification levels of CNPq, the total articles published in journals, the number of doctoral supervisions completed, as well as master dissertations directed. Surprisingly, under such aspects only, the production volume of the researchers classified by CNPq as $1C$ level is higher than that for those classified at the $1B$-level. This may be a sign of some level of inconsistency in the researcher classification. The per capita production of researchers in the category $1C$ is equivalent to the production of researchers at level $1B$, as shown in Figure 2.6.

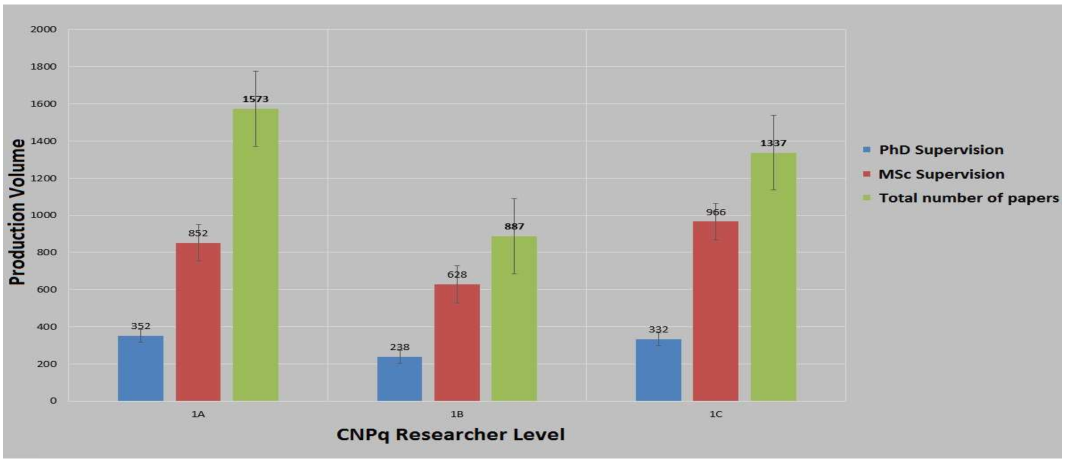


Figure 2.5: Number of doctoral and master supervisions and the total number of articles in periodicals as function of CNPq research level. Source: Lattes database.

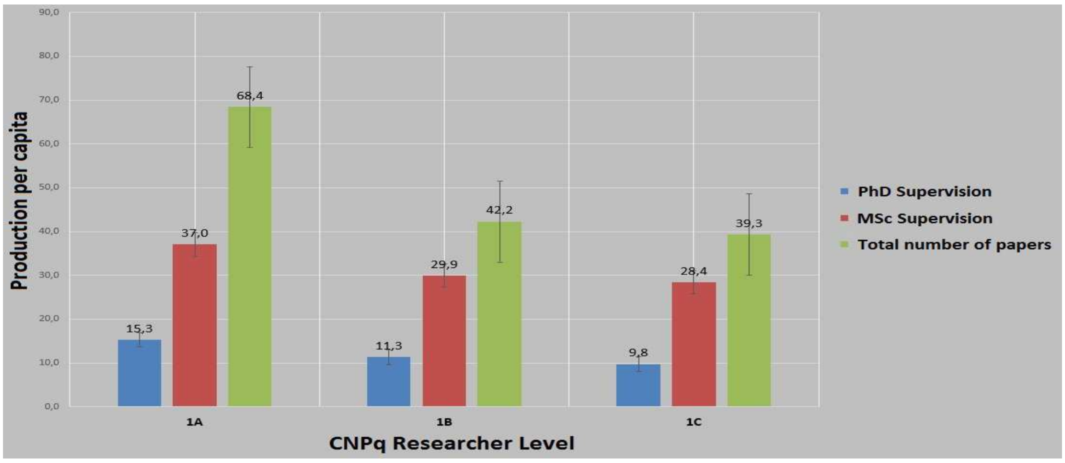


Figure 2.6: Per capita number of completed supervisions of Ph.D. and M.Sc. and the number of articles published in periodicals as function the research of CNPq. Source: Lattes database.

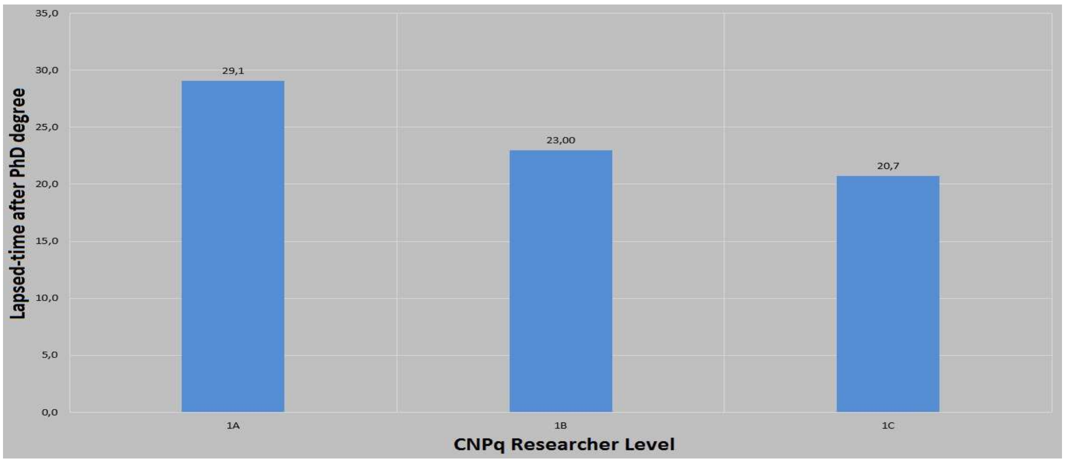


Figure 2.7: CNPq Researcher level versus lapsed-time after PhD degree. Source: Lattes database.

The time since the completion of the Ph.D. degree in the group of researchers $1C$ is also close to in the group of researchers $1B$, as shown in Figure 2.7. Figure 2.8 shows the h-index of the researchers studied. The databases were considered: ArnetMiner, Scopus and Web of Science. One way observe that the h-index of researchers who are at the level $1C$ and $1B$ as equivalent to the Web of Science database.

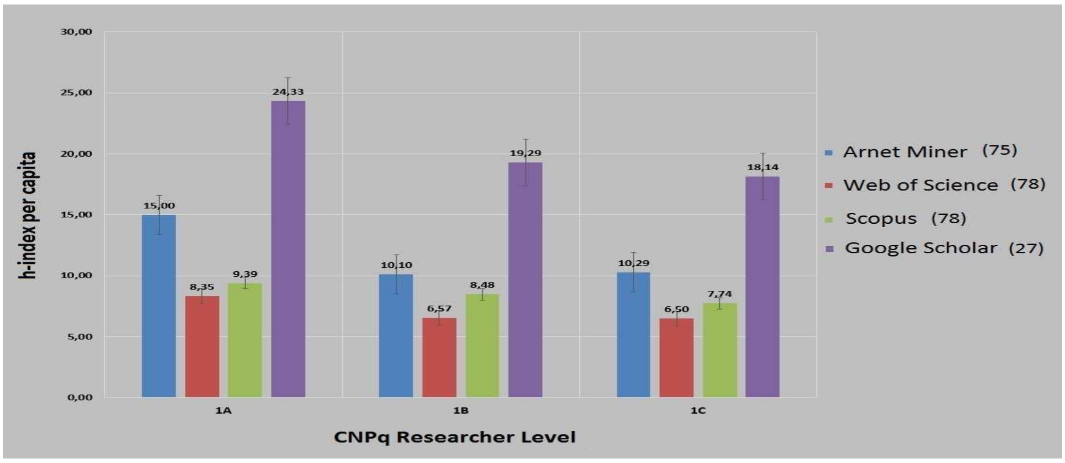


Figure 2.8: h-index per capita. Source: ArnetMiner (AM), Web of Science (WS), Scopus (Sc) and Google Scholar (GS) databases.

A more detailed statistical analysis follows:

1. The number of articles published by the researchers studied in the databases shows that the mean values differ from database to database. Therefore, the scales of the plots are not the same. In the following, several databases will be considered.

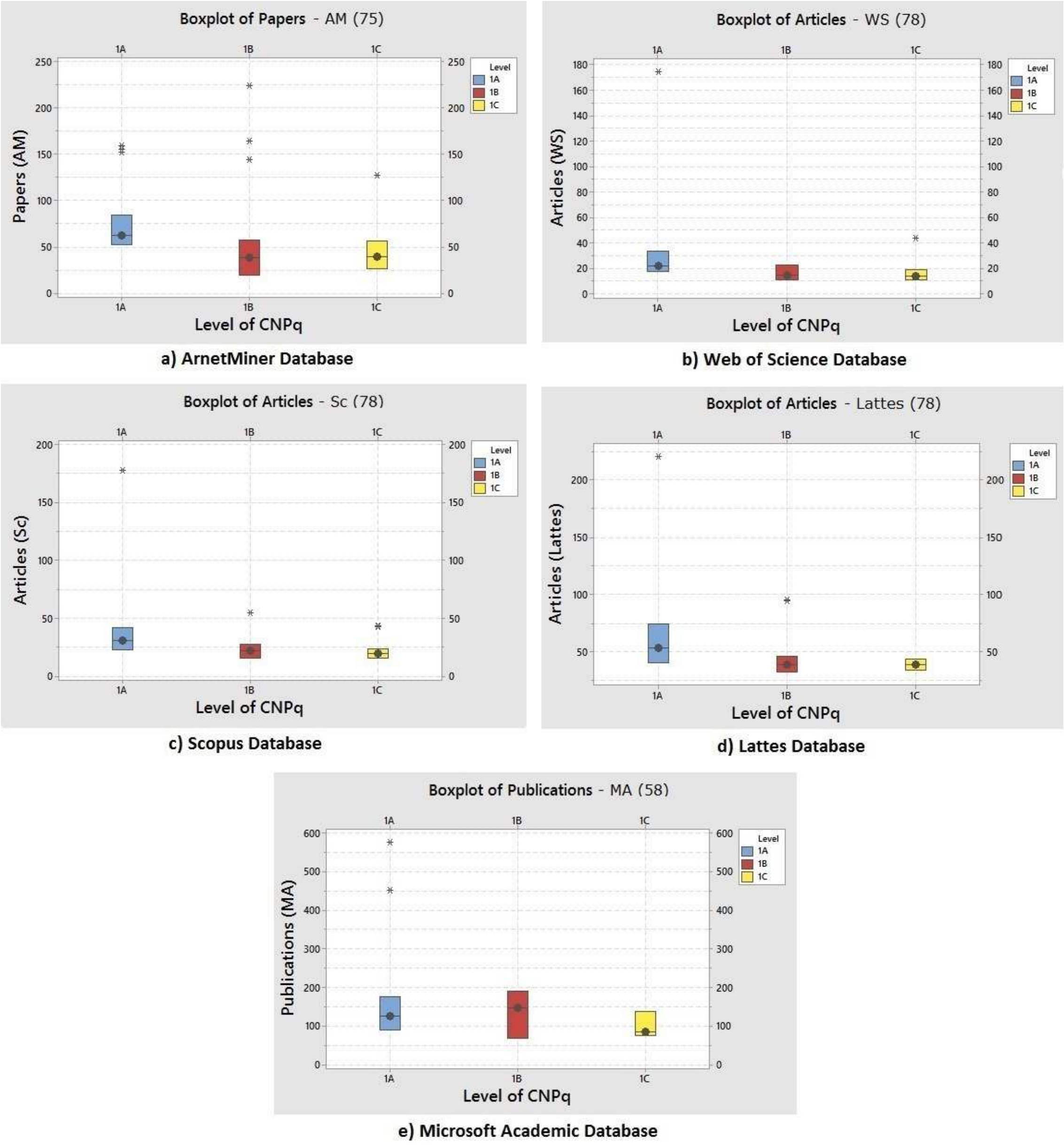
2. A normality test applied to the databases indicates that the data do not follow a normal distribution, p-value > 0.05.

3. Mood Median test was used to analyze and provided evidence that the means of the groups can be considered equal to the 5% level of significance ($p < 0.05$, rejecting the null hypothesis H_0).

4. The data from the databases Scopus and Microsoft Academic show that the samples come from populations with equal medians.

5. The data from the ArnetMiner and Web of Science databases come from populations with medians that are not equal as can be observed in Figure 2.9.

6. The number of published articles by researcher in all the studied databases have a mean and variance within the same range of values, in absolute terms, except for the outliers, as one can also see in Figure 2.9.



a) ArnetMiner Database

b) Web of Science Database

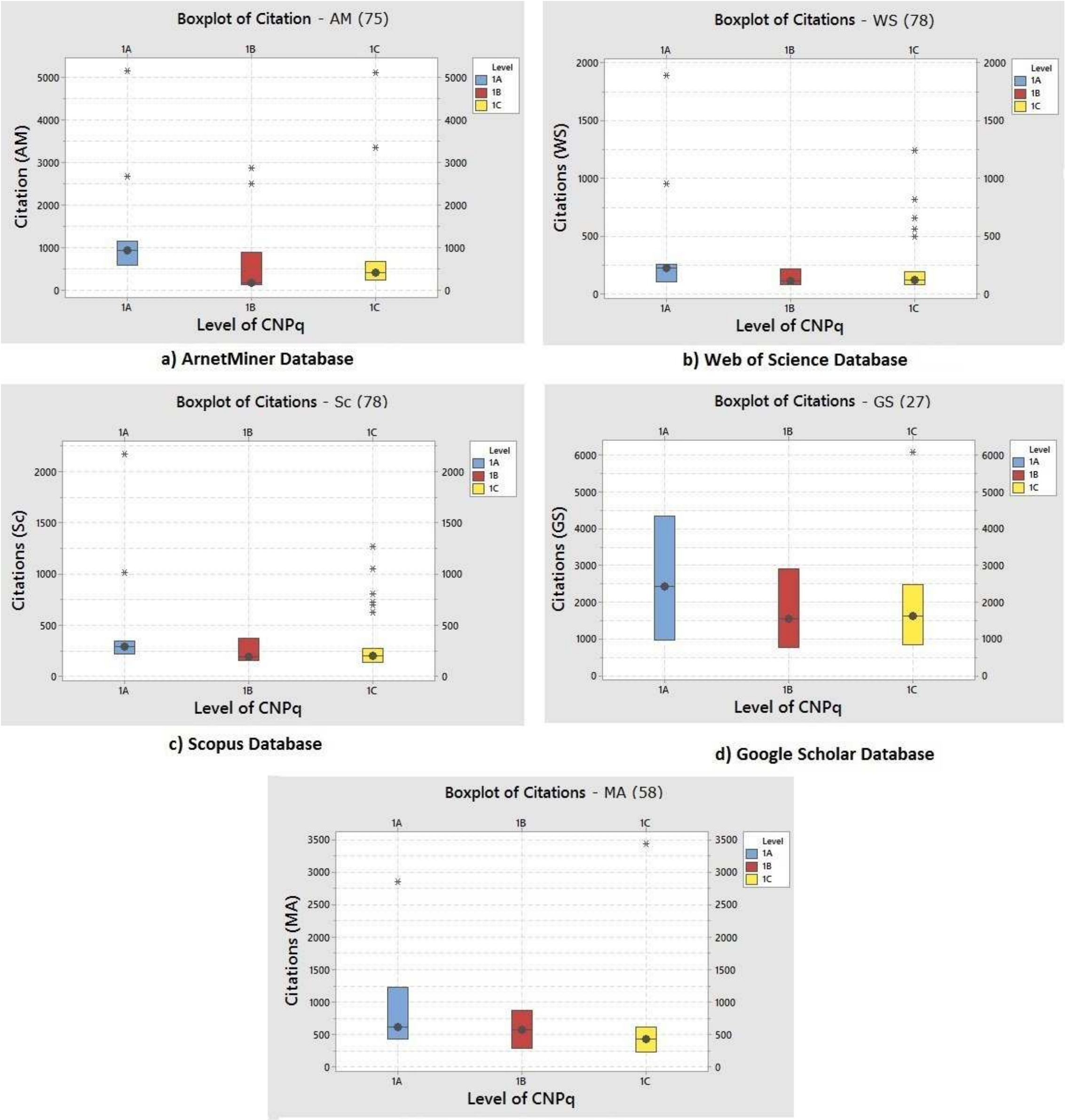
c) Scopus Database

d) Lattes Database

e) Microsoft Academic Database

Figure 2.9: Number of published articles in the different databases for the population under study.

With respect to the number of citations of articles, the data from the Web of Science, Scopus, Microsoft Academic and Google Scholar databases have samples that come from populations with close means. While the data from the ArnetMiner database come from populations with means that are different, as shown in Figure 2.10.



a) ArnetMiner Database

b) Web of Science Database

c) Scopus Database

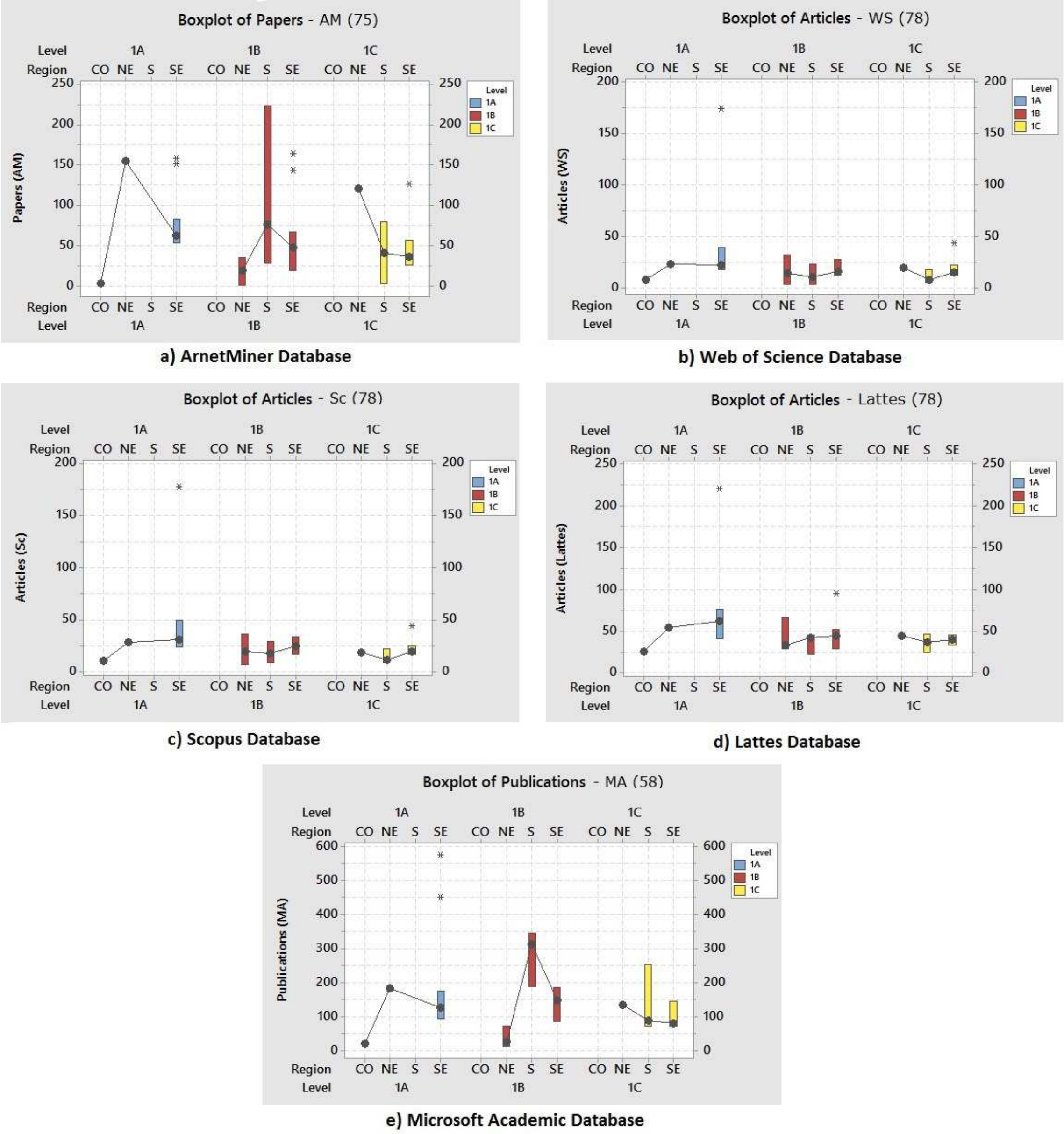
d) Google Scholar Database

e) Microsoft Academic Database

Figure 2.10: The number of citations in the different databases.

There are researchers who have a high production of articles and citations, which appear as outliers in all databases. This is the main reason why they appear in clusters with few researchers when the discriminant analysis and k-mean models are applied.

According to ArnetMiner and Microsof Academic databases, the number of articles authored by the researchers from the South region of Brazil is higher than in the production of articles from the Southeast; the average number of articles by the researchers of the Northeast is similar to the one of the researchers of the Southeast, as shown in Figure 2.11.



a) ArnetMiner Database

b) Web of Science Database

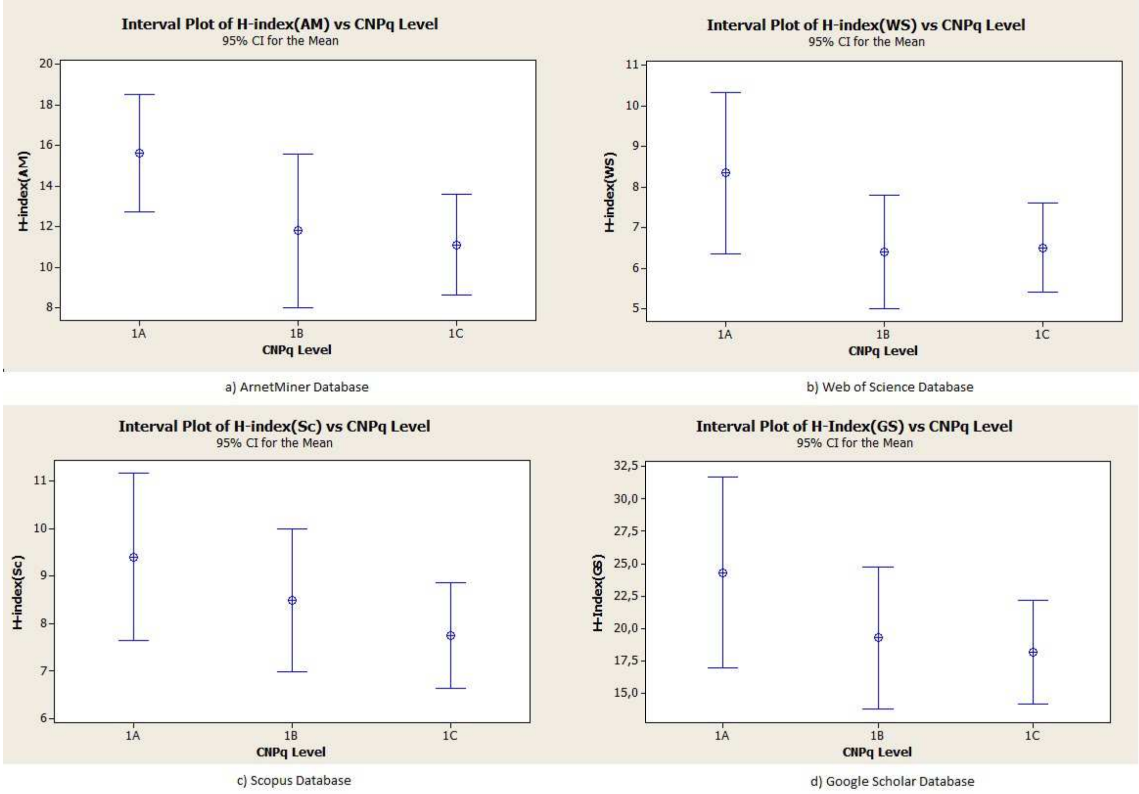
c) Scopus Database

d) Lattes Database

e) Microsoft Academic Database

Figure 2.11: Number of articles by the CNPq senior researchers in Computer Science by region.

Figure 2.12 shows h-index in the different databases. It is important to remark each of the Figures have a different scale. The Google Scholar database offers the largest absolute numbers and largest interval variation.



a) ArnetMiner Database

b) Web of Science Database

c) Scopus Database

d) Google Scholar Database

Figure 2.12: The h-index in the different databases.

2.3.2 The Principal Component Analysis Method

Using the Principal Component Analysis (PCA) method one can calculate the principal components of all the studied databases, as shown in Table 2.1. The PCA order depends on the number of variables in each database. The goal of PCA is also to order the clusters by regrouping, and to explain the maximum amount of variance with the lowest number of principal components.

Table 2.1: Eigenanalysis of the correlation matrix of all databases.

Database	Principal Component	1	2	3	4	5	6	7
ArnetMiner								
	Eigenvalue	4.970	0.847	0.506	0.294	0.239	0.114	0.028
	Proportion	0.710	0.121	0.072	0.042	0.034	0.016	0.004
	Cumulative	0.710	0.831	0.903	0.945	0.980	0.996	1.000
Web of Science								
	Eigenvalue	2.929	0.896	0.142	0.031			
	Proportion	0.732	0.224	0.036	0.008			
	Cumulative	0.732	0.956	0.992	1.000			
Scopus								
	Eigenvalue	2.954	0.870	0.148	0.026			
	Proportion	0.739	0.218	0.037	0.007			
	Cumulative	0.739	0.956	0.993	1.000			
Lattes								
	Eigenvalue	2.715	1.188	0.891	0.656	0.317	0.231	
	Proportion	0.453	0.198	0.149	0.109	0.053	0.039	
	Cumulative	0.453	0.651	0.799	0.909	0.961	1.000	
Goolgle Scholar								
	Eigenvalue	5.482	0.245	0.160	0.066	0.031	0.014	
	Proportion	0.914	0.041	0.027	0.011	0.005	0.002	
	Cumulative	0.914	0.955	0.981	0.992	0.998	1.000	
Microsoft Academic								
	Eigenvalue	2.960	0.779	0.198	0.061			
	Proportion	0.740	0.195	0.050	0.015			
	Cumulative	0.740	0.935	0.985	1.000			

The generation of the principal component as a linear combination of the main variables:

$$\begin{aligned} PC1(AM) &= 0.365*(activity)+0.382*(citation)+0.421*(h-index)+ \\ &+0.384*(papers)+0.413*(g-index)+0.293*(sociability)+ \\ &+0.374*(diversity). \end{aligned} \tag{2.3.1}$$

$$\begin{aligned} PC1(WS) &= 0.497*(articles)+0.575*(citation)+0.343*(md)+0.384*(papers)+ \\ &+0.552*(h-index). \end{aligned} \tag{2.3.2}$$

$$\begin{aligned} PC1(Sc) &= 0.481*(articles)+0.570*(citation)+0.376*(md)+0.384*(papers)+ \\ &+0.551*(h-index). \end{aligned} \tag{2.3.3}$$

$$\begin{aligned} PC1(Lattes) &= 0.389*(articles)+0.434*(papers)+0.362*(books)+ \\ &\quad +0.047*(patents)+0.517*(book\ chapters) \\ &\quad +0.510*(completed\ doctorate\ guidelines). \end{aligned} \tag{2.3.4}$$

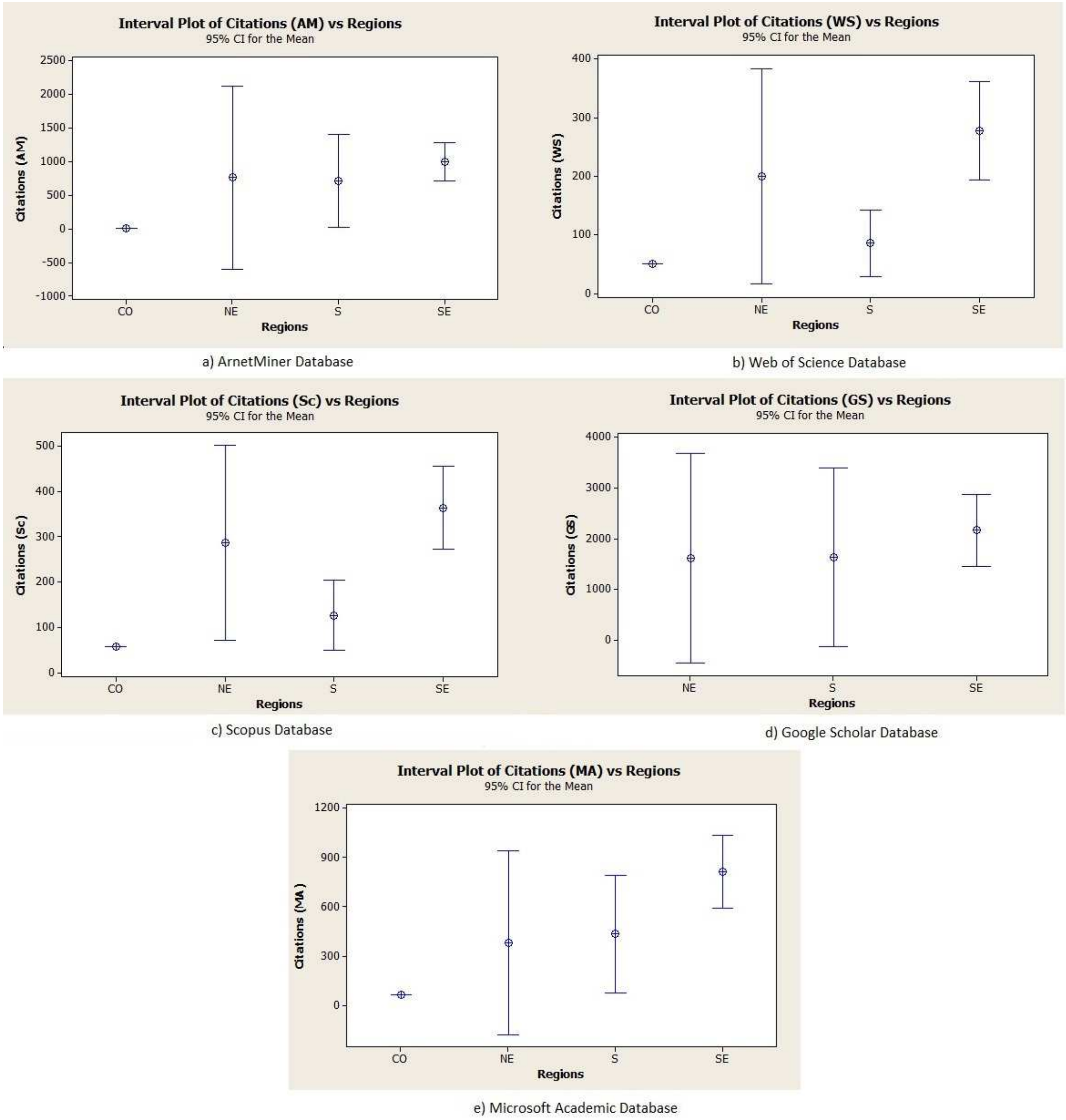
$$\begin{aligned} PC1(GS) &= 0.396*(activity)+0.417*(citations)+0.406*(h-index)+ \\ &\quad +0.416*(i10-index)+0.401*(citations(2008))+ \\ &\quad +0.413*(i10-index(2008)). \end{aligned} \tag{2.3.5}$$

$$\begin{aligned} PC1(MA) &= 0.516*(publications)+0.471*(citation)+0.461*(co-authors)+ \\ &\quad +0.547*(cited\ by). \end{aligned} \tag{2.3.6}$$

For each database, there is a variable which has the highest weight. For the Web of Science, Scopus, Google Scholar and Microsoft Academic databases the parameter with the highest weight is the number of citations of papers. While for the ArnetMiner database it is the *k*-index, and for the Lattes database is the number of book chapters published by the researcher.

Most of the coefficients in the Equations 6.0.39 to 6.0.29 have the same order of magnitude, ranging between 0.4 to 0.5, which means that the variables selected are representative. Otherwise, in the Lattes database the Equation 6.0.42, one can ignore the "number of patents" component, because it not significant, as its weight is 0.047.

The pattern of the data from the Web of Science and Scopus databases is similar, as is the number of citations of articles. The number of citations of the researchers from the Northeast and South regions is similar in the ArnetMiner, Google Scholar and Microsoft Academic databases, as shown Figure 2.13



a) ArnetMiner Database

b) Web of Science Database

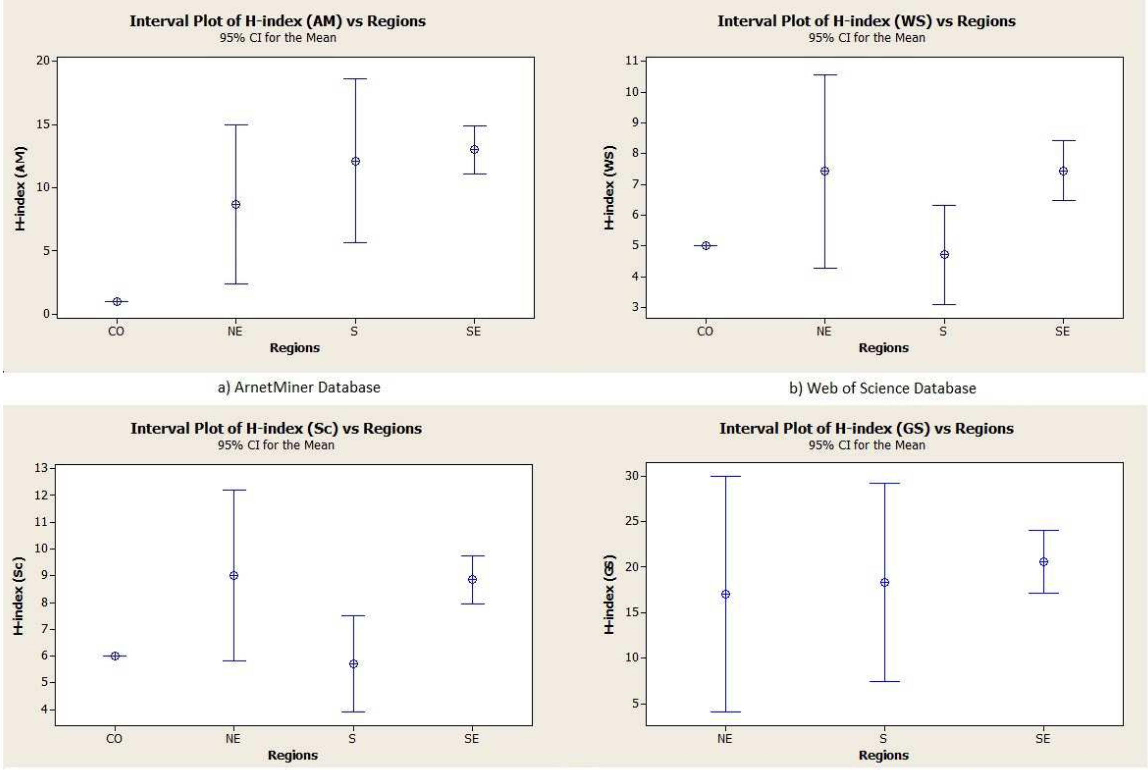
c) Scopus Database

d) Google Scholar Database

e) Microsoft Academic Database

Figure 2.13: Number of citations made to the seniors researchers in Computer Science from CNPq by region.

Figure 2.14 shows that *h*-index of the researchers from the Northeast region is comparable to the one researchers from the Southeast in the Web of Science, Scopus and Google Scholar databases.



a) ArnetMiner Database

b) Web of Science Database

c) Scopus Database

d) Google Scholar Database

Figure 2.14: h-index of the researchers by Region.

2.3.3 The Discriminant Analysis Method

To minimize the effect of the differences in scale in all the variables, their values were standardized, by subtracting the mean and dividing by the standard deviation before calculating the distance matrix. The cluster centroids and distance measures are in the standardized variable space before the distance matrix is calculated.

Using the linear discriminant analysis, all groups are assumed to have the same covariance matrix. The quadratic discrimination does not make such assumption, however.

The ratio is calculated by the following quotient, used in Tables 2.2, 2.3 and 2.4:

$$Ratio = \frac{N\ Regrouping}{N\ Total} \tag{2.3.7}$$

When the Ratio is equal to 1, it means that the accuracy rate of the method used in the classification of researchers is 100% according to the CNPq rank, for each level. The hit percentage is similar for all databases, observing the quadratic discriminant analysis model shown in Table 2.2. The linear model was applied and presented indices with a

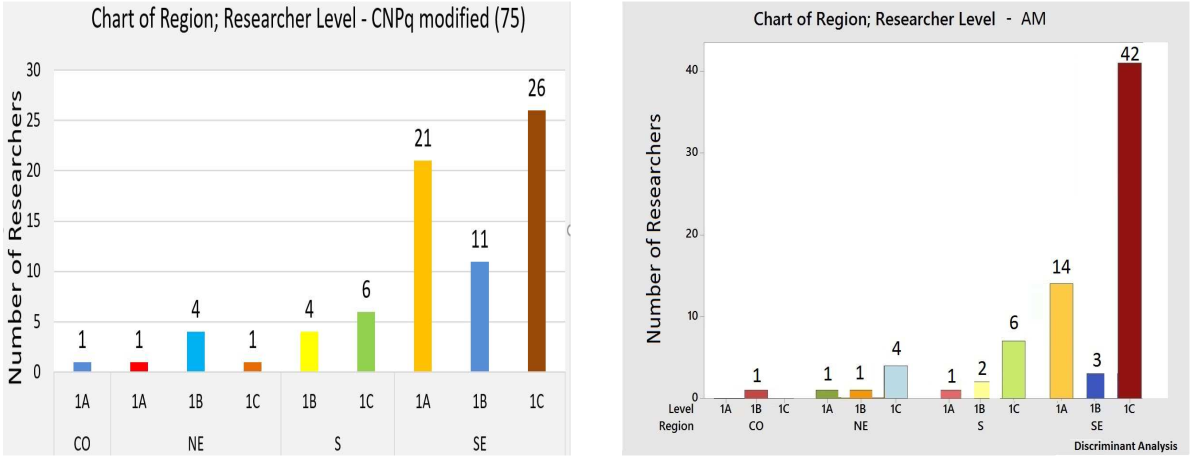
Table 2.2: Comparison of databases using the discriminant analysis method.

Database	1A	1B	1C	Sum	Correct
ArnetMiner					
N Total	23	19	33	75	
N Regrouping	16	07	52	75	
Ratio	0.695	0.368	1.575		
N Correct	12	05	28	45	
Proportion	0.522	0.263	0.848		0.600
Web of Science					
N Total	23	21	34	78	
N Regrouping	11	59	08	78	
Ratio	0.478	2.809	0.235		
N Correct	08	19	07	34	
Proportion	0.348	0.905	0.206		0.436
Scopus					
N Total	23	21	34	78	
N Regrouping	10	45	23	78	
Ratio	0.434	2.142	0.676		
N Correct	09	15	13	37	
Proportion	0.391	0.714	0.382		0.474
Lattes					
N Total	23	21	34	78	
N Regrouping	12	26	40	78	
Ratio	0.521	1.238	1.176		
N Correct	10	12	25	47	
Proportion	0.435	0.571	0.735		0.603
Google Scholar					
N Total	06	07	14	27	
N Regrouping	07	08	12	27	
Ratio	1.166	1.142	0.857		
N Correct	06	06	11	23	
Proportion	1.000	0.857	0.786		0.852
Microsoft Academic					
N Total	15	16	26	57	
N Regrouping	08	14	35	57	
Ratio	0.533	0.875	1.346		
N Correct	05	06	21	32	
Proportion	0.333	0.375	0.808		0.561

slightly better proportion.

The application of the discriminant analysis method resulted in a higher success rate in the Google Scholar database, 85.2%. Next, come the Lattes and ArnetMiner databases with 60.0%. The worst indices were obtained with Scopus and Web of Science databases which are around 47.40% and 43.60%, respectively.

Figure 2.15b shows the new distribution of reclassification of CNPq researchers using the discriminant analysis method for the ArnetMiner database. Figure 2.15a shows the distribution of CNPq classification for the 75 senior researchers, found in the ArnetMiner database to facilitate the comparison between databases. As one may observe, the new distribution is far stricter than the CNPq one as the $1A$ researcher from the CO was reclassified as $1B$. The number of $1B$ researchers from the NE dropped down to one and the number of researchers $1C$ from the SE region raised from 26 to 42. In the S region there was a decrease in the total number of researchers from 10 to 9, while in the SE there was an increase from 58 to 59 in the total number of researchers.

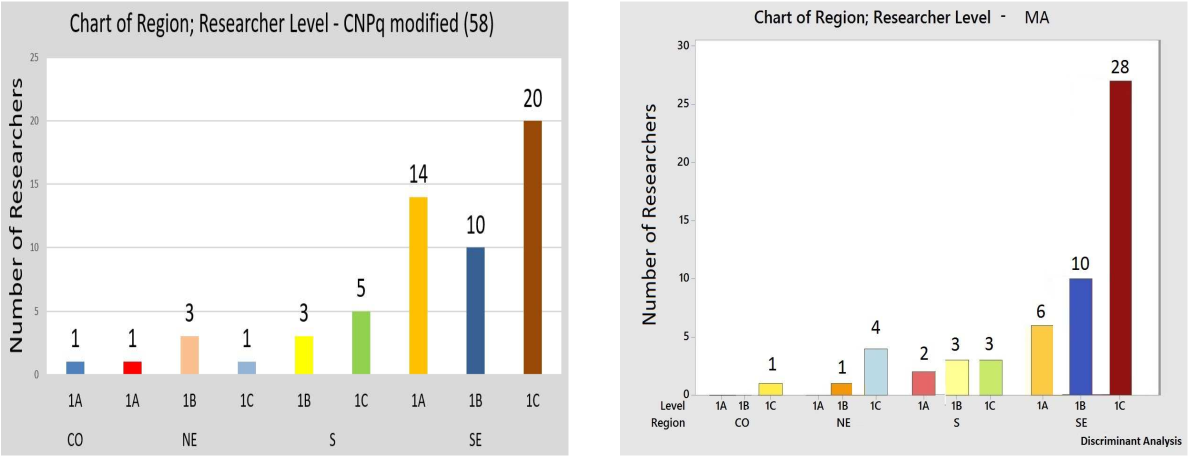


(a) Distribution of CNPq Researchers in ArnetMiner.

(b) New Distribution of CNPq Researchers.

Figure 2.15: Discriminant analysis approach. Source: ArnetMiner database.

Applying the discriminant analysis with the Microsoft Academic database, was obtained the new distribution of reclassification of researchers from CNPq, shown in Figure 2.16b. The comparison between databases can be made with the CNPq modified database shown in Figure 2.16a.



(a) Distribution of Reclassification of CNPq Modified.

(b) New Distribution of CNPq Researchers.

Figure 2.16: Discriminant analysis approach. Source: Microsoft Academic database.

2.3.4 The k-Mean Method

The k-mean approach reclassifies researchers from the data arranged in each database, considering the original classification of CNPq as the initial partition. Equations 6.0.39

to 6.0.29, using the k-mean method generated a rank with eight clusters as shown in Table 2.3, enabling the calculation of each cluster and then further regrouping into three clusters.

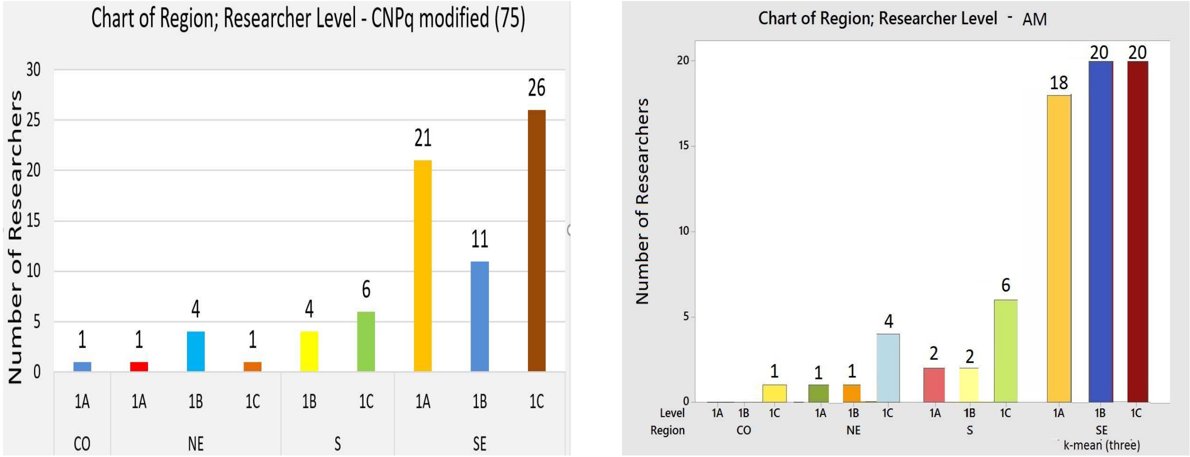
The same method was used for the analysis at the data shown in Table 2.4, except for the regrouping that was performed using four groups. At regrouping, by using the k-means approach, the database that showed the best proportion was Google Scholar. The worst proportion was provided by the Web of Science database, as shown in Table 2.3.

Table 2.3: Regrouping in three groups using k-mean method.

Database	1A	1B	1C	Sum	Correct
ArnetMiner					
N Total	23	19	33	75	
N Regrouping	21	23	31	75	
Ratio	0.913	1.210	0.939		
N Correct	10	06	11	27	
Proportion	0.434	0.315	0.333		0.360
Web of Science					
N Total	23	21	34	78	
N Regrouping	17	23	48	78	
Ratio	0.739	1.095	1.411		
N Correct	08	04	11	23	
Proportion	0.347	0.190	0.323		0.294
Scopus					
N Total	23	21	34	78	
N Regrouping	23	22	33	78	
Ratio	1.000	1.047	0.970		
N Correct	10	07	09	26	
Proportion	0.434	0.333	0.264		0.333
Lattes					
N Total	23	21	34	78	
N Regrouping	12	19	47	78	
Ratio	0.521	0.904	0.411		
N Correct	07	06	14	27	
Proportion	0.304	0.285	0.411		0.346
Google Scholar					
N Total	06	07	14	27	
N Regrouping	05	05	17	27	
Ratio	0.833	0.714	1.214		
N Correct	02	02	02	06	
Proportion	0.333	0.285	0.142		0.222
Microsoft Academic					
N Total	15	16	26	57	
N Regrouping	10	11	36	57	
Ratio	0.666	0.687	1.384		
N Correct	03	03	12	18	
Proportion	0.200	0.187	0.461		0.315

As a result, the new distribution of reclassification of the CNPq researchers, into three groups, using the k-mean approach, was applied to the ArnetMiner database, as shown in Figure 2.17b. The comparison between databases can be made with the CNPq modified database shown in Figure 2.17a.

Figure 2.18b shows the reclassification distribution of CNPq researchers, into three

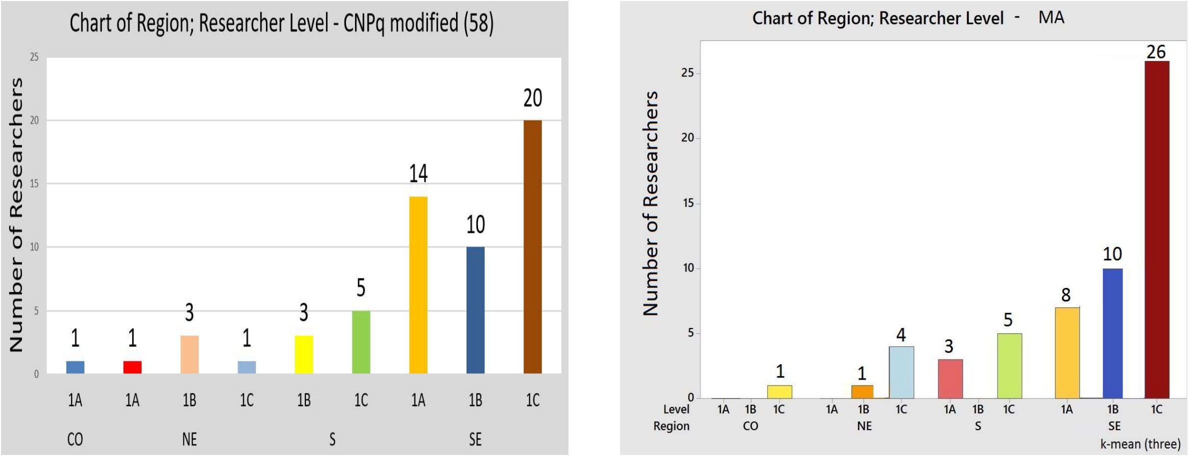


(a) Distribution of Reclassification of CNPq Modified.

(b) New Distribution of CNPq Researchers.

Figure 2.17: Discriminant analysis approach. Source: ArnetMiner database.

groups, using the discriminant analysis method for the Microsoft Academic database.



(a) Distribution of Reclassification of CNPq Modified.

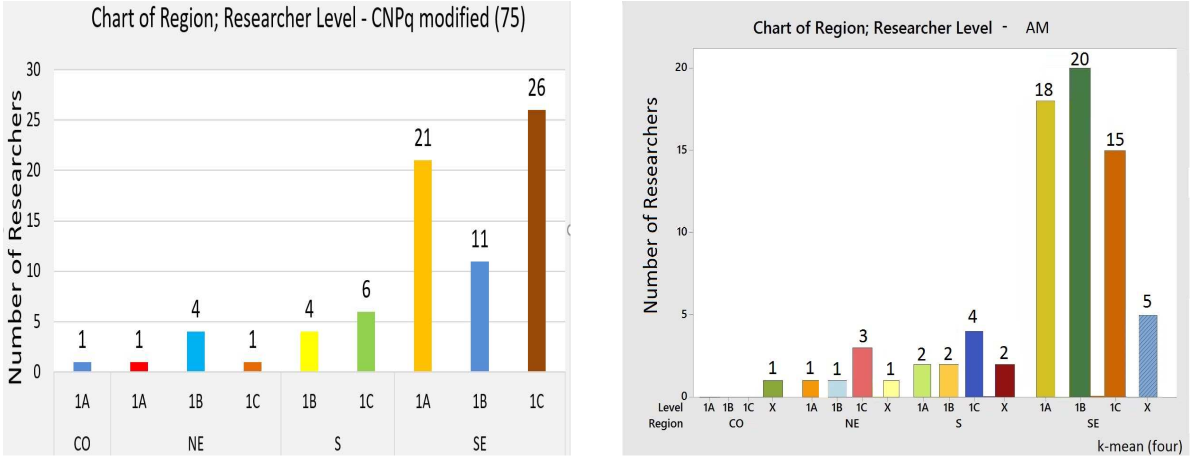
(b) New Distribution of CNPq Researchers.

Figure 2.18: Discriminant analysis approach. Source: Microsoft Academic database.

By creating a new group, called X, for the researchers who do not meet the criteria of research levels $1A$, $1B$ and $1C$, set at their centroids and by using the k-mean approach, we can make a reclassification into four groups, as shown in Table 2.4: The application of the k-mean method to the ArnetMiner database yielded a new reclassification into four groups as shown in Figure 2.19b. The comparison between databases can be made with the CNPq modified database shown in Figure 2.19a.

Table 2.4: Regrouping in four groups using k-mean method.

Database	1A	1B	1C	X	Sum	Correct
ArnetMiner						
N Total	23	19	33		75	
N Regrouping	21	23	22	09	75	
Ratio	0.913	1.210	0.667			
N Correct	10	06	11		27	
Proportion	0.434	0.315	0.333			0.360
Web of Science						
N Total	23	21	34		78	
N Regrouping	17	23	26	12	78	
Ratio	0.739	1.095	0.764			
N Correct	08	04	11		23	
Proportion	0.347	0.190	0.323			0.294
Scopus						
N Total	23	21	34		78	
N Regrouping	23	22	19	14	78	
Proportion	1.000	1.047	0.558			
N Correct	10	07	09		26	
Proportion	0.434	0.333	0.264			0.333
Lattes						
N Total	23	21	34		78	
N Regrouping	12	19	30	17	78	
Ratio	0.521	0.904	0.882			
N Correct	07	06	14		27	
Proportion	0.304	0.285	0.411			0.346
Google Scholar						
N Total	06	07	14		27	
N Regrouping	05	05	05	12	27	
Ratio	0.833	0.714	0.357			
N Correct	02	02	02		06	
Proportion	0.333	0.285	0.142			0.222
Microsoft Academic						
N Total	15	16	26		57	
N Regrouping	10	11	25	11	57	
Ratio	0.667	0.687	0.961			
N Correct	03	03	12		18	
Proportion	0.200	0.187	0.461			0.315

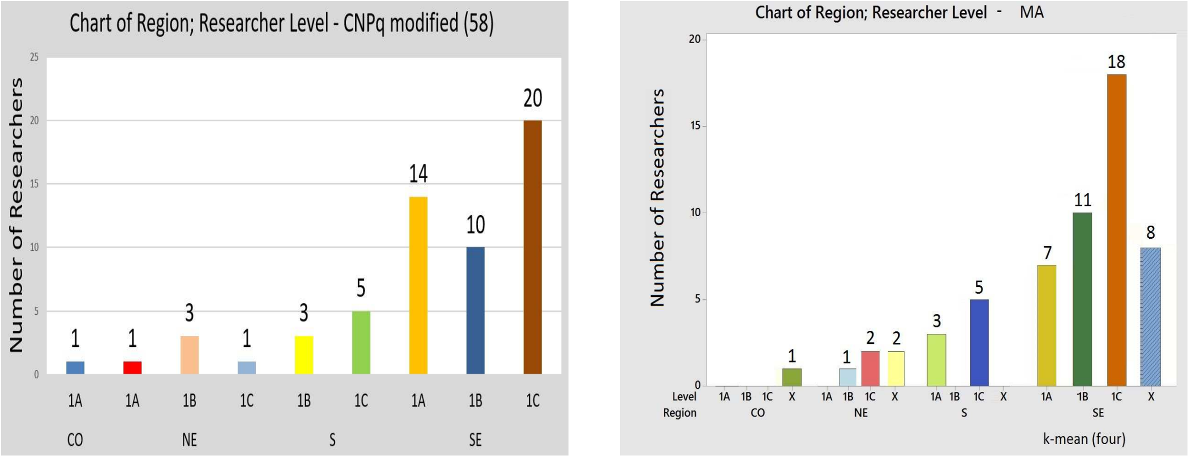


(a) Distribution of Reclassification of CNPq Modified.

(b) New Distribution of CNPq Researchers.

Figure 2.19: Discriminant analysis approach. Source: ArnetMiner database.

Figure 2.20b shows a new reclassification for researchers of CNPq, when using the *k*-mean method applied to the Microsoft Academic database.



(a) Distribution of Reclassification of CNPq Modified.

(b) New Distribution of CNPq Researchers.

Figure 2.20: Discriminant analysis approach. Source: Microsoft Academic database.

2.4 Conclusions

Taking into consideration the geographic regions of Brazil, one can see that the distribution of the researchers using the CNPq classification is uneven, with a strong bias towards the Southeast region of Brazil, and with a greater percentage of researchers in relation to the Northeast and South regions, about 70%. The Notheast region represents only 7% of the total number of researchers.

Notably for the Northeast region with regard to the volume of papers, publications by researchers in categories $1A$ and $1C$, are higher than in other regions of the country, taking as reference the ArnetMiner database. The researchers of the South region, belonging to the $1B$ level, had a number of publications above the other researchers from all regions, according to the analysis of the data in the Microsoft Academic database. In the Web of Science, Scopus and Lattes databases, the data values for publications of researchers, independently of the CNPq level, which are close to the average values, except those for the Midwest region.

Using the discriminant analysis approach, and making a comparison between the various databases, we found that the data from the Google Scholar database has the highest proportion of correct hits as 85.2%, for the researchers in classifications $1A$, $1B$ and $1C$, although the number of researchers listed in Google Scholar is small. The analysis of the Lattes and ArnetMiner databases provide equivalent results of about 60.0%. The worst hit rate is obtained with the Web of Science database, 43.6%. The highest proportion of classification similarity is using the data provided by the Google Scholar database in the case of the researchers in the $1A$ group, 100.0%. The worst classification results is obtained with the Web of Science database in group $1C$, 20.6%.

Applying the k-mean method, the databases that showed the highest correct proportion were the ArnetMiner and Lattes, with values of 36.0% and 34.6%, respectively. The results of the regrouping shown in Figure 2.21 is a histogram plot, representing the information from all the databases using the discriminant analysis and k-mean methods. For each database, there is a classification by hits.

One can find that the distribution reclassifications in the discriminant analysis method is more conservative because it meant a limited number of grants for researchers $1A$, a

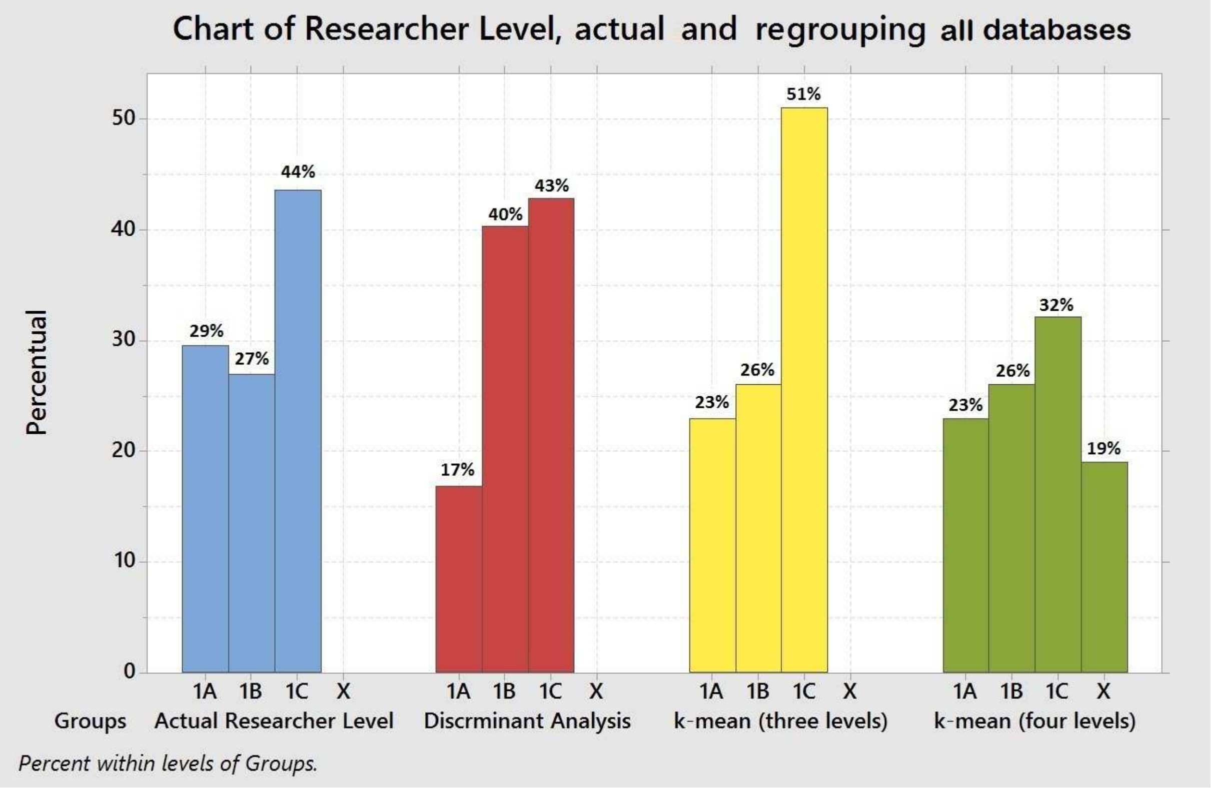


Figure 2.21: Actual CNPq researcher level and regrouping, using discriminant analysis and k-mean methods.

higher number to $1B$ and $1C$. The k-mean method for four levels, has the uniform and restricted distribution.

The discriminant analysis method was effective in identifying the appropriate database for the classification and ordering of researchers for $1A$, $1B$ and $1C$ levels, given the correct percentage of correlation to the CNPq classification. In this case, the ArnetMiner database represents the production of CNPq researchers correctly as well. Considering a set of all databases, the method remained robust with satisfactory results.

The criteria presented by CNPq does not differ much from the results presented by the methods described, discriminant analysis and k-mean. With respect to the ArnetMiner database, it showed a consistency in the set of given variables, considering the volume of data as well as the representative variables able to generate a clustering and reclassification models.

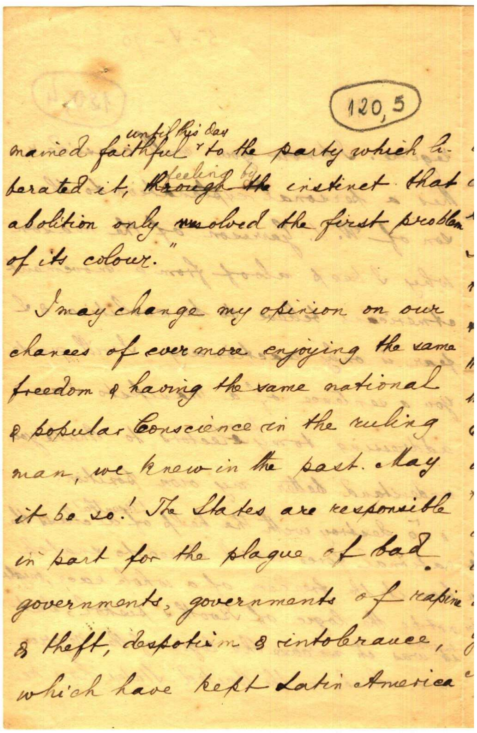
In conclusion, one can say that the CNPq rank is fair in general terms as the distortions in researcher classification are small and no regional bias was found.

3 ASSESSING BINARIZATION TECHNIQUES FOR DOCUMENT IMAGES WITH BACK-TO-FRONT INTERFERENCE

Documents written on both sides of translucent paper make visible the ink from one side on the other. This artefact, first described in the literature in reference (Lins, 1995), is called “back-to-front interference”, “bleeding”, (Kasturi, 2002), or “show-through”, (Sharma, 2001). The direct binarization of documents with such interference may yield unreadable documents. The technical literature presents several algorithms for suitably removing such an artefact, but the quality of the response given depends on a number of factors such as the strength of the interference, the hue of the paper with the aging, etc. This chapter assesses several algorithms to remove back-to-front interference.

3.1 Introduction

Whenever a document is typed or written on both sides of a sheet of paper and the opacity of the paper is such as to allow the back printing or writing to be visualized on the front side, such as in the Figure 3.1, the degree of difficulty for obtaining good segmentation increases enormously. A new set of hues of paper and printing colors appears. If the document is scanned either in true-color or gray-scale, the human brain is able to filter out that sort of noise keeping document readability, as shown in Figure 3.3, which presents a grayscale conversion of the image in Figure 3.1. This is not the case with automatic tools such as OCRs. Binarized images (black and white images) claim less storage space, allow for faster network transmission, and are more suitable to be processed by most commercial OCR tools.



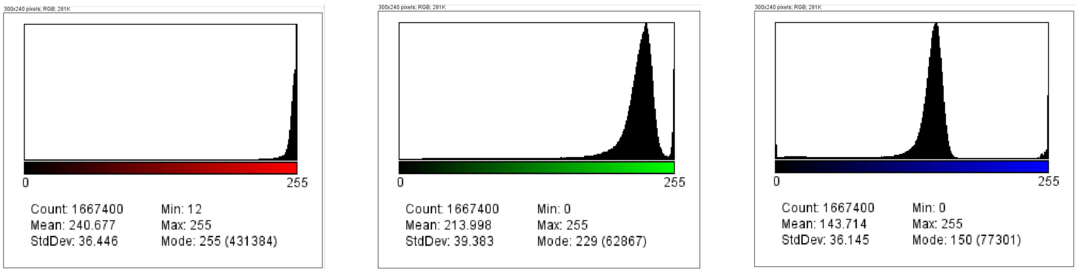
120,5

named faithful until this day to the party which liberated it, ~~through the~~ feeling by the instinct that abolition only ~~re~~solved the first problem of its colour."

I may change my opinion on our chances of ever more enjoying the same freedom & having the same national & popular Conscience in the ruling man, we knew in the past. May it be so! The States are responsible in part for the plague of bad governments, governments of rapine & theft, despotism & intolerance, which have kept Latin America

Figure 3.1: Historical Document from Nabuco Bequest with back-to-front interference.

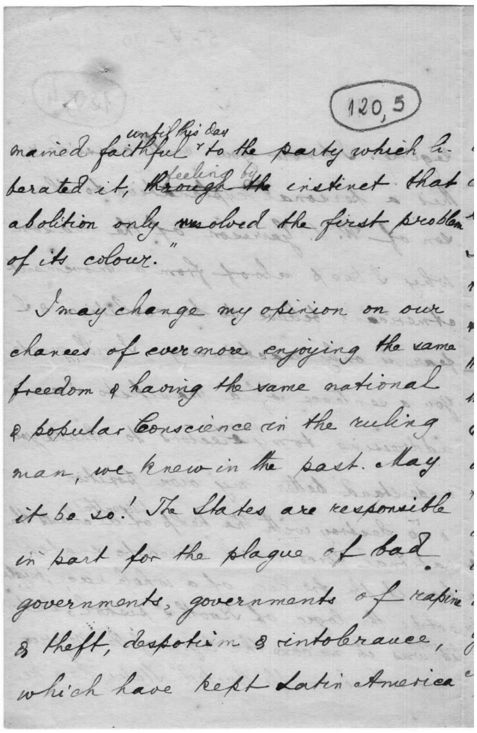
In a document such as the one presented in Figure 3.1, one expects to find three color clusters corresponding to the ink in the foreground, the paper background and the trespassed ink (the back-to-front interference). Unfortunately, no image representation provided such clustering to allow the easy filtering-out of the back-to-front interference. The RGB colour histogram may be seen in Figure 3.2 in which one may observe an overlap of the distributions of the background, the ink in the foreground, and the interfering ink from the backside.



(a) Red. (b) Green. (c) Blue.

Figure 3.2: RGB Colour Histogram of Figure 3.1.

The binarized version of the document in Figure 3.1 generated by the direct application of the binarization algorithm by using Jasc Paint Shop Pro ™ version 8 (Palette component: Gray values, Reduction component: nearest color, Palette weight: non-weighted) is completely unreadable, as one may observe in Figure 3.4. As one may also observe in Figure 3.1, the interfering ink of the backside of the paper does not look as sharp as the one in the foreground. There is a kind of blur effect that not only affects the contours of the text but also gives rise to different hues of the ink. Several papers in the literature addressed the back-to-front interference problem. Some authors use waterflow models (Oha, 2005), other researchers have used wavelet filtering (Tan, 2002), but the technique of most widespread use is thresholding (Kavallieratou & Antonopoulou, 2005), (Leedham, 2002) and Wang & Tan (2001). The most successful thresholding techniques for filtering out the back-to-front interference are based on the Shannon's entropy (Abramson, 1963) of the gray-scale document (Mello & Lins, 2000) and (Mello & Lins, 2002). Although recent advances were made in finding efficient algorithms that yield good quality images (Silva *et al.*, 2006), a final solution in the removal of the back-to-front interference is still sought off.



120,5

until this day
mained faithful to the party which li-
feeling by
berated it, ~~through the~~ instinct that
abolition only ~~un~~solved the first problem
of its colour."

I may change my opinion on our
chances of ever more enjoying the same
freedom & having the same national
& popular Conscience in the ruling
man, we knew in the past. May
it be so.' The States are responsible
in part for the plague of bad
governments, governments of rapine
& theft, despotism & intolerance,
which have kept Latin America

Figure 3.3: Gray-scale version of Figure 3.1.

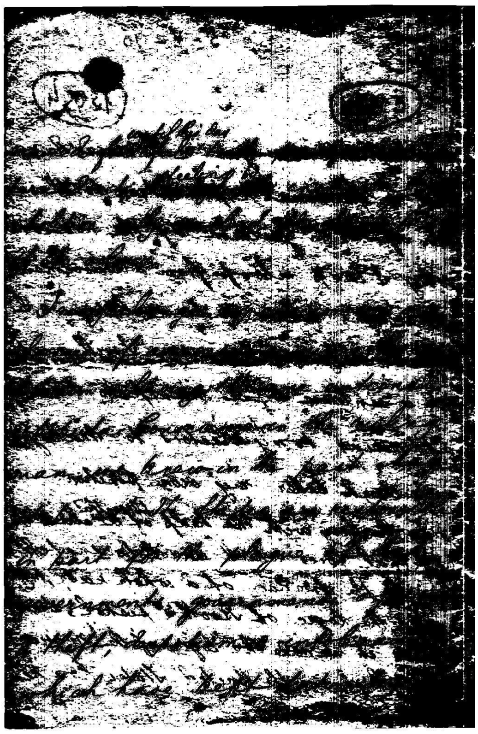


Figure 3.4: Binarized document of Figure 3.1.

The conversion from true-color image into 256 levels gray-scale images as an intermediate step towards image binarization has shown to be a valuable simplification, adopted by several binarization algorithms for removing back-to-front interference, by using the standard Equation 3.1.1 to calculate the value of the new pixel:

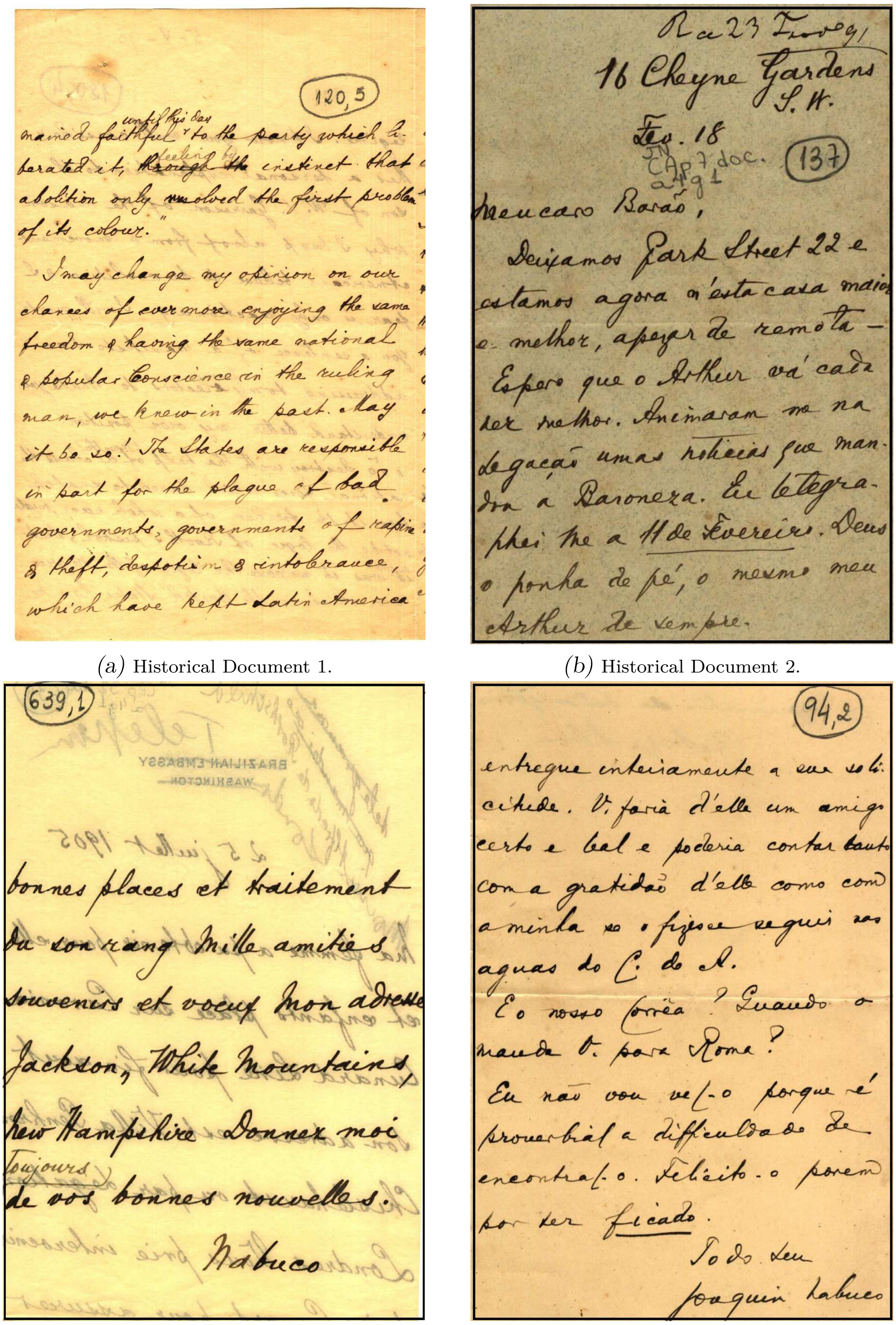
$$Gs = 0.299R + 0.587G + 0.114B, \tag{3.1.1}$$

where R, G, and B are the Red, Green and Blue values of the pixel in the true-color image, and the gray-level is the value of the pixel in the grayscale image. As one may observe from the image of the document exhibited in Figure 3.3, the grayscale conversion of documents tends to preserve their readability.

3.2 Generation and Filtering of Images with Synthetic Degradation

Historical documents with back-to-front interference are certainly the most difficult kind of document to binarize, as paper aging introduce non-uniform textures whose color distribution may overlap with the distribution of the colors from the writing in the back of the paper.

Sixteen images with different types of back-to-front interference were used in this study. Fourteen of the images are representative of a wide variety of historical documents and belong to the bequest of documents of Joaquim Nabuco shown in Figure 3.5 (a) to (n) held by the Joaquim Nabuco Foundation FUNDAJ (2016), a social science research centre in Recife, Brazil. Two other images, shown in Figure 3.6 (a) and (b) were from the LiveMemory project, Lins (2010a), and presents the variety of printed documents of the proceedings of the conferences of the SBT (Sociedade Brasileira de Telecomunicacões), The Brazilian Telecommunications Society.



120,5

mained faithful until this day to the party which li-
berated it, ~~through the~~ feeling by instinct that
abolition only ~~re~~solved the first problem
of its colour."

I may change my opinion on our
chances of ever more enjoying the same
freedom & having the same national
& popular Conscience in the ruling
man, we knew in the past. May
it be so! The States are responsible
in part for the plague of bad
governments, governments of rapine
& theft, despotism & intolerance,
which have kept Latin America

(a) Historical Document 1.

Rc 23 Fev 91

16 Cheyne Gardens
S.W.
Fev. 18

JN CAP 7 doc. a 491

137

Meu caro Barão,

Deixamos Park Street 22 e
estamos agora n'esta casa maior
e melhor, apezar de remota.
Espero que o Arthur vá cada
ser melhor. Animaram-no na
legação umas noticias que man-
dou a Baroneza. Eu telegra-
phei-lhe a 11 de Fevereiro. Deus
o ponha de pé, o mesmo meu
Arthur de sempre.

(b) Historical Document 2.

639,1

25 juillet 1905

bonnes places et traitement
du son rang. Mille amitiés
souvenirs et voeux. Mon adresse
Jackson, White Mountains,
New Hampshire. Donnez moi
Toujours
de vos bonnes nouvelles.

Nabuco

(c) Historical Document 3.

94,2

entregue inteiramente a sua soli-
citude. V. faria d'elle um amigo
certo e leal e poderia contar tanto
com a gratidão d'elle como com
a minha se o fizesse seguir nas
aguas do C. de A.

E o nosso Corrêa? Quando o
manda V. para Roma?

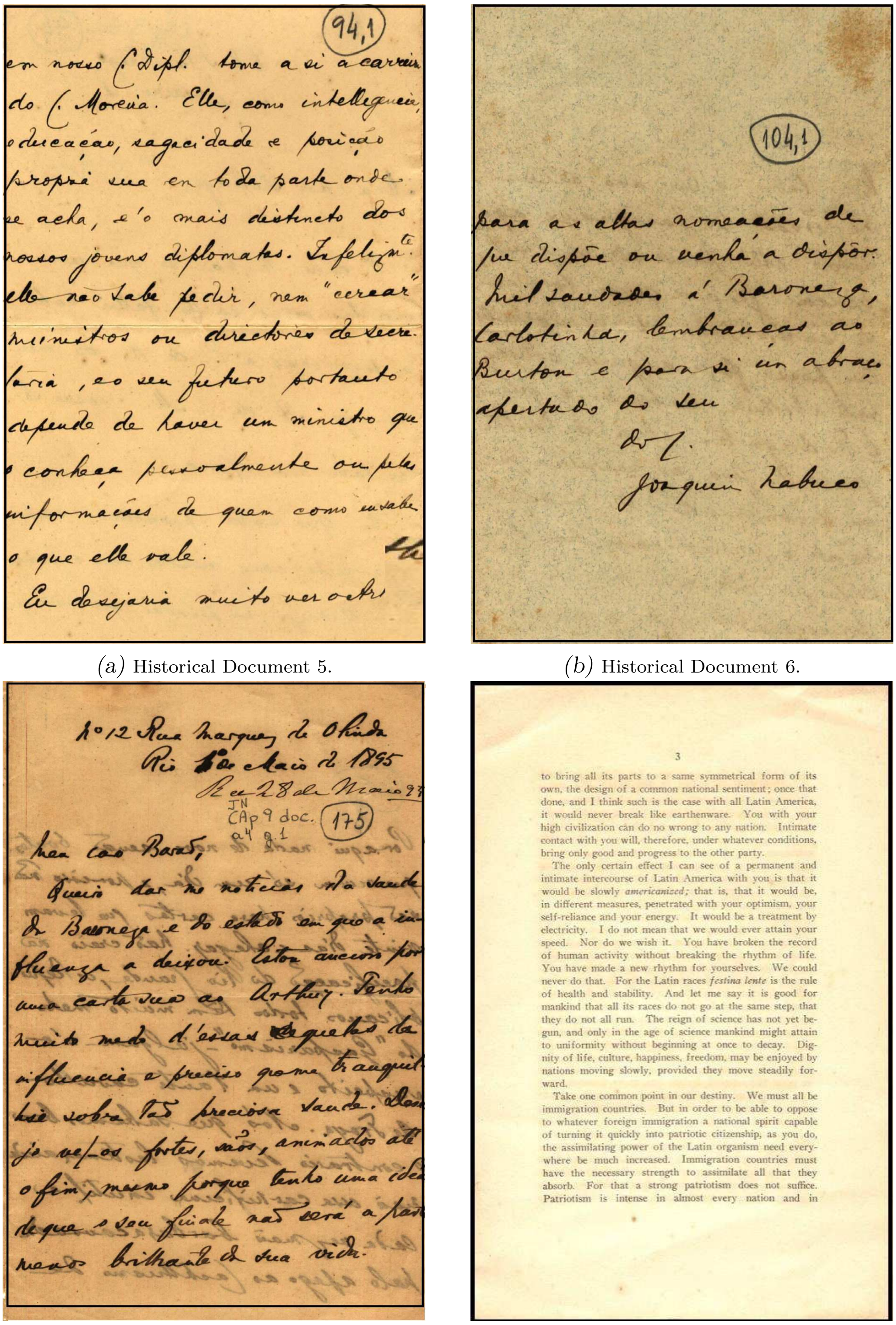
Eu não vou vel-o porque é
proverbial a difficuldade de
encontral-o. Felicito-o porem
por ter ficado.

Todo seu
Joaquim Nabuco

(d) Historical Document 4.

Figure 3.5: Historical Documents from Nabuco's bequest.

Although the technical literature presents several algorithms to filter out the back-to-front interference no algorithm is an "all case winner" as the strength of the interference may vary widely dependency on a number of factors such as the translucidity of the



94,1

em nosso C. Dipl. tome a si a carreira
do C. Moreira. Elle, como intelligencia,
educação, sagacidade e posição
propria sua em toda parte onde
se acha, é o mais distincto dos
nossos jovens diplomatas. Infelizmte.
elle não sabe pedir, nem "cercar"
ministros ou directores de secre-
taria, e o seu futuro portanto
depende de haver um ministro que
o conheça pessoalmente ou pelas
informações de quem como eu sabe
o que elle vale.
Eu desejaria muito ver o Ar

(a) Historical Document 5.

104,1

para as altas nomeações de
que dispõe ou venha a dispôr.
Mil saudades á Baroneza,
Carlotinha, lembranças ao
Burton e para si um abraço
apertado do seu
do C.
Joaquim Nabuco

(b) Historical Document 6.

N° 12 Rua Marquez de Olinda
Rio 1° de Maio de 1895
Rec 28 de Maio 9
JN
CAp 9 doc.
a4 g.1
175

Meu caro Barão,
Queira dar-me noticias da saude
da Baroneza e do estado em que a in-
fluenza a deixou. Estou ancioso por
uma carta sua ao Arthur. Tenho
muito medo d'essas sequelas da
influencia e preciso que me tranquil-
lise sobre tão preciosa saude. Dese-
jo vel-os fortes, sãos, animados até
o fim, mesmo porque tenho uma idea
de que o seu final não será a part
menos brilhante da sua vida.

(c) Historical Document 7.

3

to bring all its parts to a same symmetrical form of its own, the design of a common national sentiment; once that done, and I think such is the case with all Latin America, it would never break like earthenware. You with your high civilization can do no wrong to any nation. Intimate contact with you will, therefore, under whatever conditions, bring only good and progress to the other party.

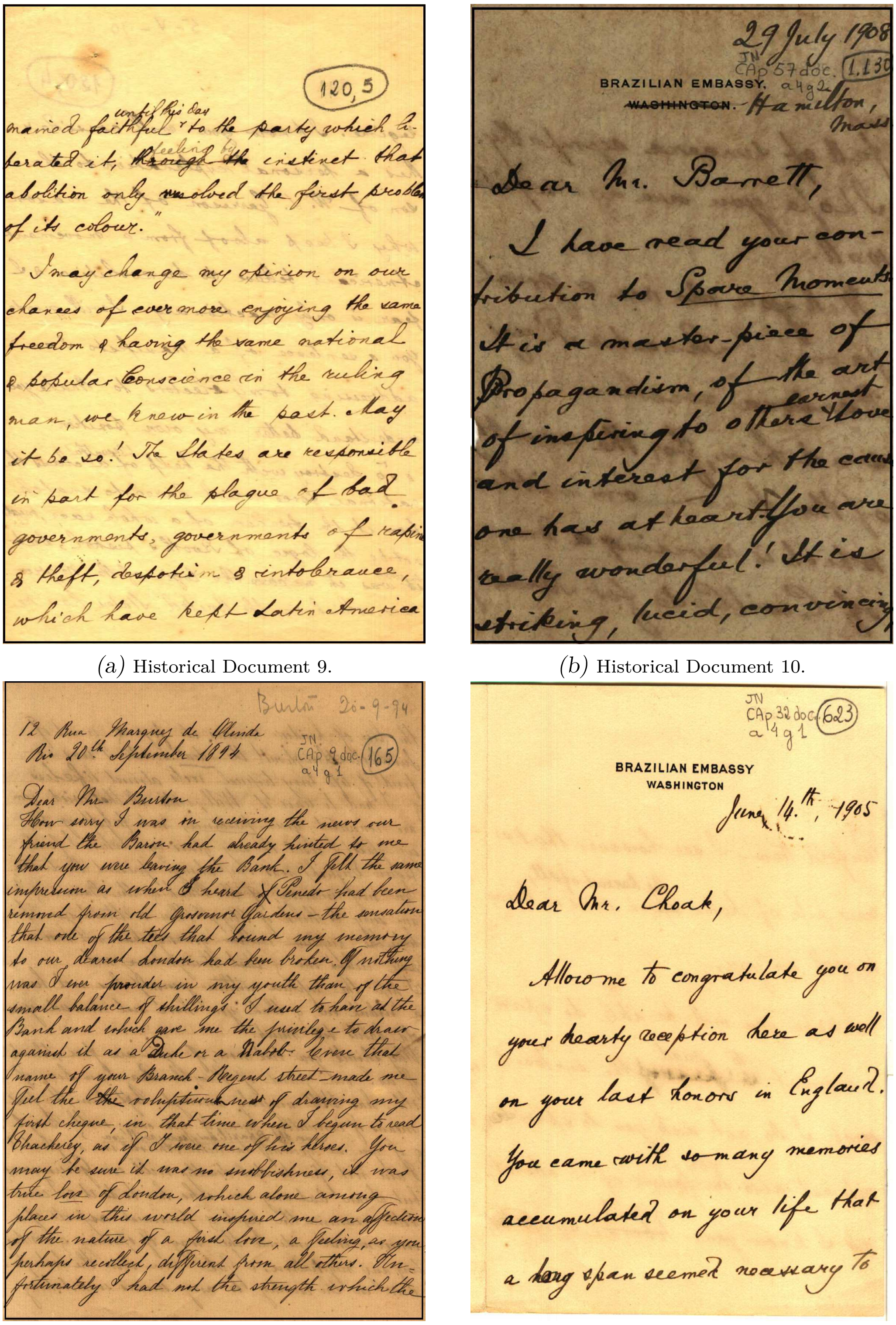
The only certain effect I can see of a permanent and intimate intercourse of Latin America with you is that it would be slowly *americanized;* that is, that it would be, in different measures, penetrated with your optimism, your self-reliance and your energy. It would be a treatment by electricity. I do not mean that we would ever attain your speed. Nor do we wish it. You have broken the record of human activity without breaking the rhythm of life. You have made a new rhythm for yourselves. We could never do that. For the Latin races *festina lente* is the rule of health and stability. And let me say it is good for mankind that all its races do not go at the same step, that they do not all run. The reign of science has not yet begun, and only in the age of science mankind might attain to uniformity without beginning at once to decay. Dignity of life, culture, happiness, freedom, may be enjoyed by nations moving slowly, provided they move steadily forward.

Take one common point in our destiny. We must all be immigration countries. But in order to be able to oppose to whatever foreign immigration a national spirit capable of turning it quickly into patriotic citizenship, as you do, the assimilating power of the Latin organism need everywhere be much increased. Immigration countries must have the necessary strength to assimilate all that they absorb. For that a strong patriotism does not suffice. Patriotism is intense in almost every nation and in

(d) Historical Document 8.

Figure 3.6: Historical Documents from Nabuco's bequest.

paper, its porosity, the kind of the ink used (water or oil based), the color of ink, the age of document, the conditions of document storage, etc.



120,5

mained faithful until this day to the party which li-
berated it, ~~through the~~ feeling by cristinct that
abolition only resolved the first problem
of its colour."

I may change my opinion on our
chances of ever more enjoying the same
freedom & having the same national
& popular Conscience in the ruling
man, we knew in the past. May
it be so! The States are responsible
in part for the plague of bad
governments, governments of rapine
& theft, despotism & intolerance,
which have kept Latin America

(a) Historical Document 9.

29 July 1908
JN
CAp 57 doc. 1.130
a4g2

BRAZILIAN EMBASSY,
~~WASHINGTON.~~ Hamilton,
Mass.

Dear Mr. Barrett,

I have read your con-
tribution to Spare Moments
It is a master-piece of
Propagandism, of the art
of inspiring to others earnest love
and interest for the cause
one has at heart. You are
really wonderful! It is
striking, lucid, convincing

(b) Historical Document 10.

Burton 20-9-94

12 Rua Marquez de Olinda
Rio 20th September 1894

JN
CAp 9 doc. 165
a4g1

Dear Mr Burton
How sorry I was on receiving the news our
friend the Baron had already hinted to me
that you were leaving the Bank. I felt the same
impression as when I heard Penedo had been
removed from old Grosvenor Gardens – the sensation
that one of the ties that bound my memory
to our dearest London had been broken. Of nothing
was I ever prouder in my youth than of the
small balance of shillings I used to have at the
Bank and which gave me the privilege to draw
against it as a Duke or a Nabob. Even that
name of your Branch – Regent street – made me
feel the ~~the~~ voluptuousness of drawing my
first cheque, in that time when I begun to read
Thackeray, as if I were one of his heroes. You
may be sure it was no snobbishness, it was
true love of London, which alone among
places in this world inspired me an affection
of the nature of a first love, a feeling, as you
perhaps recollect, different from all others. Un-
fortunately I had not the strength which the

(c) Historical Document 11.

JN
CAp 32 doc. 623
a4g1

BRAZILIAN EMBASSY
WASHINGTON

June 14th, 1905

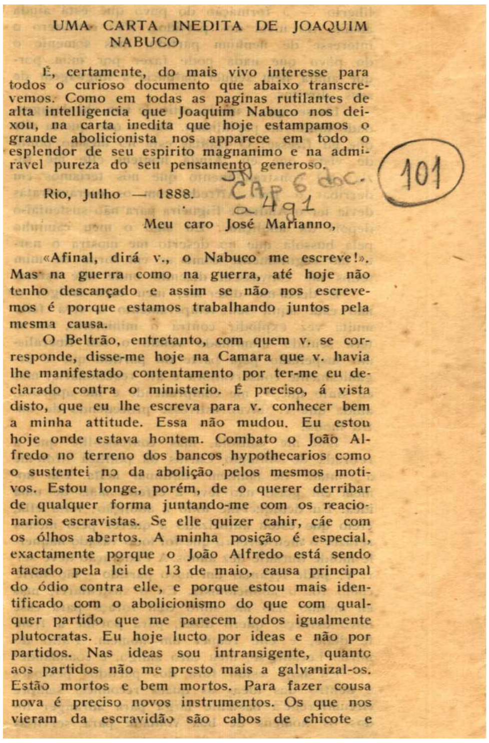
Dear Mr. Choak,

Allow me to congratulate you on
your hearty reception here as well
on your last honors in England.
You came with so many memories
accumulated on your life that
a long span seemed necessary to

(d) Historical Document 12.

Figure 3.7: Historical Documents from Nabuco's bequest.

The goal here is to create a controlled environment, to generate synthesized images to filter them using the most successful algorithms in the literature to be able to detect the parameters in which their performance and their efficiency reach then the best results.



UMA CARTA INEDITA DE JOAQUIM NABUCO

É, certamente, do mais vivo interesse para todos o curioso documento que abaixo transcrevemos. Como em todas as paginas rutilantes de alta intelligencia que Joaquim Nabuco nos deixou, na carta inedita que hoje estampamos o grande abolicionista nos apparece em todo o esplendor de seu espirito magnanimo e na admiravel pureza do seu pensamento generoso.

JN CAP 6 doc. a 491

101

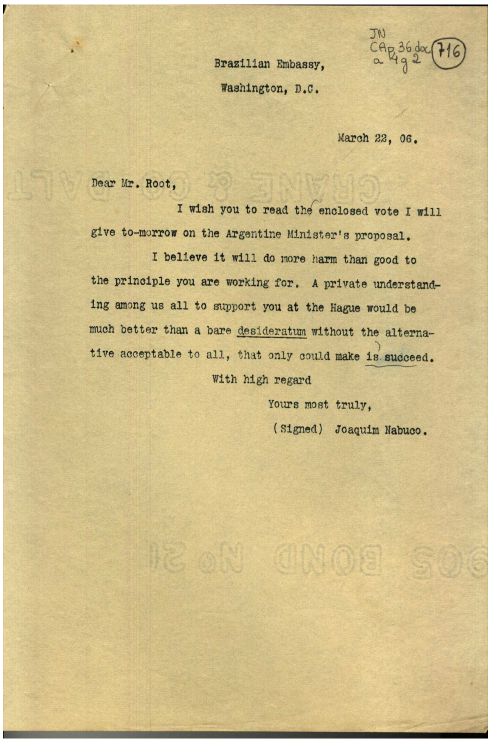
Rio, Julho — 1888.

Meu caro José Marianno,

«Afinal, dirá v., o Nabuco me escreve!». Mas na guerra como na guerra, até hoje não tenho descançado e assim se não nos escrevemos é porque estamos trabalhando juntos pela mesma causa.

O Beltrão, entretanto, com quem v. se corresponde, disse-me hoje na Camara que v. havia lhe manifestado contentamento por ter-me eu declarado contra o ministerio. É preciso, á vista disto, que eu lhe escreva para v. conhecer bem a minha attitude. Essa não mudou. Eu estou hoje onde estava hontem. Combato o João Alfredo no terreno dos bancos hypothecarios como o sustentei no da abolição pelos mesmos motivos. Estou longe, porém, de o querer derribar de qualquer forma juntando-me com os reacionarios escravistas. Se elle quizer cahir, cáe com os ólhos abertos. A minha posição é especial, exactamente porque o João Alfredo está sendo atacado pela lei de 13 de maio, causa principal do ódio contra elle, e porque estou mais identificado com o abolicionismo do que com qualquer partido que me parecem todos igualmente plutocratas. Eu hoje lucto por ideas e não por partidos. Nas ideas sou intransigente, quanto aos partidos não me presto mais a galvanizal-os. Estão mortos e bem mortos. Para fazer cousa nova é preciso novos instrumentos. Os que nos vieram da escravidão são cabos de chicote e

(a) Historical Document 13.



JN CAP 36 doc a 492

716

Brazilian Embassy,
Washington, D.C.

March 22, 06.

Dear Mr. Root,

I wish you to read the enclosed vote I will give to-morrow on the Argentine Minister's proposal.

I believe it will do more harm than good to the principle you are working for. A private understanding among us all to support you at the Hague would be much better than a bare desideratum without the alternative acceptable to all, that only could make is succeed.

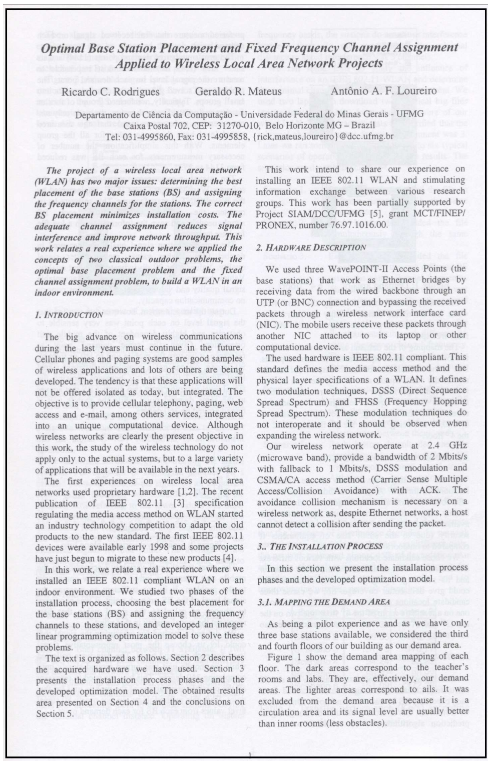
With high regard

Yours most truly,

(Signed) Joaquim Nabuco.

(b) Historical Document 14.

Figure 3.8: Historical Documents from Nabuco's bequest.



Optimal Base Station Placement and Fixed Frequency Channel Assignment Applied to Wireless Local Area Network Projects

Ricardo C. Rodrigues Geraldo R. Mateus Antônio A. F. Loureiro

Departamento de Ciência da Computação - Universidade Federal do Minas Gerais - UFMG
Caixa Postal 702, CEP: 31270-010, Belo Horizonte MG – Brazil
Tel: 031-4995860, Fax: 031-4995858, {rick,mateus,loureiro}@dcc.ufmg.br

The project of a wireless local area network (WLAN) has two major issues: determining the best placement of the base stations (BS) and assigning the frequency channels for the stations. The correct BS placement minimizes installation costs. The adequate channel assignment reduces signal interference and improve network throughput. This work relates a real experience where we applied the concepts of two classical outdoor problems, the optimal base placement problem and the fixed channel assignment problem, to build a WLAN in an indoor environment.

1. INTRODUCTION

The big advance on wireless communications during the last years must continue in the future. Cellular phones and paging systems are good samples of wireless applications and lots of others are being developed. The tendency is that these applications will not be offered isolated as today, but integrated. The objective is to provide cellular telephony, paging, web access and e-mail, among others services, integrated into an unique computational device. Although wireless networks are clearly the present objective in this work, the study of the wireless technology do not apply only to the actual systems, but to a large variety of applications that will be available in the next years.

The first experiences on wireless local area networks used proprietary hardware [1,2]. The recent publication of IEEE 802.11 [3] specification regulating the media access method on WLAN started an industry technology competition to adapt the old products to the new standard. The first IEEE 802.11 devices were available early 1998 and some projects have just begun to migrate to these new products [4].

In this work, we relate a real experience where we installed an IEEE 802.11 compliant WLAN on an indoor environment. We studied two phases of the installation process, choosing the best placement for the base stations (BS) and assigning the frequency channels to these stations, and developed an integer linear programming optimization model to solve these problems.

The text is organized as follows. Section 2 describes the acquired hardware we have used. Section 3 presents the installation process phases and the developed optimization model. The obtained results area presented on Section 4 and the conclusions on Section 5.

This work intend to share our experience on installing an IEEE 802.11 WLAN and stimulating information exchange between various research groups. This work has been partially supported by Project SIAM/DCC/UFMG [5], grant MCT/FINEP/PRONEX, number 76.97.1016.00.

2. HARDWARE DESCRIPTION

We used three WavePOINT-II Access Points (the base stations) that work as Ethernet bridges by receiving data from the wired backbone through an UTP (or BNC) connection and bypassing the received packets through a wireless network interface card (NIC). The mobile users receive these packets through another NIC attached to its laptop or other computational device.

The used hardware is IEEE 802.11 compliant. This standard defines the media access method and the physical layer specifications of a WLAN. It defines two modulation techniques, DSSS (Direct Sequence Spread Spectrum) and FHSS (Frequency Hopping Spread Spectrum). These modulation techniques do not interoperate and it should be observed when expanding the wireless network.

Our wireless network operate at 2.4 GHz (microwave band), provide a bandwidth of 2 Mbits/s with fallback to 1 Mbits/s, DSSS modulation and CSMA/CA access method (Carrier Sense Multiple Access/Collision Avoidance) with ACK. The avoidance collision mechanism is necessary on a wireless network as, despite Ethernet networks, a host cannot detect a collision after sending the packet.

3.. THE INSTALLATION PROCESS

In this section we present the installation process phases and the developed optimization model.

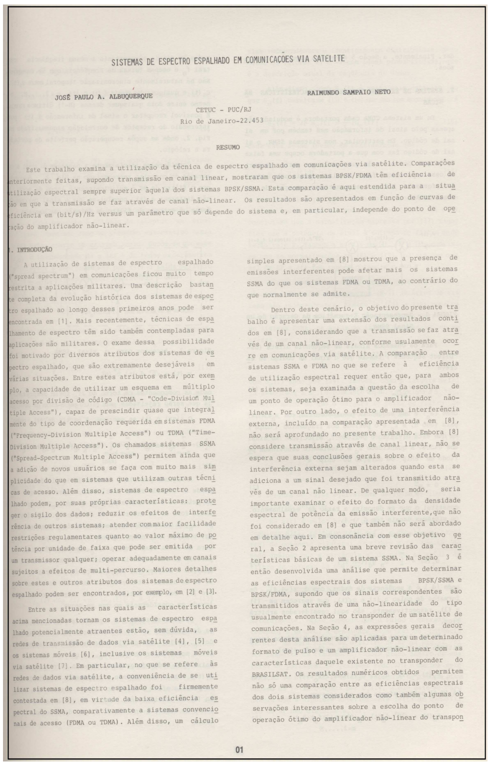
3.1. MAPPING THE DEMAND AREA

As being a pilot experience and as we have only three base stations available, we considered the third and fourth floors of our building as our demand area.

Figure 1 show the demand area mapping of each floor. The dark areas correspond to the teacher's rooms and labs. They are, effectively, our demand areas. The lighter areas correspond to ails. It was excluded from the demand area because it is a circulation area and its signal level are usually better than inner rooms (less obstacles).

1

(a) Printed Document 15.



SISTEMAS DE ESPECTRO ESPALHADO EM COMUNICAÇÕES VIA SATELITE

JOSÉ PAULO A. ALBUQUERQUE **RAIMUNDO SAMPAIO NETO**

CETUC - PUC/RJ
Rio de Janeiro-22.453

RESUMO

Este trabalho examina a utilização da técnica de espectro espalhado em comunicações via satélite. Comparações nteriormente feitas, supondo transmissão em canal linear, mostraram que os sistemas BPSK/FDMA têm eficiência de tilização espectral sempre superior àquela dos sistemas BPSK/SSMA. Esta comparação é aqui estendida para a situaão em que a transmissão se faz através de canal não-linear. Os resultados são apresentados em função de curvas de ficiência em (bit/s)/Hz versus um parâmetro que só depende do sistema e, em particular, independe do ponto de operação do amplificador não-linear.

1. INTRODUÇÃO

A utilização de sistemas de espectro espalhado ("spread spectrum") em comunicações ficou muito tempo restrita a aplicações militares. Uma descrição bastante completa da evolução histórica dos sistemas de espectro espalhado ao longo desses primeiros anos pode ser encontrada em [1]. Mais recentemente, técnicas de espalhamento de espectro têm sido também contempladas para aplicações não militares. O exame dessa possibilidade foi motivado por diversos atributos dos sistemas de espectro espalhado, que são extremamente desejáveis em várias situações. Entre estes atributos está, por exemplo, a capacidade de utilizar um esquema em múltiplo acesso por divisão de código (CDMA - "Code-Division Multiple Access"), capaz de prescindir quase que integralmente do tipo de coordenação requerida em sistemas FDMA ("Frequency-Division Multiple Access") ou TDMA ("Time-Division Multiple Access"). Os chamados sistemas SSMA ("Spread-Spectrum Multiple Access") permitem ainda que a adição de novos usuários se faça com muito mais simplicidade do que em sistemas que utilizam outras técnicas de acesso. Além disso, sistemas de espectro espalhado podem, por suas próprias características: proteger o sigilo dos dados; reduzir os efeitos de interferência de outros sistemas; atender com maior facilidade restrições regulamentares quanto ao valor máximo de potência por unidade de faixa que pode ser emitida por um transmissor qualquer; operar adequadamente em canais sujeitos a efeitos de multi-percurso. Maiores detalhes sobre estes e outros atributos dos sistemas de espectro espalhado podem ser encontrados, por exemplo, em [2] e [3].

Entre as situações nas quais as características acima mencionadas tornam os sistemas de espectro espalhado potencialmente atraentes estão, sem dúvida, as redes de transmissão de dados via satélite [4], [5] e os sistemas móveis [6], inclusive os sistemas móveis via satélite [7]. Em particular, no que se refere às redes de dados via satélite, a conveniência de se utilizar sistemas de espectro espalhado foi firmemente contestada em [8], em virtude da baixa eficiência espectral do SSMA, comparativamente a sistemas convencionais de acesso (FDMA ou TDMA). Além disso, um cálculo simples apresentado em [8] mostrou que a presença de emissões interferentes pode afetar mais os sistemas SSMA do que os sistemas FDMA ou TDMA, ao contrário do que normalmente se admite.

Dentro deste cenário, o objetivo do presente trabalho é apresentar uma extensão dos resultados contidos em [8], considerando que a transmissão se faz através de um canal não-linear, conforme usualmente ocorre em comunicações via satélite. A comparação entre sistemas SSMA e FDMA no que se refere à eficiência de utilização espectral requer então que, para ambos os sistemas, seja examinada a questão da escolha de um ponto de operação ótimo para o amplificador não-linear. Por outro lado, o efeito de uma interferência externa, incluído na comparação apresentada em [8], não será aprofundado no presente trabalho. Embora [8] considere transmissão através de canal linear, não se espera que suas conclusões gerais sobre o efeito da interferência externa sejam alterados quando esta se adiciona a um sinal desejado que foi transmitido através de um canal não linear. De qualquer modo, seria importante examinar o efeito do formato da densidade espectral de potência da emissão interferente, que não foi considerado em [8] e que também não será abordado em detalhe aqui. Em consonância com esse objetivo geral, a Seção 2 apresenta uma breve revisão das características básicas de um sistema SSMA. Na Seção 3 é então desenvolvida uma análise que permite determinar as eficiências espectrais dos sistemas BPSK/SSMA e BPSK/FDMA, supondo que os sinais correspondentes são transmitidos através de uma não-linearidade do tipo usualmente encontrado no transponder de um satélite de comunicações. Na Seção 4, as expressões gerais decorrentes desta análise são aplicadas para um determinado formato de pulso e um amplificador não-linear com as características daquele existente no transponder do BRASILSAT. Os resultados numéricos obtidos permitem não só uma comparação entre as eficiências espectrais dos dois sistemas considerados como também algumas observações interessantes sobre a escolha do ponto de operação ótimo do amplificador não-linear do transpon

01

(b) Printed Document 16.

Figure 3.9: Printed Documents from the SBrT file.

The block diagram in Figure 3.10 sketches the environment developed. The method used for generating synthesized documents followed reference Lins (2010b):

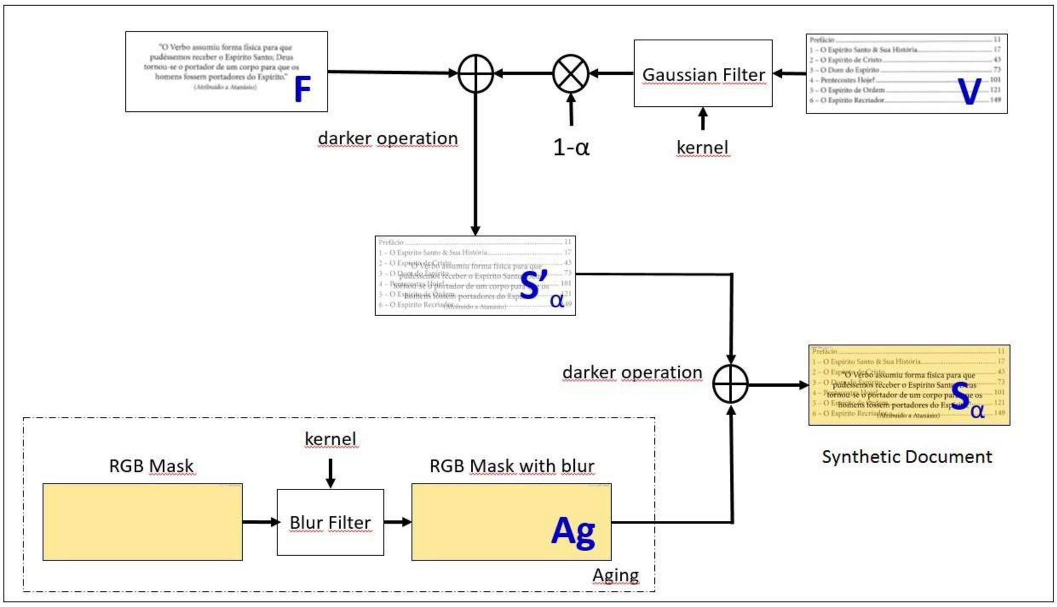


Figure 3.10: Block Diagram of the synthetic image generator.

Step 1 - The ground-truth images

The first step of the generation of synthetic images was to produce a set of images that covers all the universe of text documents: typed in mechanical typewriters, printed in inkjet, laser, offset in most usual colors (black, blue, red), handwritten with different kinds of pen (fountain, ballpen, felt pen) from different manufacturers, using black and blue ink. Such documents were typed/printed/written in good quality A4 white papers. Such images were scanned using a flatbed scanner set to a resolution of 300 dpi in true-color (24 bits RGB) yielding raster images standardized in $1,120 \times 466$ pixels, such as the ones in Figure 3.11 are used. The images obtained were binarized using the standard binarization algorithm in Jasc Paint Shop Pro version 8 and are used as ground truth images and also in the generation of the synthetic images. Currently, twenty different images were generated.

Step 2 - A third image S'_α is generated by composing the images F and V.

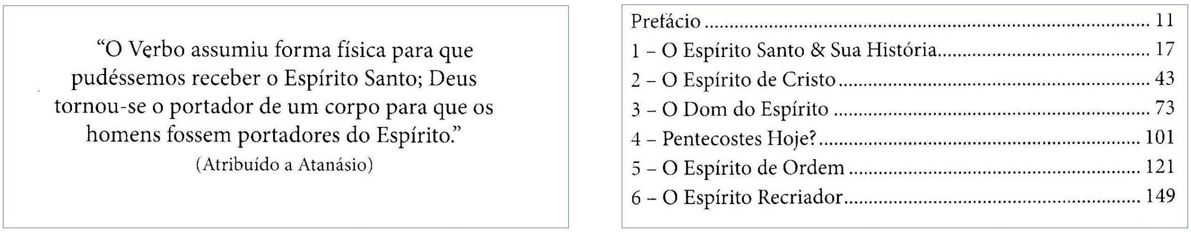
To model the unsharpness of the text from the back side when seen from the front side, a blur filter that is a kind of linear filter is applied to the reverse image, then, the image V (Back or Verse) was multiplied by the factor $(1-\alpha)$, where the index α represents the opacity coefficient of the image that assumes values from 0 to 100%. Otherwise, $(1-\alpha)$ indicates the transparency coefficient. The new components of this image are given by:

$$S'_{\alpha} = F \bigoplus (1-\alpha)V, \tag{3.2.1}$$

The resulting image shown in Figure 3.12 contains an example of image of document with back-to-front interference for $\alpha = 0.5$.

Step 3 - Adding paper texture

The texture of the paper has a strong influence the performance of binarization algorithms. Thus, it is of paramount importance to get a set of paper textures that are representative of the universe of documents intended to be modeled, from late 19th century to today, which will be used in the assessment of binarization algorithms. To do so 160 document images were used, of which 140 were from Nabuco bequest and the other 20 were obtained from five years of the LiveMemory Project, which generated a digital library of all proceeding the SBrT - Brazilian Telecommunications Symposium. The images were automatically scanned looking for a window of 141 × 114 pixels. The automatic window selection was human checked to guarantee that the area has no ink or other sort of noises. The 160 textures were statistically analyzed to get a set of paper textures representative of the whole universe of documents. For each texture sample a vector of features was built taking into account each RGB-channel of the sample, the image average filtered (R+G+B)/3, and its grayscale equivalent. For each of those 16 images the



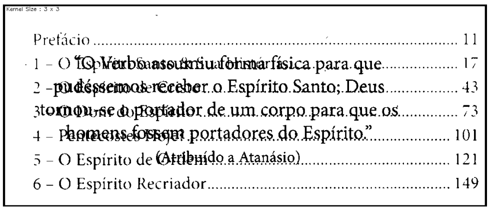
"O Verbo assumiu forma física para que
pudéssemos receber o Espírito Santo; Deus
tornou-se o portador de um corpo para que os
homens fossem portadores do Espírito."
(Atribuído a Atanásio)

Prefácio 11
1 – O Espírito Santo & Sua História 17
2 – O Espírito de Cristo 43
3 – O Dom do Espírito 73
4 – Pentecostes Hoje? 101
5 – O Espírito de Ordem 121
6 – O Espírito Recriador 149

(a) Front Image - F

(b) Back Image - V

Figure 3.11: Original images without back-to-front interference.



Kernel Size : 3 x 3

Prefácio 11
1 – O [illegible] 17
2 – [illegible] 43
3 – [illegible] 73
4 – Pentecostes Hoje? 101
5 – O Espírito de Ordem 121
6 – O Espírito Recriador 149

"O Verbo assumiu forma física para que
pudéssemos receber o Espírito Santo; Deus
tornou-se o portador de um corpo para que os
homens fossem portadores do Espírito."
(Atribuído a Atanásio)

Figure 3.12: Image of a synthetic document with back-to-front interference.

following 6 statistic measures were taken and placed in a vector:

- Mean;
- Standard deviation;
- Mode;
- Minimum value;
- Maximum value;
- Median;

Figure 3.13 presents a texture block from document in Figure 3.1.

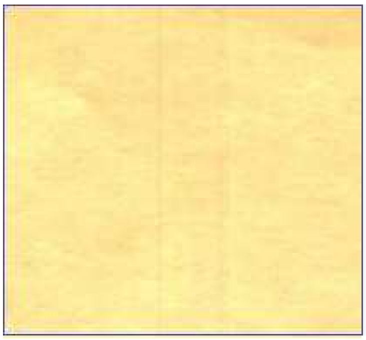


Figure 3.13: Sample texture of historical document.

Table 3.1 shows the results of this evaluation, the value of the mean and standard deviation of the RGB components which contain all the information needed to create a synthetic image of the texture of the paper. Mello & Lins (2002) used the similar parameters to generate color synthetic images in a compression scheme. These values feed the blur filter kernel for aging of the paper in the block diagram as shown in Figure 3.10. The grayscale intensity is calculated by Equation 3.1.1. The

Table 3.1: Texture information of historical document.

Label	Mean	Standard Deviation	Mode	Min	Max
Red	252.740	2.506	255	226	255
Green	233.209	6.252	234	193	248
Blue	153.654	4.907	155	126	167
GrayScale Intensity	229.99	4.65	231	196	241

paper background image generated after applied the aging information for presented in the Table 3.1, it is shown in Figure 3.14.

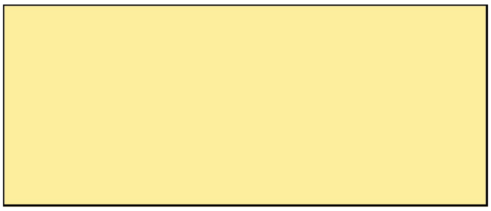
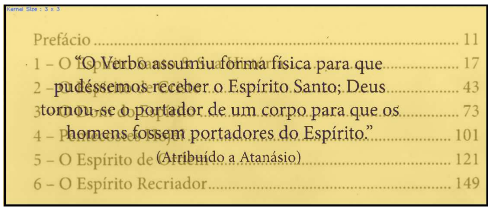


Figure 3.14: Aging mask generated by image synthesizer.

Step 4 - Generating synthetic images.

Finally, a "darker operation" is performed between the S' and A_g images, generating the final image S_α, such operation is done pixel by pixel, comparing the luminance of the correspondent pixels on both images, as represented in Figure 3.15. Thus, two



Kernel Size : 3 x 3

Prefácio 11
1 – O [illegible] 17
2 – [illegible] 43
3 [illegible] 73
4 – P[illegible] 101
5 – O Espírito de [illegible] 121
6 – O Espírito Recriador 149

"O Verbo assumiu forma física para que
pudéssemos receber o Espírito Santo; Deus
tornou-se o portador de um corpo para que os
homens fossem portadores do Espírito."
(Atribuído a Atanásio)

Figure 3.15: Synthesized image with back-to-front interference generated by the scheme in Figure 3.10.

images of documents of different nature (typed, handwritten with different pens, printed, etc.) are taken: F - front and V - verso (back). One of them is "blurred" by passing though Gaussian filters that simulate the low-pass effect of the translucidity of the verso as seen in the front part of the paper. The "blurred" verso image is now faded with a coefficient α varying between 0 and 1 in steps of 0.1. The two images are overlapped by performing a "darker" operation pixel-by-pixel in the images.

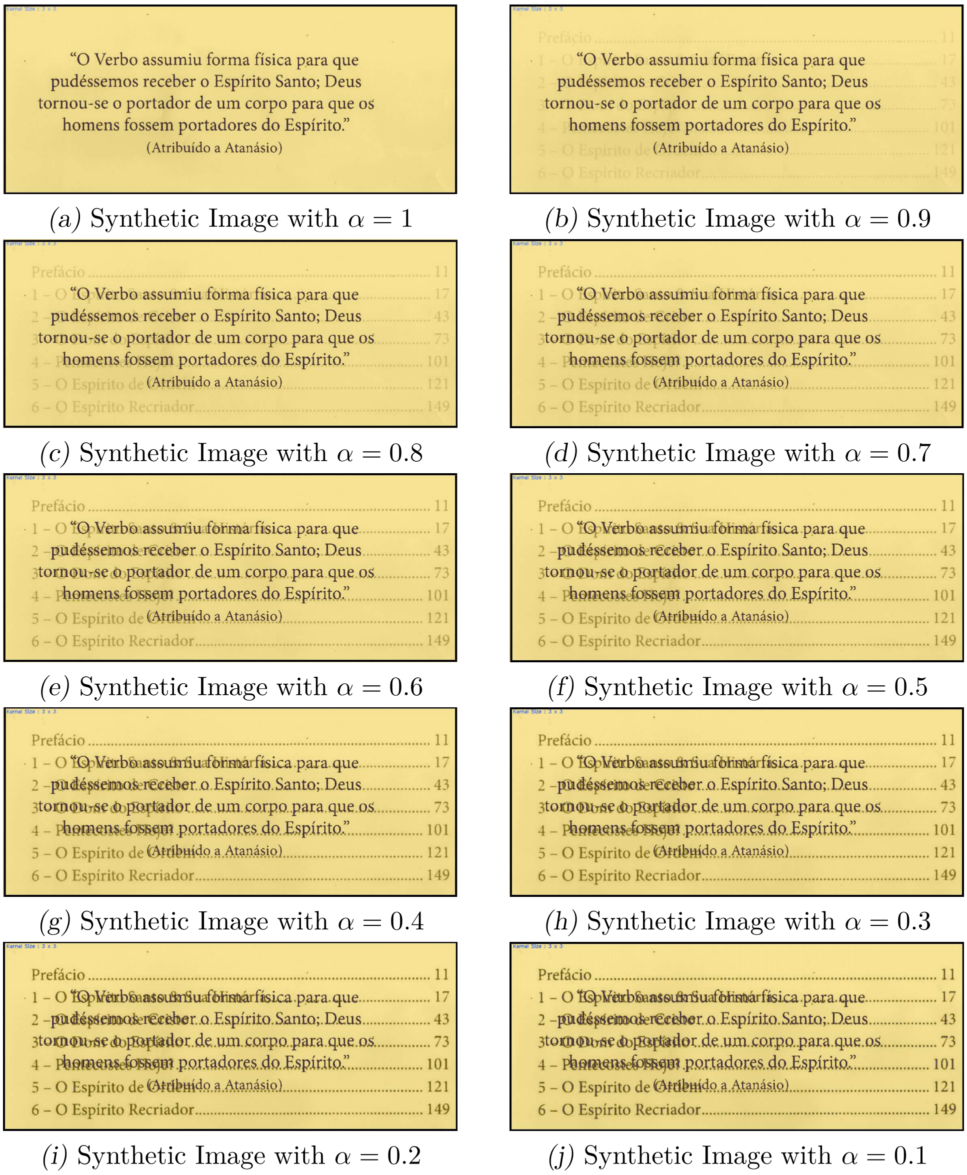
Paper texture is added to the image to simulate the effect of document aging. The steps in the generation of the synthetic images are explained next. It is important to remark that the two major concerns here: the first one is to have ground-truth images to be able to assess the performance of the several different binarization algorithms, the second one is to be able to have a very large set of synthetic images that will be used to train a classifier that will be able to automatically match a "real-world" image with the synthetic one.

3.2.1 Removing the Back-to-Front Interference

This section presents how test images with back-to-front interference were generated. The basic idea is to introduce such interference in a well controlled way, thus one is able to really know which pixels ought to be removed and which should not be removed in the filtered image. The mismatching pixels from the reference and filtered images were used to calculate the quality factors of the algorithm, allowing a fair comparison between the results obtained.

Ten documents similar to the one shown in Figure 3.16 were created using the scheme in Figure 3.10 with the parameters found in the historical file of letters by Joaquim Nabuco. The opacity coefficient α, was varied between 0.1 to 1 in steps by 0.1, representing different strength of the back-to-front interference in the document.

The blur - defocusing was modeled as a Gaussian filter, representing the point spread function, with standard error of blur in units of output pixel size. The blur size was chosen, in pixel, as 3×3, simulating a realistic aging effect in a document.



(a) Synthetic Image with $\alpha = 1$ (b) Synthetic Image with $\alpha = 0.9$

(c) Synthetic Image with $\alpha = 0.8$ (d) Synthetic Image with $\alpha = 0.7$

(e) Synthetic Image with $\alpha = 0.6$ (f) Synthetic Image with $\alpha = 0.5$

(g) Synthetic Image with $\alpha = 0.4$ (h) Synthetic Image with $\alpha = 0.3$

(i) Synthetic Image with $\alpha = 0.2$ (j) Synthetic Image with $\alpha = 0.1$

Figure 3.16: Synthetic images generated.

The 160 generated images have 1120 × 466 pixels and resolution of 300 dpi and 24 bits RGB. The mean and standard deviation values for the paper background texture are shown in Table 3.1. The block diagram shown in Figure 3.10 was implemented and simulated in C^{++} using Visual Studio Platform with “OpenCV” libraries. The program developed in C^{++} is presented in Appendix C.

An example of the generated synthetic images is shown in Figure 3.16.

3.2.2 Co-Occurrence Matrices as a Measure of Quality

This Section analyzes various types of filtering techniques to remove the back-to-front interference, after the images shown in Figure 3.16. The techniques used were: IsoData, Pun, Kapoo-Sahoo-Wong, Johannsen-Bille, Yen-Chang-Chang, Otsu, Mello-Lins, Roe-Mello, Silva-Lins Rocha and Wu-Lu methods, where the filter input synthesized images by the process described in Figure 3.10. The output image was binarized and the threshold was calculated from each algorithm. The measure to evaluate the quality of the binarized images makes use of the matrix of co-occurrence probabilities.

In the binarization process after filtering, the co-occurrence probabilities of the original image and of the binary image were calculated. More specifically, the four quadrants of the co-occurrence matrix were considered, as illustrated in Figure 3.17. The first quadrant stands for the number of pixels background-to-background (b, b), and the correct mapping of background pixels from the original image that correspond to background pixels in the filtered one. Similarly, the third quadrant stands for the foreground-to-foreground (f, f) correct mappings. The second and fourth quadrants denote, respectively, the background-to-foreground (b, f) and the foreground-to-background (f, b) uncorrect transformations. Using the co-occurrence probabilities $p_{i,j}$, that is, the score of i to j gray-level transitions normalized by the total number of transitions, the quadrant probabilities were:

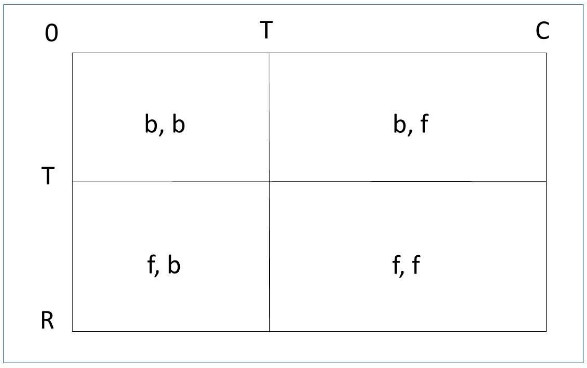


Figure 3.17: Co-occurrence matrix.

$$P_{bb}(T) = \sum_{i=0}^{T}\sum_{j=0}^{T} p_{ij}. \tag{3.2.2}$$

$$P_{bf}(T) = \sum_{i=0}^{T} \sum_{j=T+1}^{C} p_{ij}. \tag{3.2.3}$$

$$P_{fb}(T) = \sum_{i=T+1}^{R} \sum_{j=0}^{T} p_{ij}. \tag{3.2.4}$$

$$P_{ff}(T) = \sum_{i=T+1}^{R} \sum_{j=T+1}^{C} p_{ij}. \tag{3.2.5}$$

In conclusion, the performance underwent a comparison among the various types of filters.

The procedure performed on filter analysis is shown in Figure 3.18, and was implemented and simulated using ImageJ plugins. The several synthetic images went through

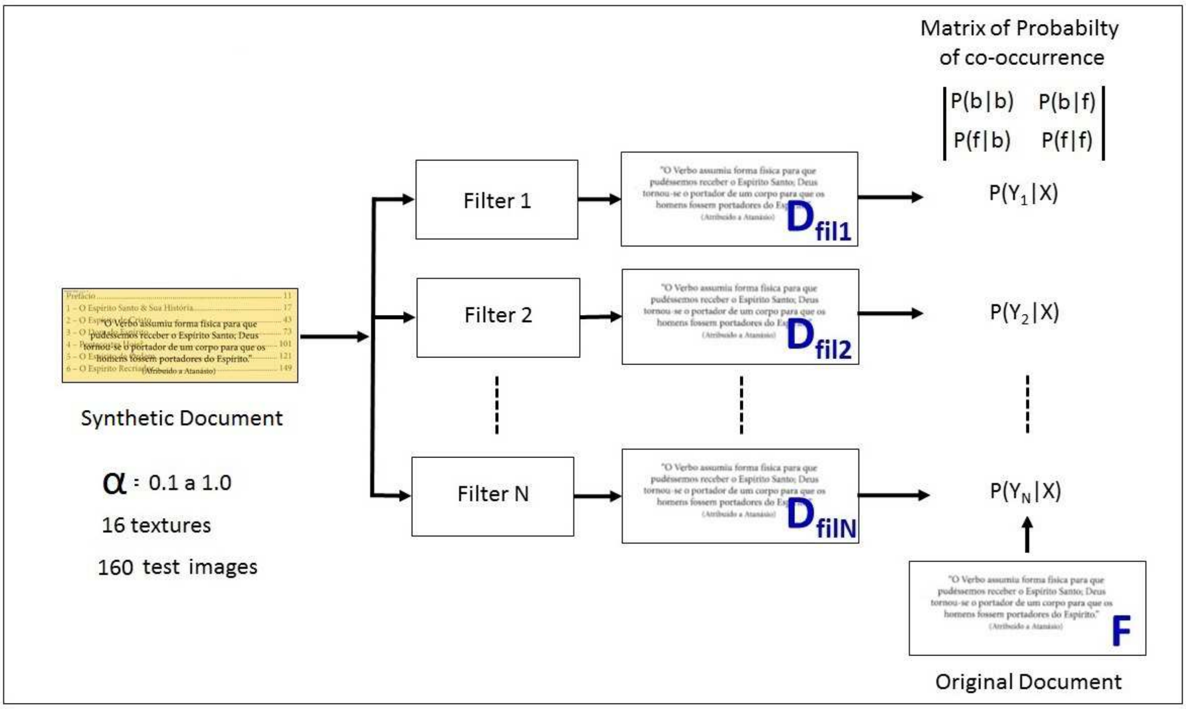


Figure 3.18: *Procedures for the analysis of filters.*

the different binarization filters, and the output image of each filter. Then, the matrix of co-occurrence probability - $P(Y|X)$, where X was the input and Y the output images, is calculated as represented in Equation 3.2.6.

$$P(Y|X) = \begin{array}{c} \\ b \\ f \end{array} \begin{array}{c} \begin{array}{cc} b & f \end{array} \\ \begin{pmatrix} P(b|b) & P(b|f) \\ P(f|b) & P(f|f) \end{pmatrix} \end{array} \tag{3.2.6}$$

3.3 Thresholding Algorithms

Thresholding algorithms can be classified into *global* or *local* algorithms. Global algorithms define a unique threshold value for the complete image. Local algorithms first split the image into regions according to some criterion and then define threshold values for each region. In general, global algorithms are faster than local algorithms, although, in general, local algorithms may provide better results. Otsu is the most widely used global thresholding algorithm. Otsu's algorithm is adaptive and requires no adjustment setting. It considers that there are two classes, separated by a threshold value. The inter and intra class variances are evaluated and the final threshold value is the one that maximizes the relation between them.

Global thresholding based on clustering (Otsu, 1979), entropy minimization (Kapur *et al.*, 1985), and valley seeking in the intensity histogram (Weszka, 1979) as well as feature-based (Dawoud & Kamel, 2004) and model-based (Liu & Srihari, 1997) methods have been presented in the literature. Such methods are efficient for images in which the gray levels of the text and background pixels are separable. If the histogram of the text overlaps with the one of the background, the result are poor quality binary images, (Valizadeh & Kabir, 2012).

Sezgin & Sankur (2004) presented a comprehensive overview and comparison of thresholding algorithms, clustering them according to their nature. From the almost forty algorithms presented six schemes were deemed to be suitable to work in documents with back-to-front interference such as the ones studied here: Pun (1981), Kapur *et al.* (1985), Johannsen & Bille (1982), Yen (1995), Wu (1998), and Otsu (1979). The Roe & Mello (2013) algorithm used a local image equalization based on color constancy, and an extension to the standard difference of Gaussians edge detection operator, XDoG and Otsu binarization algorithm. The first five algorithms are based on the entropy of the image, whereas the last one uses discriminator analysis. Iterative Self Organizing Data Analysis Technique - IsoData was added. It is a method of unsupervised classification, and the computer runs the algorithm through many iterations until threshold is reached; all algorithms work with thresholding techniques. Appendix *A* outlines how such algorithms work.

Besides those algorithms, two algorithms based on Shannon's entropy, that were created in the scope of the Nabuco Project to filter that interference were also accessed: Mello & Lins (2002), Mello & Lins (2000) and Silva *et al.* (2006).

The Appendix B of this thesis presents an updated annotated bibliography of the algorithms found in the literature for document image binarization techniques, including documents with back-to-front interference.

3.3.1 The IsoData Method

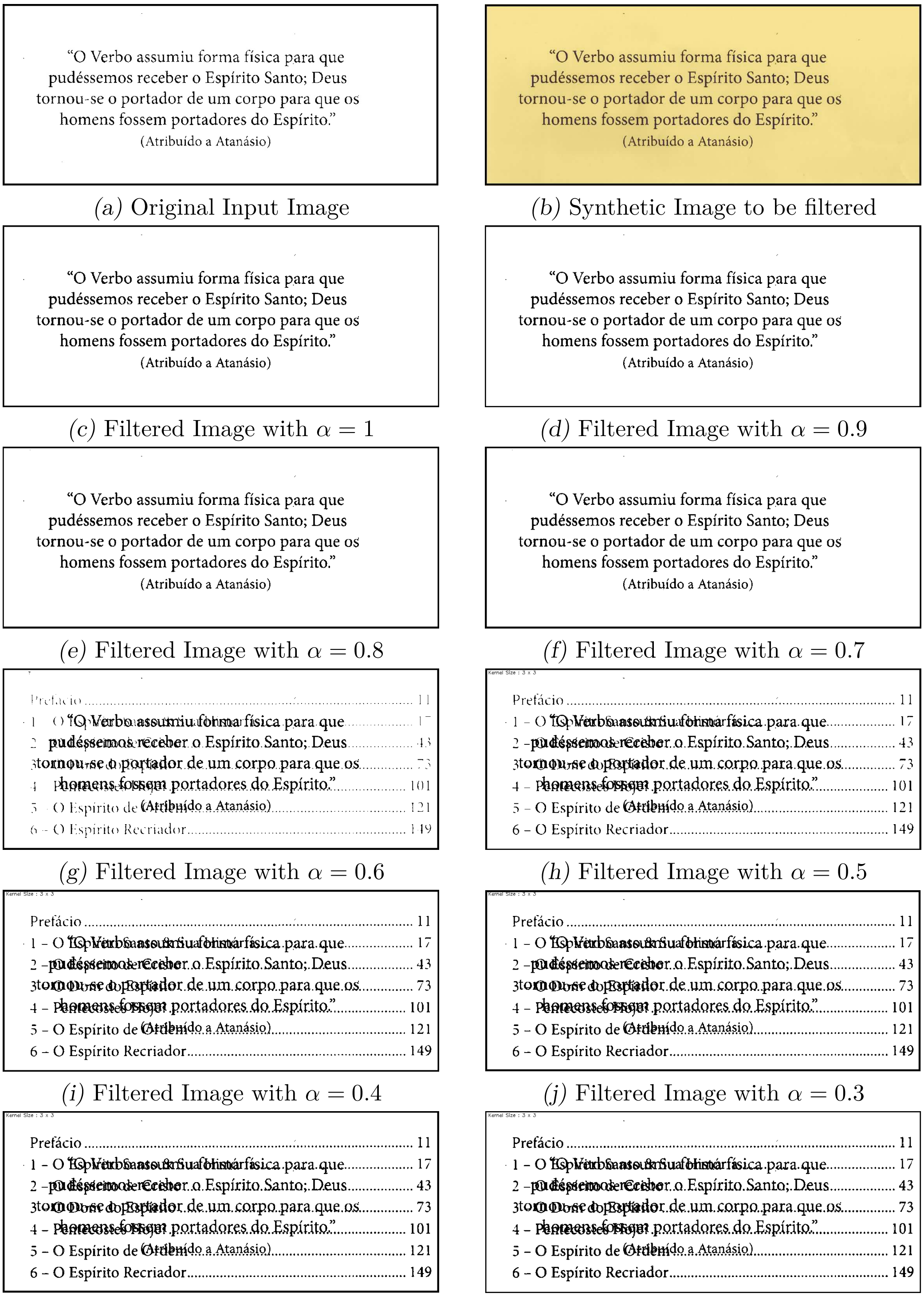
Clustering is an unsupervised classification as no a priori knowledge (such as samples of known classes) is assumed to be available. The ISODATA Algorithm (Iterative Self-Organizing Data Analysis Technique Algorithm) allows the number of clusters to be adjusted automatically during the iteration by merging similar clusters and splitting clusters with large standard deviations, according to reference (Memarsadeghi, 2007) and (Ball & Hall, 1965). The algorithm is highly heuristic. In the case of using the IsoData algorithm for binarizing document images the pixels in the image are iteratively sent to two clusters which will correspond to the black and white pixels.

Unsupervised clustering is a fundamental tool in image processing for geoscience and remote sensing applications. For example, unsupervised clustering is often used to obtain vegetation maps of an area of interest according to reference (Memarsadeghi, 2007).

The most common unsupervised image classification are Isodata, Support Vector Machine (SVM) and k-Means methods, according to (Abburu & Golla, 2015).

Figure 3.19 (a) shows a standard image without front-to-back interference, which will be compared to the binarized image at the filter output, whereas Figure 3.19 (b) shows the colored synthetic image to be filtered.

The result of applying IsoData Filter using the document images of Figure 3.16 is shown in Figure 3.19. Table 3.2 presents the result of the binarization of the test set images using the IsoData algorithm.



(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.19: ISODATA filtering of the images in Figure 3.16.

Table 3.2 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

Table 3.2: IsoData Filter Result.

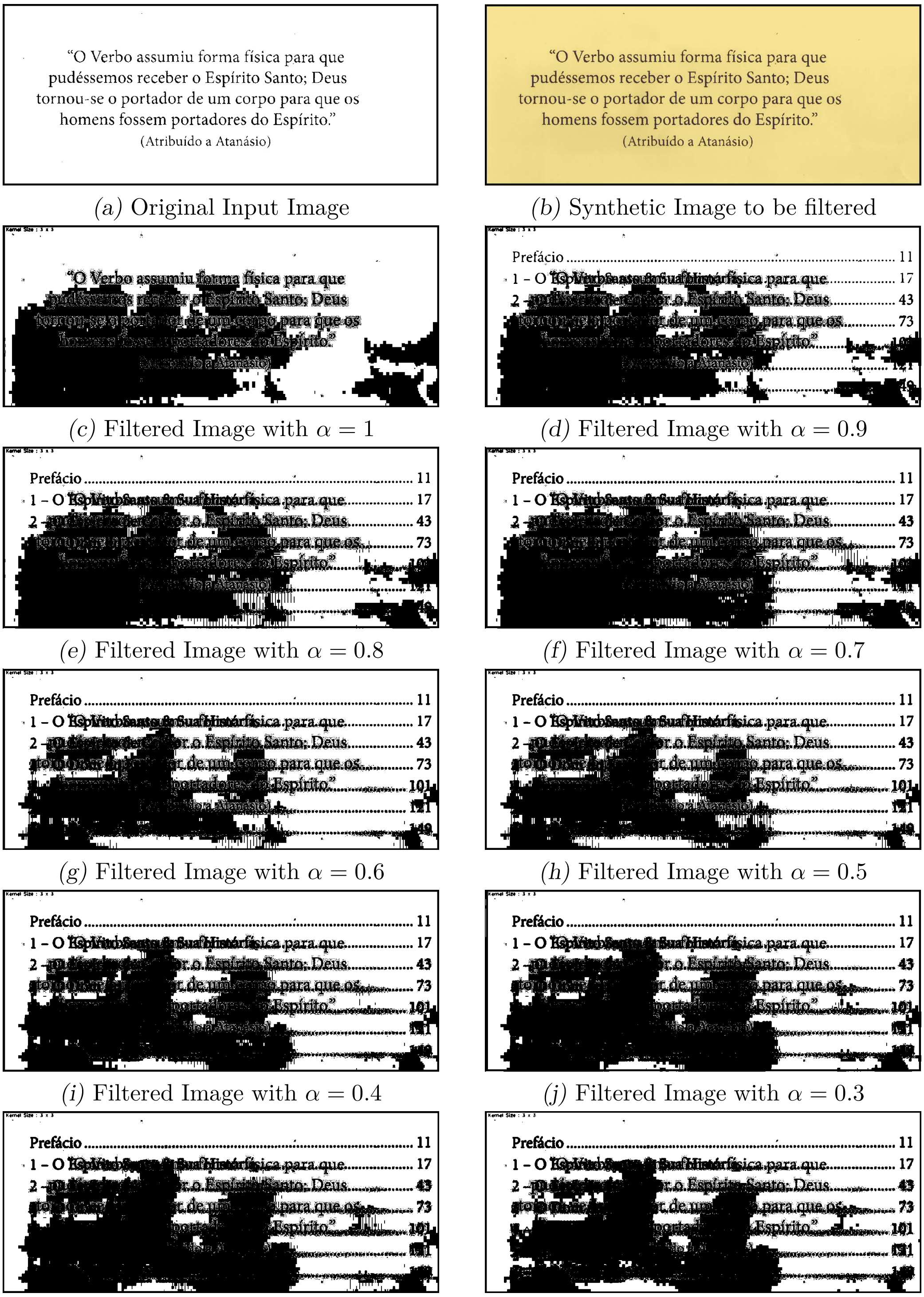
α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	142	94.49%	5.51%	99.52%	0.48%
0.2	1x1	153	233	252	3x3	142	94.84%	5.16%	99.53%	0.47%
0.3	1x1	153	233	252	3x3	144	95.14%	4.86%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	146	95.54%	4.46%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	147	96.22%	3.78%	100.00%	0.00%
0.6	1x1	153	233	252	3x3	144	97.85%	2.15%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	136	99.89%	0.11%	98.87%	1.13%
0.8	1x1	153	233	252	3x3	137	99.94%	0.06%	99.23%	0.77%
0.9	1x1	153	233	252	3x3	137	99.98%	0.02%	99.20%	0.80%
1.0	1x1	153	233	252	3x3	138	100.00%	0.00%	99.56%	0.44%

Analyzing the quality of the binarized images produced by the Isodata filter, it seems reasonable to consider important features for removing back-to-front interference: where the interference fade varied between $0.7 \leq \alpha \leq 1.0$, the value of the background-background mapping yielded an error of less than 0.11% as $99.89\% \leq P(b|b) \leq 100.00\%$. The foreground to foreground matching percentage P(f|f) had a small variation between 99.56% and 98.87%, a error less than 1.13%. It is interesting to notice that for very weak back-to-front interference ($\alpha = 0.1, \alpha = 0.2$) over 5% of the pixels from the paper texture were mapped onto the foreground, degrading the quality of the image. The filtering threshold varied between 136 and 147.

3.3.2 Pun Method

The algorithm proposed by Pun (1981) takes as input a gray levels image considered as produced by a source with an alphabet consisting of 256 statistically independent symbols. Pun considers the ratio between the a posteriori entropy and the total entropy as the image threshold. Table 6 presents the results of applying Pun's algorithm to the gray-level version of the synthetic images in the test set.

Figure 3.20 presents the result of applying Pun's algorithm from the generated synthetic images presented in Figure 3.16.


"O Verbo assumiu forma física para que
pudéssemos receber o Espírito Santo; Deus
tornou-se o portador de um corpo para que os
homens fossem portadores do Espírito."
(Atribuído a Atanásio)
Prefácio 11
1 – O 17
2 43
73
101
121
149

(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.20: Pun filtering of the images in Figure 3.16.

Table 3.3 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

Pun algorithm is not suitable for the binarization of the test set of images although the P(f|f) was of 100.00% for all alphas, the P(b|b) was around 60%, reaching 55.51% for

Table 3.3: Pun Filter Result.

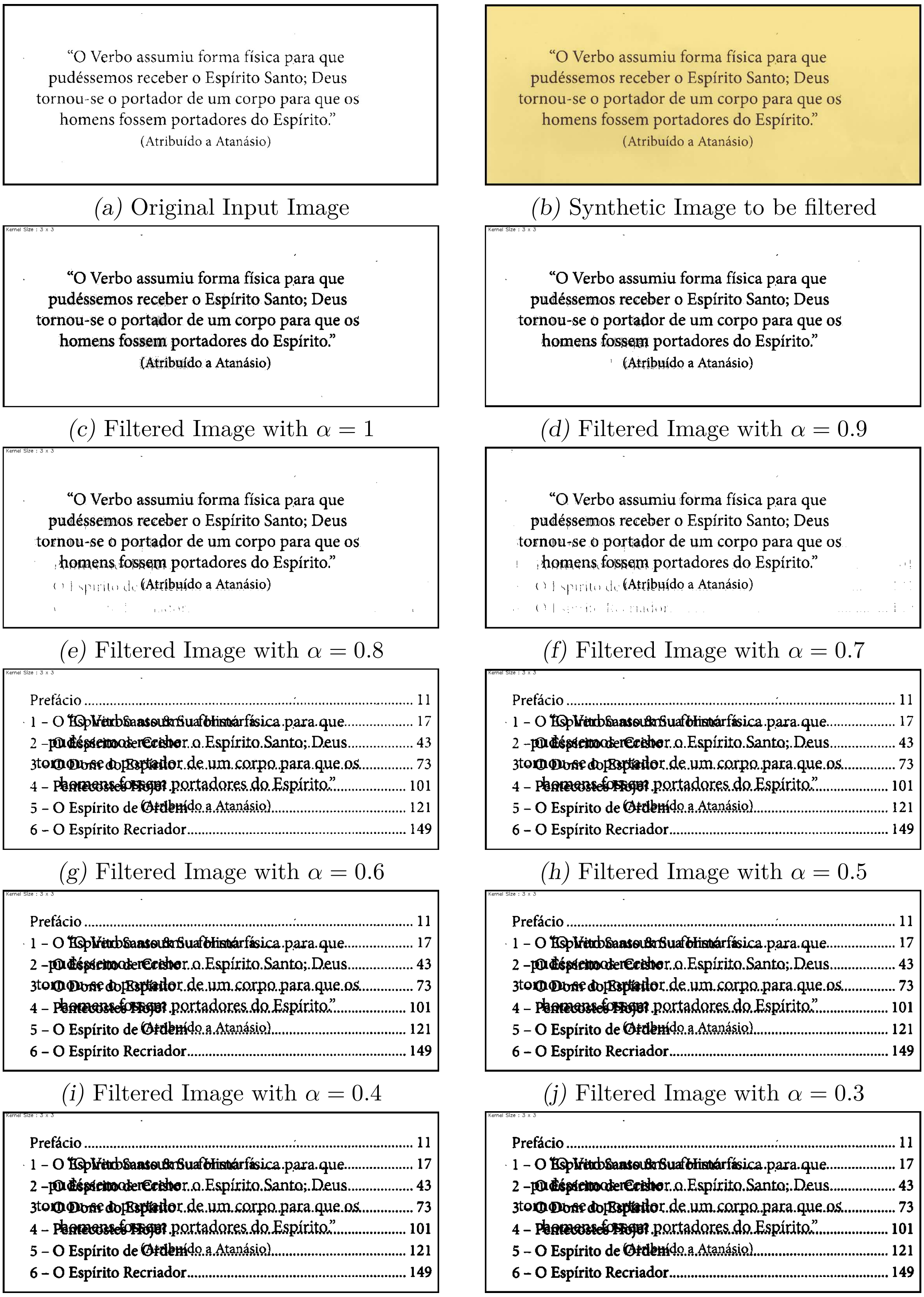
(α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	195	61.99%	38.01%	100.00%	0.00%
0.2	1x1	153	233	252	3x3	196	57.97%	42.03%	100.00%	0.00%
0.3	1x1	153	233	252	3x3	196	59.15%	40.85%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	196	61.64%	38.36%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	196	65.20%	34.80%	100.00%	0.00%
0.6	1x1	153	233	252	3x3	196	67.16%	32.84%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	198	55.51%	44.49%	100.00%	0.00%
0.8	1x1	153	233	252	3x3	198	58.39%	41.61%	100.00%	0.00%
0.9	1x1	153	233	252	3x3	198	60.52%	39.48%	100.00%	0.00%
1.0	1x1	153	233	252	3x3	199	59.76%	40.24%	100.00%	0.00%

$\alpha = 0.7$, meaning that are large number of background pixels were mapped onto black pixels of the monochromatic image.

3.3.3 Kapur-Sahoo-Wong Filter

The algorithm by Kapur *et al.* (1985) considers the foreground and background images as two distinct sources, such that whenever the addition of the two entropies reach a maximum, its argument t reaches the optimal value.

Figure 3.21 shows the result of applying Kapur-Sahoo-Wong Filter from the synthetically generated images showed in Figure 3.16.



(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.21: Kapur-Sahoo-Wong filtering of the images in Figure 3.16.

Table 3.4 presents an analysis of the evolution of the different opacity coefficient (α values) versus the matrix of co-occurrence probability.

The analysis of the data in Table 3.4 reveals that there was the partial elimination the back-to-front interference, for $0.7 \leq \alpha \leq 1.0$ as the value of background-background

Table 3.4: Kapur-Sahoo-Wong Filter Result.

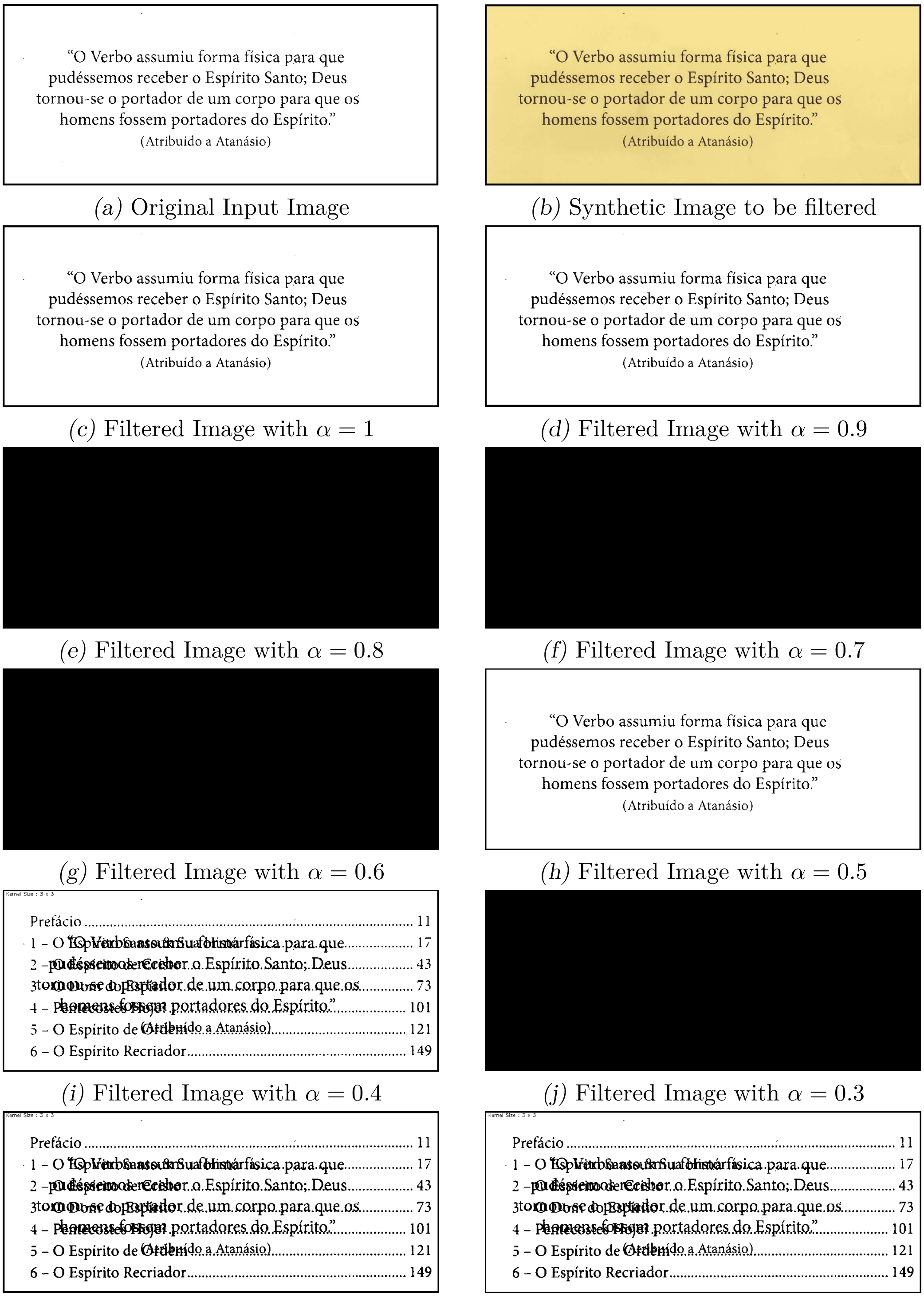
α	kernel-Gaussian	Red	Green	Blue	kernel-Blur	Threshold	P(b|b)	P(b|f)	P(f|f)	P(f|b)
0.1	1x1	153	233	252	3x3	176	90.88%	9.12%	100.00%	0.00%
0.2	1x1	153	233	252	3x3	174	91.50%	8.50%	100.00%	0.00%
0.3	1x1	153	233	252	3x3	174	91.86%	8.15%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	174	92.29%	7.71%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	173	92.98%	7.02%	100.00%	0.00%
0.6	1x1	153	233	252	3x3	174	93.49%	6.51%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	147	99.25%	0.75%	100.00%	0.00%
0.8	1x1	153	233	252	3x3	162	98.87%	1.13%	100.00%	0.00%
0.9	1x1	153	233	252	3x3	175	98.59%	1.41%	100.00%	0.00%
1.0	1x1	153	233	252	3x3	182	98.36%	1.64%	100.00%	0.00%

probability P(b|b) varied between 99.25% and 98.36%, an error less than 1.64%, considering that the foreground-foreground matching percentage P(b|b) was of 100.00%. Table 3.4 clearly shows that this algorithm reaches the best performance for the image with $\alpha = 0.7$, with a P(b|f) of 0.75%.

3.3.4 Johannsen-Bille Method

This method Johannsen & Bille (1982) uses the entropy of the gray level histogram of the digital image. Essentially, it divides the set of gray into two parts, to minimize the interdependence between them. Table 3.5 presents the performance obtained by this filter for the test set.

Figure 3.22 shows the result of applying Johannsen-Bille method from the synthetically generated images showed in Figure 3.16.


"O Verbo assumiu forma física para que
pudéssemos receber o Espírito Santo; Deus
tornou-se o portador de um corpo para que os
homens fossem portadores do Espírito."
(Atribuído a Atanásio)
Kernel Size : 3 x 3
Prefácio 11
1 – O Espírito Santo e Sua história 17
2 – O Espírito de Cristo 43
3 – O Dom do Espírito 73
4 – Pentecostes Hoje 101
5 – O Espírito de Ordem 121
6 – O Espírito Recriador 149

(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.22: Johannsen-Bille filtering of the images in Figure 3.16.

Table 3.5 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

The results shown demonstrate that the Johanssen-Bille filter is very unstable depending on the opacity coefficient α, as when its values were 0.3, 0.6, 0.7, and 0.8 the output

Table 3.5: Johannsen-Bille Filter Result.

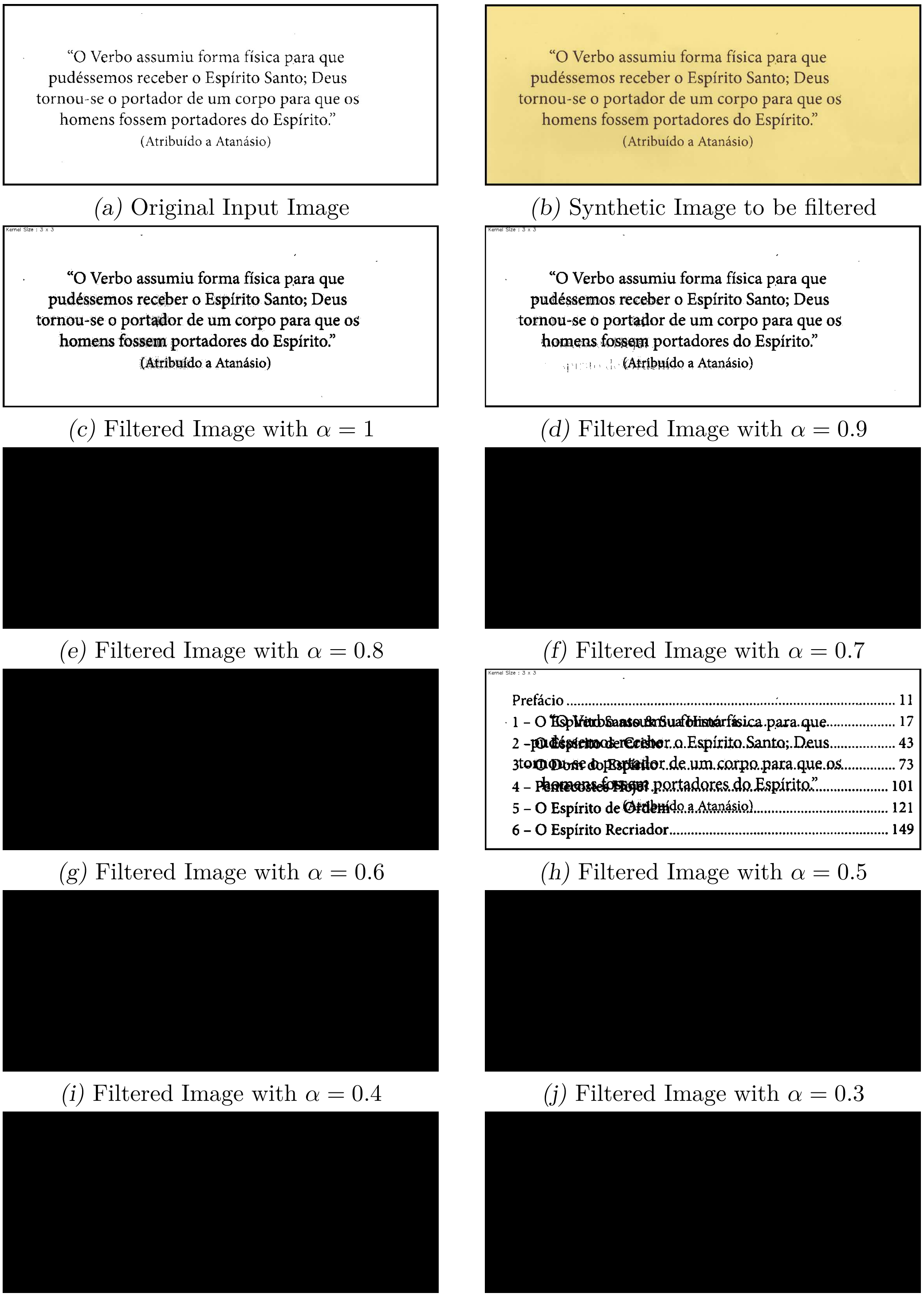
α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	142	94.49%	5.51%	99.52%	0.48%
0.2	1x1	153	233	252	3x3	149	94.23%	5.77%	100.00%	0.00%
0.3	1x1	153	233	252	3x3	210	0.00%	100.00%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	150	95.15%	4.85%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	100	99.97%	0.03%	84.63%	15.37%
0.6	1x1	153	233	252	3x3	211	0.00%	100.00%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	211	0.00%	100.00%	100.00%	0.00%
0.8	1x1	153	233	252	3x3	211	0.00%	100.00%	100.00%	0.00%
0.9	1x1	153	233	252	3x3	112	100.00%	0.00%	88.39%	11.61%
1.0	1x1	153	233	252	3x3	112	100.00%	0.00%	88.11%	11.89%

was completely black images. The Johanssen-Bille algorithm presented in some of the cases ($\alpha = 0.5, 0.9, 1.0$) an information loss, as over 10% of the foreground pixels were mapped onto background ones. The value of background-background probability $P(b|b)$ varied between 100.00% and 99.97%, while the value of foreground-foreground probability changed between 88.11% and 84.63%, offering a loss in the quality of the text.

3.3.5 Yen-Chang-Chang Method

The binarization algorithm by Yen (1995) follows the same ideas as the one by Kapur *et al.* (1985) in respect to the entropy distributions. The result of applying Yen-Chang-Chang Method to the test set of document images is showed in Table 3.6.

The result of applying Yen-Chang-Chang Method to the document image of Figure 3.16 is showed in Figure 3.23.



(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.23: Yen-Chang-Chang filtering of the images in Figure 3.16.

Table 3.6 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

Figure 3.23 (e), (f), (g), (i), (j), (k) and (l) which corresponds to an $\alpha = 0.8, 0.7, 0.6, 0.4, 0.3, 0.2$ and 0.1 value, respectively, the inadequacy of applying the Yen-Chang-Chang filter

Table 3.6: Yen-Chang-Chang Filter Result.

α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	210	0.00%	100.00%	100.00%	0.00%
0.2	1x1	153	233	252	3x3	210	0.00%	100.00%	100.00%	0.00%
0.3	1x1	153	233	252	3x3	210	0.00%	100.00%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	210	0.00%	100.00%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	178	92.14%	7.86%	100.00%	0.00%
0.6	1x1	153	233	252	3x3	211	0.00%	100.00%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	211	0.00%	100.00%	100.00%	0.00%
0.8	1x1	153	233	252	3x3	211	0.00%	100.00%	100.00%	0.00%
0.9	1x1	153	233	252	3x3	176	98.47%	1.53%	100.00%	0.00%
1.0	1x1	153	233	252	3x3	183	98.23%	1.77%	100.00%	0.00%

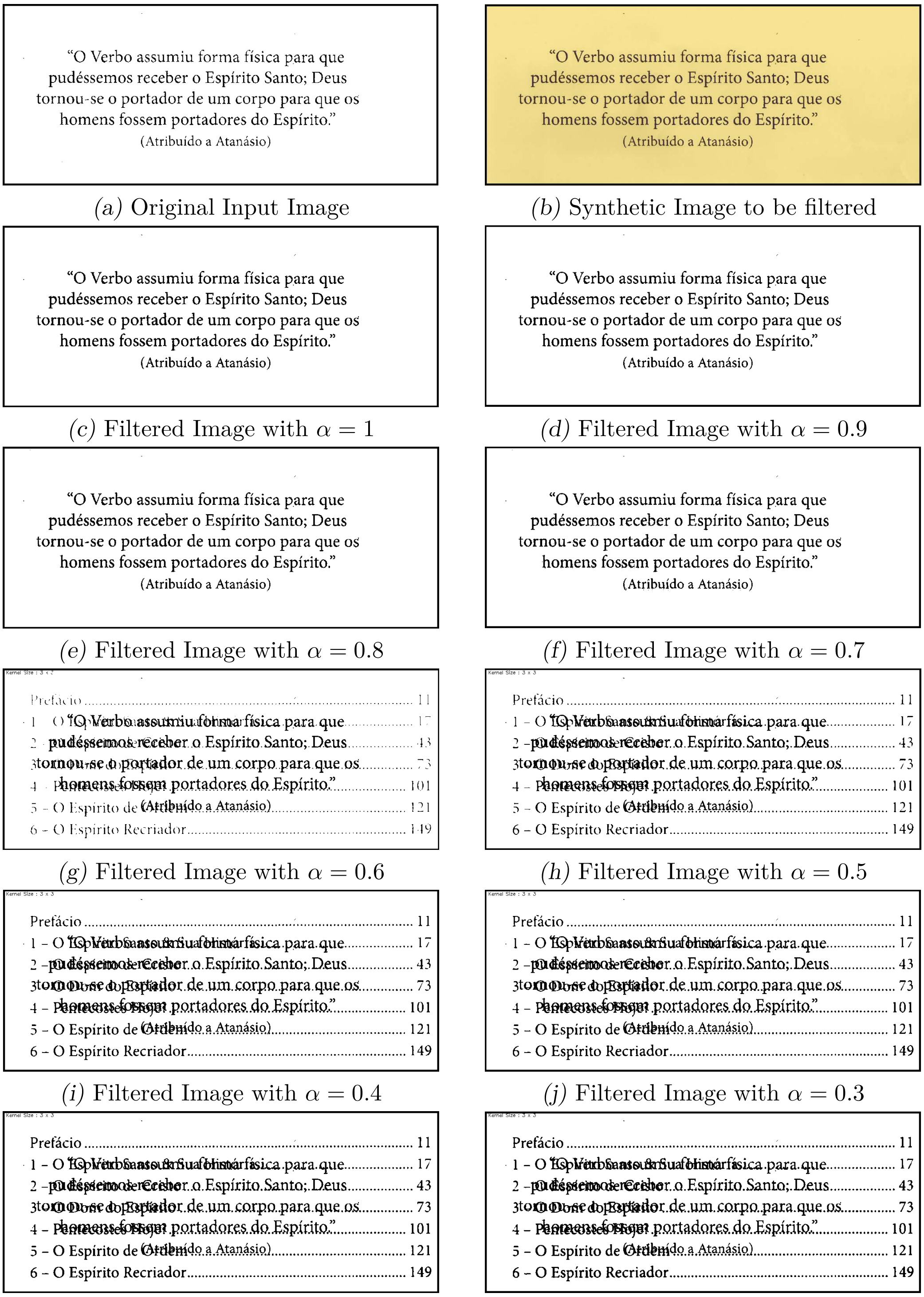
from the set of synthesised image presented in Figure 3.16, however, for the $\alpha = 1.0, 0.9$ and 0.5 values the Yen-Chang-Chang filter has a reasonable response. Only for $\alpha = 1$ and 0.9 the result of binarized image quality and the co-occurrence probability were acceptable.

The results above show that Yen-Chang-Chang algorithm is not suitable to binarize the test set images as seven out of ten images were mapped onto completely back images.

3.3.6 Otsu Threshold Method

Otsu (1979) is the most widely used global thresholding algorithm. Otsu's algorithm is adaptive and requires no adjustment setting. It considers that there are two classes, separated by a threshold value. Otsu's algorithm makes use of Sahoo discriminator analysis for defining whether a gray level t is mapped onto foreground or background information.

Figure 3.24 presents the result of the application of Otsu's algorithm to the image presented in Figure 3.16. The result of this algorithm applied to the synthetic images with different alphas is shown in Table 3.7.



(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.24: Otsu filtering of the images in Figure 3.16.

Table 3.7 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

Although Otsu algorithm was originally developed for ultrasound images, the results above show that it performs well with document images. Table 3.7 shows that

Table 3.7: Otsu Filter Result.

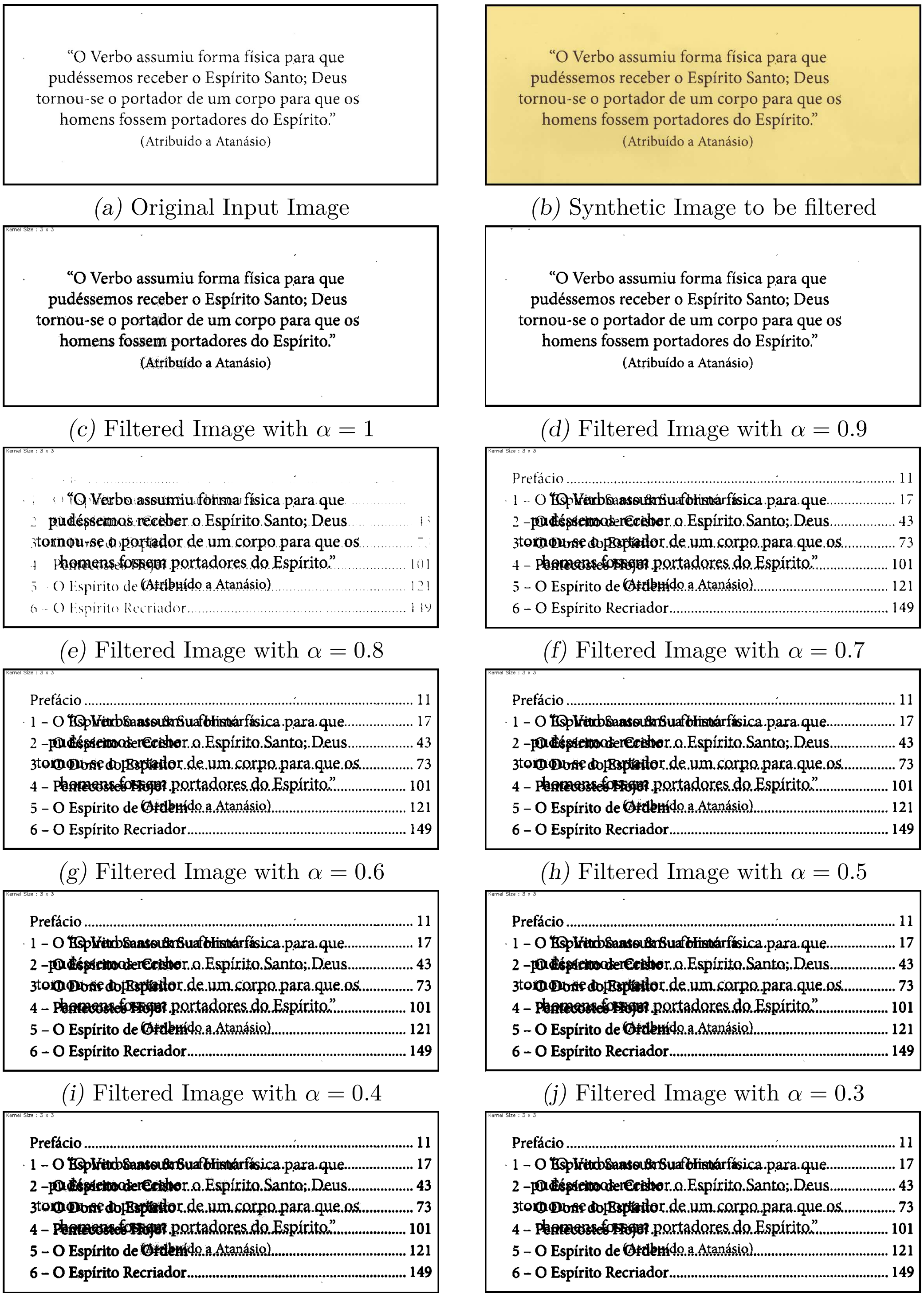
α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	145	94.19%	5.81%	100.00%	0.00%
0.2	1x1	153	233	252	3x3	145	94.57%	5.43%	100.00%	0.00%
0.3	1x1	153	233	252	3x3	145	95.05%	4.95%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	149	95.24%	4.76%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	149	96.00%	4.00%	100.00%	0.00%
0.6	1x1	153	233	252	3x3	146	97.51%	2.49%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	138	99.87%	0.13%	99.54%	0.46%
0.8	1x1	153	233	252	3x3	138	99.94%	0.06%	99.56%	0.44%
0.9	1x1	153	233	252	3x3	138	99.97%	0.03%	99.53%	0.47%
1.0	1x1	153	233	252	3x3	140	99.95%	0.05%	99.55%	0.45%

for $0.7 \leq \alpha \leq 1.0$, the value of background-background correct mapping percentage was $99.87\% \leq P(b|b) \leq 99.97\%$ yielding error less than 0.13%, while the foreground-foreground percentage $99.53\% \leq P(f|f) \leq 99.56\%$, an error less than 0.47%. Comparing the data presenting in Table 3.7 and 3.4 one may conclude that Otsu presented better results than Kapur-Sahoo-Wong filter for that specific set of images.

3.3.7 Mello-Lins Algorithm

The algorithm by Mello & Lins (2002) and Mello & Lins (2000) is based on Shannon entropy to calculate a global threshold and looks for the most frequent gray level of the image and takes it like initial threshold to evaluate the values H_b, H_w and H, shown in Appendix A. It was developed with the aim of filtering out the back-to-front interference. The results obtained for the images in the test set are presented in Table 3.8. All the pixels of the foreground in the test images were correctly mapped onto pixels of the foreground in the ground case images, as P(f|f)=100% for all values of α. The P(b|b) values were very high, reaching its best performance for $\alpha = 0.9$.

The result of applying Mello-Lins Algorithm to the document image of Figure 3.16 is showed in Figure 3.25.



(*a*) Original Input Image

(*b*) Synthetic Image to be filtered

(*c*) Filtered Image with $\alpha = 1$

(*d*) Filtered Image with $\alpha = 0.9$

(*e*) Filtered Image with $\alpha = 0.8$

(*f*) Filtered Image with $\alpha = 0.7$

(*g*) Filtered Image with $\alpha = 0.6$

(*h*) Filtered Image with $\alpha = 0.5$

(*i*) Filtered Image with $\alpha = 0.4$

(*j*) Filtered Image with $\alpha = 0.3$

(*k*) Filtered Image with $\alpha = 0.2$

(*l*) Filtered Image with $\alpha = 0.1$

Figure 3.25: Mello-Lins filtering of the images in Figure 3.16.

Table 3.8 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

The analysis of the binarized images in Figure 3.25 (c) to (f) reveals that there was the partial elimination the back-to-from interference, mainly figures (e) and (f), which corre-

Table 3.8: Mello-Lins Filter Result.

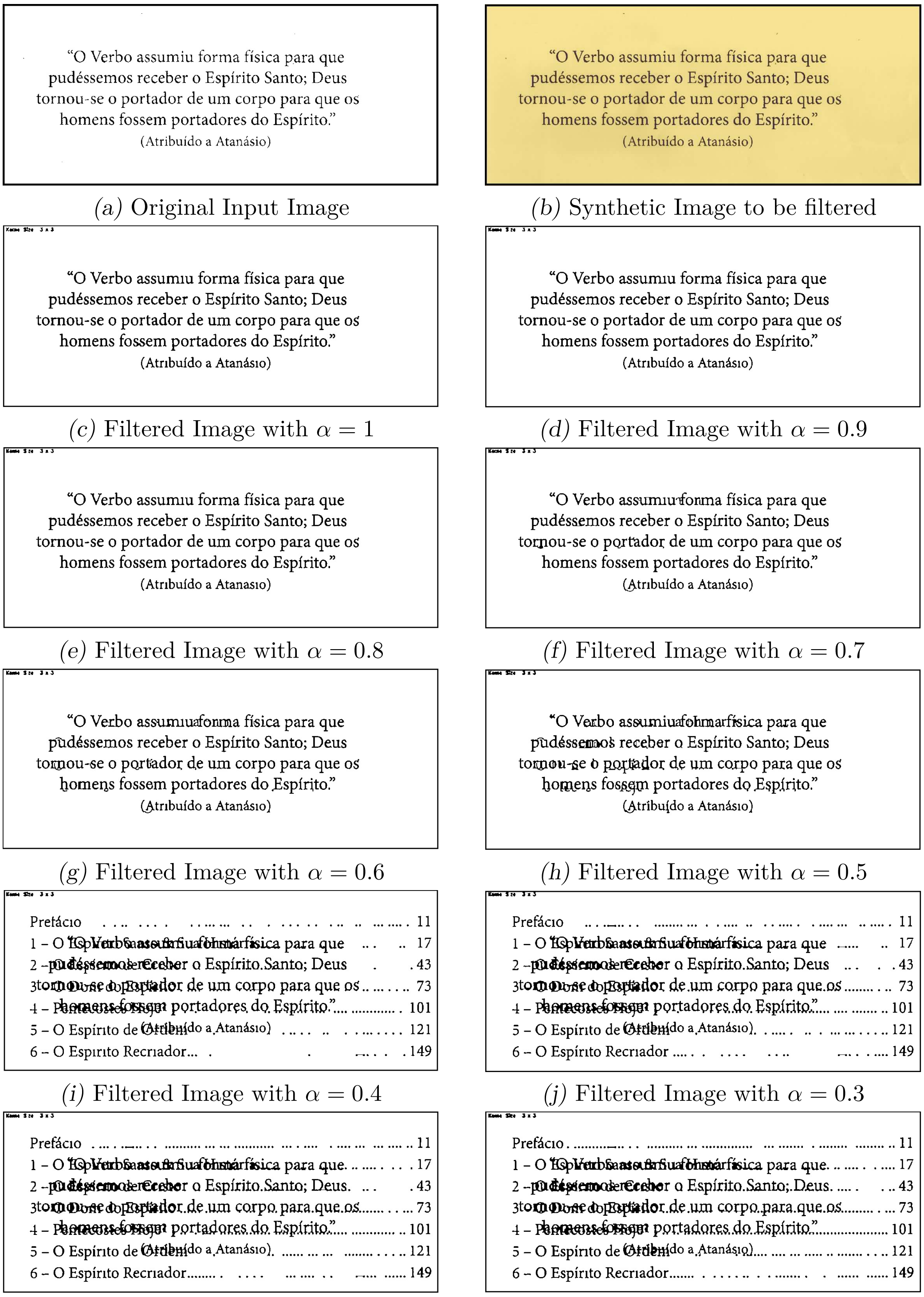
α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	174	91.19%	8.81%	100.00%	0.00%
0.2	1x1	153	233	252	3x3	183	89.76%	10.24%	100.00%	0.00%
0.3	1x1	153	233	252	3x3	181	90.58%	9.42%	100.00%	0.00%
0.4	1x1	153	233	252	3x3	180	91.21%	8.78%	100.00%	0.00%
0.5	1x1	153	233	252	3x3	178	92.14%	7.86%	100.00%	0.00%
0.6	1x1	153	233	252	3x3	176	93.14%	6.86%	100.00%	0.00%
0.7	1x1	153	233	252	3x3	174	94.47%	5.53%	100.00%	0.00%
0.8	1x1	153	233	252	3x3	170	97.30%	2.70%	100.00%	0.00%
0.9	1x1	153	233	252	3x3	165	99.19%	0.81%	100.00%	0.00%
1.0	1x1	153	233	252	3x3	181	98.45%	1.55%	100.00%	0.00%

sponds in the range where the alpha varied between 1.0 and 0.7, the value of background-background probability $P(b|b)$ varied between 94.47% and 99.19%, a error less than 5.53%, considering that the foreground-foreground probability $P(f|f)$ was of 100.00%.

3.3.8 Roe-Mello Algorithm

The Roe & Mello (2013) (also presented in Appendix A) algorithm performs a local image equalization based on color constancy, and an extension to the standard difference of Gaussian edge detection operator, XDoG and Otsu binarization algorithm. The last two algorithms assessed are based on the entropy of the image, whereas the Roe-Mello one uses discriminator analysis. The threshold used by the algorithm showed very little variation, as may be observed in Table 3.9.

The result of applying Roe-Mello Algorithm to the document image of Figure 3.16 is presented in Figure 3.26.



(*a*) Original Input Image

(*b*) Synthetic Image to be filtered

(*c*) Filtered Image with $\alpha = 1$

(*d*) Filtered Image with $\alpha = 0.9$

(*e*) Filtered Image with $\alpha = 0.8$

(*f*) Filtered Image with $\alpha = 0.7$

(*g*) Filtered Image with $\alpha = 0.6$

(*h*) Filtered Image with $\alpha = 0.5$

(*i*) Filtered Image with $\alpha = 0.4$

(*j*) Filtered Image with $\alpha = 0.3$

(*k*) Filtered Image with $\alpha = 0.2$

(*l*) Filtered Image with $\alpha = 0.1$

Figure 3.26: Roe-Mello filtering of the images in Figure 3.16.

Table 3.9 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

Analyzing the quality of the binarized images produced by the Roe-Mello filter in Figure 3.26 (*c*) to (*h*), it seems reasonable to consider important features for removing

Table 3.9: Roe-Mello Filter Result.

α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	181	88.16%	11.84%	39.39%	60.61%
0.2	1x1	153	233	252	3x3	181	88.41%	11.59%	39.10%	60.90%
0.3	1x1	153	233	252	3x3	181	88.73%	11.27%	38.11%	61.89%
0.4	1x1	153	233	252	3x3	181	89.23%	10.76%	36.45%	63.55%
0.5	1x1	153	233	252	3x3	181	94.84%	5.16%	23.70%	76.30%
0.6	1x1	153	233	252	3x3	181	95.41%	4.59%	22.46%	77.54%
0.7	1x1	153	233	252	3x3	181	95.55%	4.45%	22.10%	77.90%
0.8	1x1	153	233	252	3x3	181	95.63%	4.37%	22.04%	77.96%
0.9	1x1	153	233	252	3x3	181	95.63%	4.37%	22.03%	77.97%
1.0	1x1	153	233	252	3x3	181	98.58%	4.42%	22.13%	77.87%

back-to-front interference, beside the Figure 3.26 (h) presents a confusing text, in the range, where the α varied between 1.0 to 0.5, the value of background-background probability $P(b|b)$ varied between 98.50% and 94.84%, respectively, a error less than 5.16%.

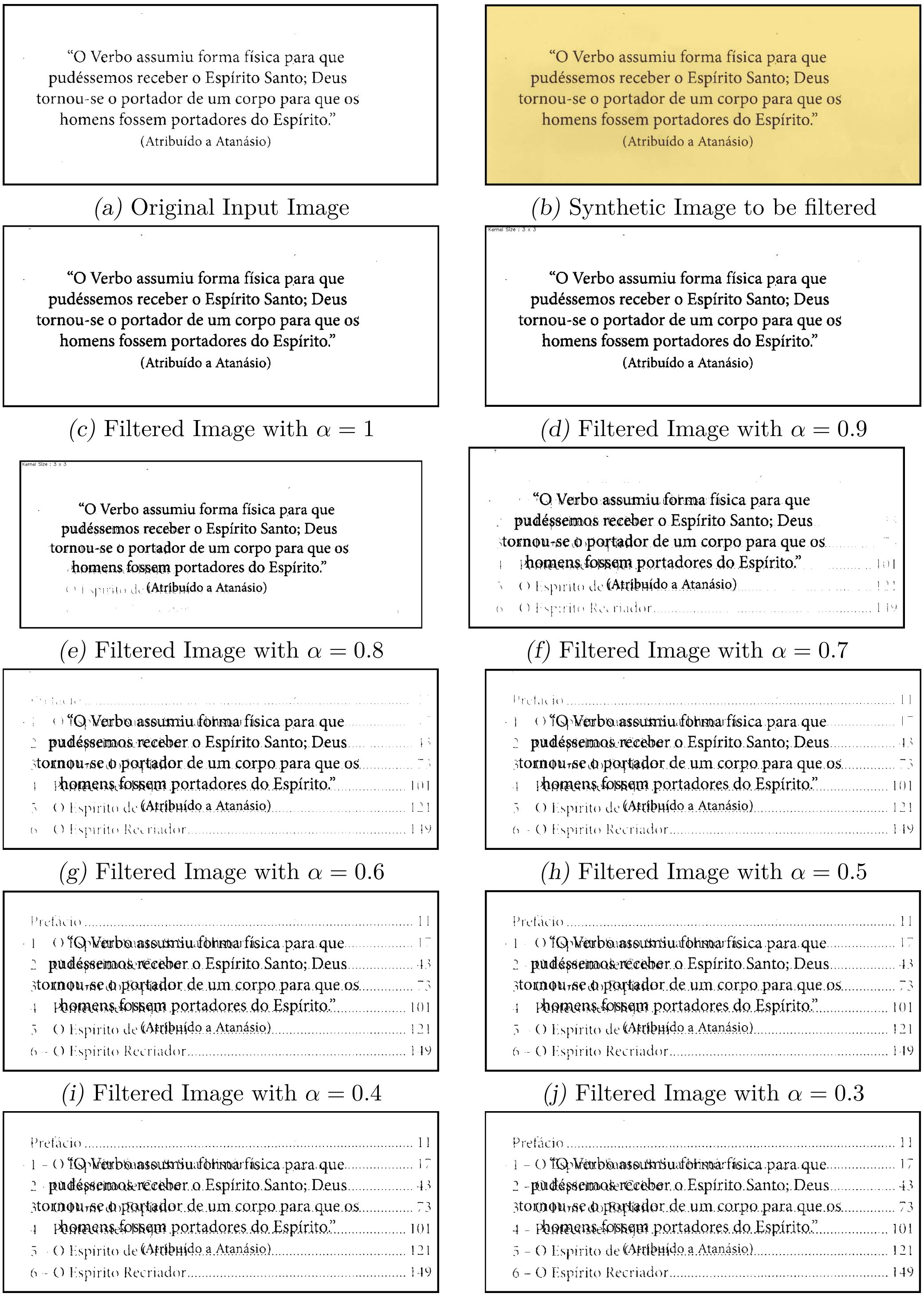
The value of foreground-foreground probability $P(f|f)$ varied between 22.13% and 23.70%, registering an error up to 77.87% in the range, where the α varied between 1.0 to 0.5, displaying a loss of information in the text, when we do a visual analysis.

The results obtained by the Roe-Mello algorithm may be considered unsuitable for the binarization of the test set used.

3.3.9 Silva-Lins-Rocha Algorithm

The algorithm developed by Silva *et al.* (2006) was developed to further improve the Mello-Lins algorithm. It considers the histogram distribution as the 256-symbol source (a priori source) distribution. It is assumed the hypothesis that all the symbols are statistically independent. In the case of real images one knows that this hypothesis does not hold. However, according to Silva *et al.* (2006), this largely simplifies the algorithm and was supposed to yield better results than its predecessors. The result of applying Silva-Lins-Rocha algorithm to the test images provided the results presented in Table 3.10.

The result of applying Silva-Lins-Rocha Approach to the document image of Figure 3.16 is showed in Figure 3.27.



(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.27: Silva-Lins-Rocha filtering of the images in Figure 3.16.

Table 3.10 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

For visual analysis of the binarized images in Figure 3.27, it seems reasonable to consider important features such as partial elimination of back-to-front interference in

Table 3.10: Silva-Lins-Rocha Filter Result.

α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	89	97.60%	2.40%	78.73%	21.27%
0.2	1x1	153	233	252	3x3	95	97.77%	2.23%	82.80%	17.20%
0.3	1x1	153	233	252	3x3	105	97.94%	2.06%	86.73%	13.27%
0.4	1x1	153	233	252	3x3	115	98.17%	1.83%	90.60%	9.40%
0.5	1x1	153	233	252	3x3	126	98.44%	1.56%	94.96%	5.04%
0.6	1x1	153	233	252	3x3	137	98.80%	1.20%	99.22%	0.78%
0.7	1x1	153	233	252	3x3	150	98.80%	1.20%	100.00%	0.00%
0.8	1x1	153	233	252	3x3	161	98.98%	1.02%	100.00%	0.00%
0.9	1x1	153	233	252	3x3	167	99.07%	0.93%	100.00%	0.00%
1.0	1x1	153	233	252	3x3	165	99.26%	0.74%	100.00%	0.00%

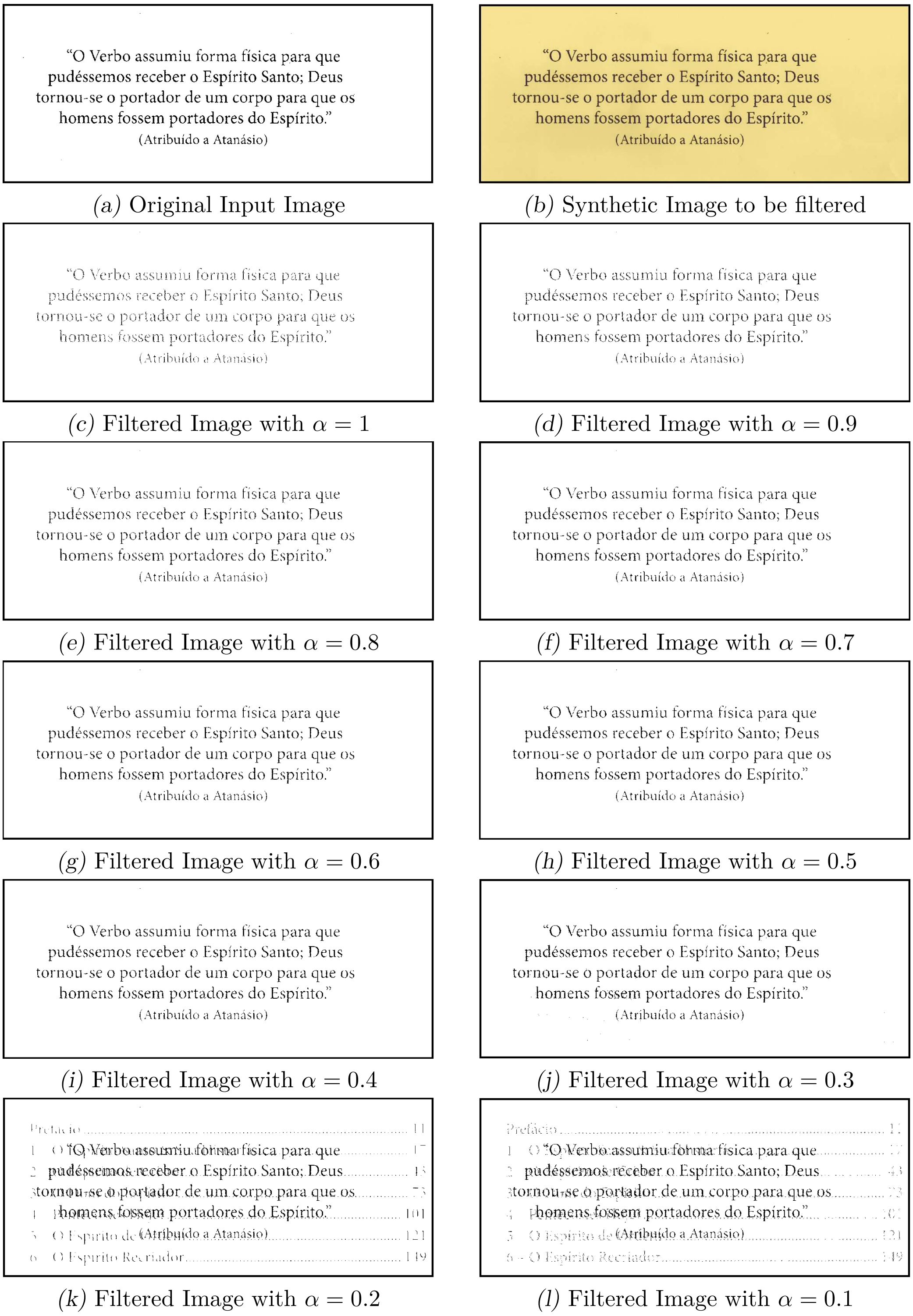
Figure 3.27 (c) to (f), which corresponds, in the range, where the alpha varied between 1.0 to 0.7, the value of background-background probability $P(b|b)$ varied between 99.26% and 98.80%, respectively, an error less than 1.20%, considering that the foreground-foreground probability $P(f|f)$ was of 100.00%. While from the images in Figure 3.27 (g) to (l) the back-to-front interference is present, the proportion of an α reduction.

As one may observe, considering the test set used, the Silva-Lins-Rocha actually performed better than the Mello-Lins algorithm for all values of fading coefficient but $\alpha = 0.9$, for some reason.

3.3.10 Wu-Lu Algorithm

The Wu-Lu binarization algorithm Wu (1998) was also originally developed for ultrasound images and seems to work particularly well in images with few contrast values. It is based on Shannon entropy and uses the lower difference between the minimum entropy of the objects and the entropy of the background as threshold value. Table 3.11 presents the results obtained in using Wu-Lu algorithm in the binarization of the test set images.

The result of applying Wu-Lu Algorithm to the document image of Figure 3.16 is showed in Figure 3.28.


"O Verbo assumiu forma física para que pudéssemos receber o Espírito Santo; Deus tornou-se o portador de um corpo para que os homens fossem portadores do Espírito."
(Atribuído a Atanásio)

(a) Original Input Image

(b) Synthetic Image to be filtered

(c) Filtered Image with $\alpha = 1$

(d) Filtered Image with $\alpha = 0.9$

(e) Filtered Image with $\alpha = 0.8$

(f) Filtered Image with $\alpha = 0.7$

(g) Filtered Image with $\alpha = 0.6$

(h) Filtered Image with $\alpha = 0.5$

(i) Filtered Image with $\alpha = 0.4$

(j) Filtered Image with $\alpha = 0.3$

(k) Filtered Image with $\alpha = 0.2$

(l) Filtered Image with $\alpha = 0.1$

Figure 3.28: Wu-Lu filtering of the images in Figure 3.16.

Table 3.11 presents an analysis of the evolution of the different opacity coefficient α values versus the matrix of co-occurrence probability.

Analyzing the results presented in Table 3.11, one may see that, although the value of the percentage of background-background mapping P(b|b) did not vary much and is

Table 3.11: Wu-Lu Filter Result.

α	**kernel-Gaussian**	**Red**	**Green**	**Blue**	**kernel-Blur**	**Threshold**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	153	233	252	3x3	75	99.13%	0.87%	62.81%	37.19%
0.2	1x1	153	233	252	3x3	75	99.00%	1.00%	62.45%	37.55%
0.3	1x1	153	233	252	3x3	74	99.96%	0.04%	61.00%	39.00%
0.4	1x1	153	233	252	3x3	73	100.00%	0.00%	59.72%	40.28%
0.5	1x1	153	233	252	3x3	72	100.00%	0.00%	57.70%	42.30%
0.6	1x1	153	233	252	3x3	71	100.00%	0.00%	55.86%	44.14%
0.7	1x1	153	233	252	3x3	70	100.00%	0.00%	54.23%	45.77%
0.8	1x1	153	233	252	3x3	68	100.00%	0.00%	50.21%	49.79%
0.9	1x1	153	233	252	3x3	66	100.00%	0.00%	45.99%	54.01%
1.0	1x1	153	233	252	3x3	62	100.00%	0.00%	36.61%	63.39%

either 100.00% or very close to that value for all the alphas, the P(f|f) value of foreground-foreground mapping varied between 36.61% and 59.72%, registering an error up to 63.39%, a strong loss of information in the text. That indicates that the Wu-Lu algorithm is possibly not suitable to binarize such set of document images.

3.4 Classification of Binarized Images of the Filters Studied

The assessment presented in the last section for the ten selected binarization algorithms presented for one test set formed by ten synthetic images obtained with ten different fading coefficients α varying from 0.1 to 1.0 in steps of 0.1 showed that the performance of the algorithms is highly dependent of the features of the document image.

The 16 documents shown in Figures 3.5 and 3.6, gave rise to 16 textures used to generate synthetic documents. Each of them were overlapped with another image to generate the document with 10 different intensities of α. The resulting 160 synthetic images were filtered using the following ten methods to remove back-to-front interference: Isodata, Pun, Kapur-Sahoo-Wong, Johannsen-Bille, Yen-Chang-Chang, Otsu, Mello-Lins, Roe-Mello, Silva-Lins-Rocha and Wu-Lu. At yielding 1,600 binarized images. To access the quality of the resulting images, the 1,600 matrices of co-occurrence probability were calculated.

Analyzing those $1,600$ matrices of co-occurrence probability, one can find the best filters for the different textures of historical documents with back-to-front interference. The average of the results of P(b|b)% and P(f|f)% were taken for each of the filters assessed for each value of α. The filters that showed both P(b|b)% and P(f|f)% average values higher than 99% and are presented in Table 3.12. For each historical document, one can select a filter with the optimal threshold. The results are shown in Table 3.12. The σ-aging represents the standard deviation of the texture.

Table 3.12: Overall algorithm classification for 1,600 synthetic images with $0 \leq \alpha \leq 1$ in steps of 0.1.

Red	**Green**	**Blue**	**Grayscale**	σ	**Mode**	σ**-aging**	**alpha**	**p(b|b)**	**p(f|f)**	**Filter**	**Threshold**
252	233	153	230	5	231	3	0.8	99.94	99.23	IsoData	137
252	233	153	230	5	231	3	0.9	99.98	99.20	IsoData	137
252	233	153	230	5	231	3	1.0	100.00	99.56	IsoData	138
252	233	153	230	5	231	3	0.7	99.25	100.00	Kapur-Sahoo-Wong	147
252	233	153	230	5	231	3	0.7	99.87	99.54	Otsu	138
252	233	153	230	5	231	3	0.8	99.94	99.56	Otsu	138
252	233	153	230	5	231	3	0.9	99.97	99.53	Otsu	138
252	233	153	230	5	231	3	1.0	99.95	99.56	Otsu	140
252	233	153	230	5	231	3	0.8	98.98	100.00	Silva-Lins-Rocha	161
252	233	153	230	5	231	3	0.9	99.07	100.00	Silva-Lins-Rocha	167
252	233	153	230	5	231	3	1.0	99.26	100.00	Silva-Lins-Rocha	165
252	233	153	230	5	231	3	0.9	99.19	100.00	Mello-Lins	165
253	238	158	233	3	233	3	0.7	99.75	99.45	Otsu	141
253	238	158	233	3	233	3	0.8	99.92	99.16	Otsu	140
253	238	158	233	3	233	3	0.9	99.97	99.14	Otsu	140
253	238	158	233	3	233	3	0.9	99.04	100.00	Silva-Lins-Rocha	171
253	238	158	233	3	233	3	0.9	99.29	100.00	Mello-Lins	165
241	200	136	205	4	206	5	0.7	99.19	99.53	IsoData	122
241	200	136	205	4	206	5	0.8	99.84	99.55	IsoData	120
241	200	136	205	4	206	5	0.9	99.87	99.55	IsoData	121
241	200	136	205	4	206	5	0.8	99.79	99.55	Otsu	122
241	200	136	205	4	206	5	0.9	99.81	99.55	Otsu	123
241	200	136	205	4	206	5	0.9	99.04	100.00	Silva-Lins-Rocha	147
245	204	141	209	4	211	5	0.7	99.59	99.60	IsoData	124
245	204	141	209	4	211	5	0.8	99.87	99.65	IsoData	124
245	204	141	209	4	211	5	0.9	99.92	99.64	IsoData	124
245	204	141	209	4	211	5	0.7	99.19	99.60	Otsu	127
245	204	141	209	4	211	5	0.8	99.84	99.65	Otsu	125
245	204	141	209	4	211	5	0.9	99.87	99.64	Otsu	126
245	204	141	209	4	211	5	0.9	99.04	100.00	Silva-Lins-Rocha	152
201	181	134	182	7	184	7	0.9	99.09	100.00	Silva-Lins-Rocha	127
254	247	208	244	4	246	5	0.7	99.48	100.00	Kapur-Sahoo-Wong	164
254	247	208	244	4	246	5	0.8	99.17	100.00	Kapur-Sahoo-Wong	180
254	247	208	244	4	246	5	0.9	99.04	100.00	Silva-Lins-Rocha	126
254	247	208	244	4	246	5	1.0	99.27	100.00	Silva-Lins-Rocha	123
254	247	208	244	4	246	5	0.8	99.67	100.00	Mello-Lins	168
254	247	208	244	4	246	5	0.9	99.84	100.00	Mello-Lins	164
254	247	208	244	4	246	5	1.0	99.45	100.00	Mello-Lins	179
253	232	152	229	4	229	3	0.7	99.78	99.21	IsoData	136
253	232	152	229	4	229	3	0.8	99.92	99.21	IsoData	136
253	232	152	229	4	229	3	0.9	99.97	99.20	IsoData	136
253	232	152	229	4	229	3	1.0	100.00	99.51	IsoData	137
253	232	152	229	4	229	3	0.8	99.07	100.00	Kapur-Sahoo-Wong	157
253	232	152	229	4	229	3	0.7	99.59	99.53	Otsu	139
253	232	152	229	4	229	3	0.8	99.87	99.54	Otsu	139
253	232	152	229	4	229	3	0.9	99.92	99.52	Otsu	139
253	232	152	229	4	229	3	1.0	99.95	99.51	Otsu	139
253	232	152	229	4	229	3	0.9	99.04	100.00	Silva-Lins-Rocha	167
253	232	152	229	4	229	3	1.0	99.27	100.00	Silva-Lins-Rocha	164
253	232	152	229	4	229	3	0.9	99.14	100.00	Mello-Lins	165
253	223	156	224	6	227	5	1.0	100.00	99.09	IsoData	135
253	223	156	224	6	227	5	0.8	99.07	100.00	Kapur-Sahoo-Wong	156
253	223	156	224	6	227	5	0.7	99.75	99.45	Otsu	136
253	223	156	224	6	227	5	0.8	99.92	99.14	Otsu	136
253	223	156	224	6	227	5	0.9	99.97	99.46	Otsu	136
253	223	156	224	6	227	5	1.0	99.95	99.48	Otsu	138
253	223	156	224	6	227	5	0.9	99.04	100.00	Silva-Lins-Rocha	166
253	223	156	224	6	227	5	1.0	99.27	100.00	Silva-Lins-Rocha	163
253	223	156	224	6	227	5	0.9	99.09	100.00	Mello-Lins	165
248	197	132	205	6	207	7	1.0	100.00	99.09	IsoData	124
248	197	132	205	6	207	7	0.8	99.07	100.00	Kapur-Sahoo-Wong	140
248	197	132	205	6	207	7	0.7	99.75	99.45	Otsu	132
248	197	132	205	6	207	7	0.8	99.92	99.14	Otsu	125
248	197	132	205	6	207	7	0.9	99.97	99.46	Otsu	126
248	197	132	205	6	207	7	1.0	99.95	99.48	Otsu	126
248	197	132	205	6	207	7	0.9	99.04	100.00	Silva-Lins-Rocha	149
248	197	132	205	6	207	7	1.0	99.27	100.00	Silva-Lins-Rocha	146
248	197	132	205	6	207	7	0.9	99.09	100.00	Mello-Lins	165
212	192	141	192	7	193	7	0.7	99.34	99.55	IsoData	121
212	192	141	192	7	193	7	1.0	99.91	99.58	IsoData	112
212	192	141	192	7	193	7	0.8	99.79	99.56	Otsu	113
212	192	141	192	7	193	7	0.9	99.84	99.58	Otsu	113
212	192	141	192	7	193	7	1.0	99.88	99.58	Otsu	113
212	192	141	192	7	193	7	0.9	99.09	100.00	Silva-Lins-Rocha	137
212	192	141	192	7	193	7	1.0	99.30	100.00	Silva-Lins-Rocha	134
242	243	247	243	5	246	5	0.7	99.59	100.00	Kapur-Sahoo-Wong	171
242	243	247	243	5	246	5	0.8	99.17	100.00	Kapur-Sahoo-Wong	188
242	243	247	243	5	246	5	0.9	99.04	100.00	Silva-Lins-Rocha	199
242	243	247	243	5	246	5	1.0	99.27	100.00	Silva-Lins-Rocha	196
242	243	247	243	5	246	5	0.7	99.34	100.00	Mello-Lins	173
242	243	247	243	5	246	5	0.8	99.89	100.00	Mello-Lins	170
242	243	247	243	5	246	5	1.0	99.67	100.00	Mello-Lins	182
226	219	207	220	4	219	3	0.7	99.48	100.00	Kapur-Sahoo-Wong	145
226	219	207	220	4	219	3	0.8	99.17	100.00	Kapur-Sahoo-Wong	161
226	219	207	220	4	219	3	0.9	99.04	100.00	Silva-Lins-Rocha	172
226	219	207	220	4	219	3	1.0	99.27	100.00	Silva-Lins-Rocha	169
226	219	207	220	4	219	3	0.9	99.29	100.00	Mello-Lins	166

The data presented in Table 3.12 corroborate the hypothesis formulated that the performance of binarization algorithms depends heavily on the “intrinsic nature” of the document image, and that a small variation in the image may yield completely different performance figures. In that sense, the data presented in this section must be read as a simple indicator of the quality of the images generated by those algorithms using a controlled test set, not being adequate to read the results as a quality classification rank for the compared algorithms. No binarization algorithm is an “all-kind-of-document” winner. Several factors such as paper texture, aging, thickness, tranlucidity, permability, the kind of ink, its fluidity, color, aging, etc all may influence the performance of each algorithm.

3.5 Conclusions

The 16 historical documents from the bequest of Joaquim Nabuco bequest and the SBrT file were used to generate 160 synthesised images with different textures and 10 different strengths of back-to-front interference ($0.1 \leq \alpha \leq 1.0$) such images were passed through the nine filters studied: Isodata, Pun, Kapur-Sahoo-Wong, Johannsen-Bille, Yen-Chang-Chang, Otsu, Mello-Lins, Silva-Lins-Rocha and Wu-Lu, generating a database of 1,600 binarized documents.

Of the 16 historical documents analyzed, 12 could have some treatment by the studied filters, under the imposed conditions of the co-occurrence probabilities $P(b|b) \geq 99.00\%$ and P $(f|f) \geq 99.00\%$ shown as an example in Section 3.4. In each binarized document one has one or more filters selected with the optimum threshold and corresponding co-occurrence probabilities $P(b|b)$ and $(f|f)$, providing a measure of the quality of the binarized images. The remaining 4 historical documents the filters were not able to meet the restrictions above.

The Pun method was not efficient in filtering the 160 images generated synthetically, because the co-occurrence probability $P(b|f)$ or maximum error was above 32.00%. The visual inspection of the binarized documents also corroborate to this.

Thus, from the visual aspect of the binarized images studied as well as the matrix co-occurrence probability, one finds that the remaining algorithms are good thresholding methods to remove back-to-from interference in historical documents, depending on the texture of the historical document and the intensity of α of the back-to-front interference, according to the results shown in Table 3.12.

The filters that met the requirements of co-occurrence probabilities in the given example were: Isodata, Otsu, Kapur-Sahoo-Wong, Silva-Lins-Rocha and Mello-Lins. The opacity factor value α varied from 0.7 to 1.0.

4 A NEW BINARIZATION ALGORITHM FOR IMAGES WITH BACK-TO-FRONT INTERFERENCE

The previous Chapter presented the problem of back-to-front interference and assessed some of the different techniques to filter out such a noise in document images. This Chapter presents a new algorithm to remove such a noise and assesses its applicability.

4.1 The New Algorithm

The technical literature presents for document image processing several thresholding algorithms, both with and without back-to-front interference. None of them has been shown to be effective for all strengthes of interference, textures of the background, etc, (Leedham, 2002). This Chapter presents a new algorithm for filtering out the back-to-front interference, which examines both pixels of the foreground and the background regions.

The proposed filter adopted a global approach using four steps, as shown the block diagram in Figure 4.1.

1. Filtering using the Bilateral filter.
2. Split image in RGB components.
3. Decision-making block for each RGB channel.
4. Classify the binarized images.

The present study investigates the impact of selecting of proposed filter variables in order to remove the noise and back-to-front interference in the historical documents. A bilateral filter (whose details are described in a Appendix A) was used due its property of de-noising digital images while preserving the edges of the written/typed areas of the document. The RGB components of the image were used to analyze which of them best preserved the text information in the foreground. An adaptive binarization method inspired by Otsu's method was implemented, with a minute choice of thresholding level, as a priori information to differentiate between text and non-text regions, and the last

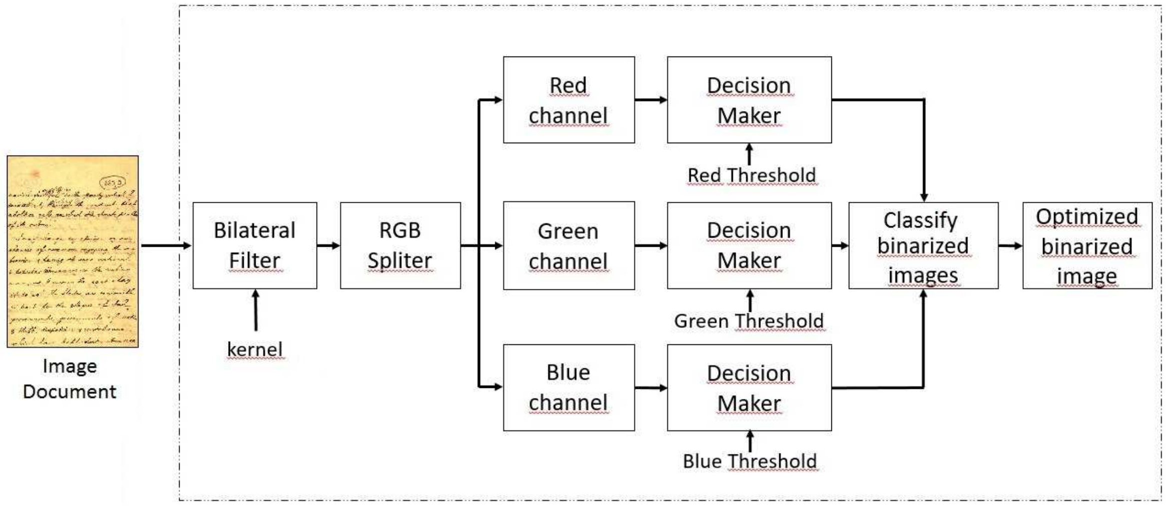


Figure 4.1: Block Diagram of the proposed method.

step was used to analyze and to decide which of the RGB components best preserved the text information in the foreground.

4.2 The Bilateral Filter

The bilateral filter is technique to smoothen images while preserving their edges. The filter output at each pixel is a weighted average of its neighbors. The weight assigned to each neighbor decreases with both the distance values among pixels of the image plane (the spatial domain S) and the distance on the intensity axis (the range domain R). The filter applies spatial weighted averaging without smoothing the edges. It combines two Gaussian filters; one filter works in the spatial domain, the other filter works in the intensity domain. Therefore, not only the spatial distance but also the intensity distance is important for the determination of weights. The bilateral filter combines two stages of filtering. These are the geometric closeness (i.e., filter domain) and the photometric similarity (i.e., filter range) among pixels in an $N \times N$ window size.

4.3 The Decision Making Block

After passing through the bilateral filter, the image is split into its Red, Green and Blue components, as shown the block diagram in Figure 4.1. Once the RGB channels are generated, the decision-making block is applied to process and the optimal threshold is calculated for each RGB channel, then three binary images are generated. The background-background probability is a function that needs to be optimized in the decision making block, mapping background pixels (paper) from the original image onto white pixels of the binary image. It depends of all the parameters of the original image texture, strength of the back to front interference simulated by the coefficient α , paper translucidity, etc. for each RGB channel. Thus, one can represent this dependence as shown in the Equation 4.3.1:

$$P(b,b) = f(\alpha, R, G, B) \tag{4.3.1}$$

The optimal threshold t^* for each channel is calculated in the decision-making block, maximizing P(b,b):

$$t^* = \text{Max}\, P(b|b) \tag{4.3.2}$$

subject to a given criterion

$$P(f|f) \geq L$$

. The criterion used here was L=99%, that is at most 1% of the pixels may be incorrectly mapped. The matrix of co-occurrence probability is calculated and the decision maker chooses the best binarized image.

4.4 Image Classification

The image classification block analyses the three binary images in each of the channels and outputs the one that is considered the best one. The decision was made by an “intelligent” automatic classifier based on the results obtained in the synthetic images used in the training process.

4.5 Results of the Proposed Method

Figure 3.16 in Chapter 3 shows the document synthesized with back-to-front interference and aging for various α opacity coefficient values. The master original image has $521,920$ pixels of which $20,529$ are black and $501,391$ are white. Thus, the probability distribution is P(white pixels)= 96.07% and P(black pixels)= 3.93%.

The proposed method was applied to the document image in Figure 3.16 following the block diagram in Figure 4.1.

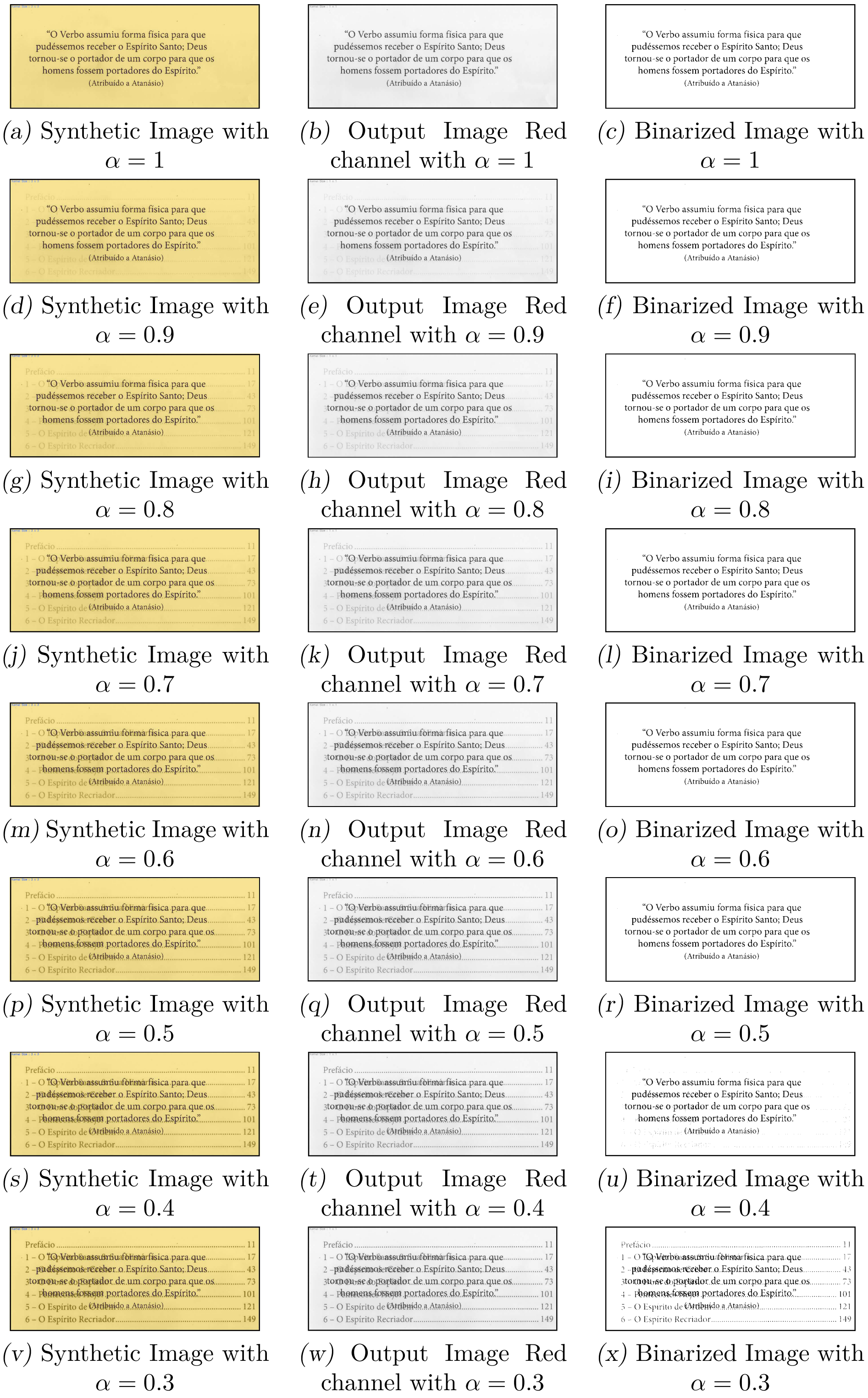
Several sets of experiments were ran on synthetically generated images sets to analyze the performance of the method here. Figure 4.2 show the results of this evaluation for the Red channel. For visual analysis of the binarized images in Figure 4.2 (c), (f), (i), (l), (o), (r) and (u), it seems reasonable to consider important features for removing back-to-front interference, which corresponds, in the range, where the alpha varied between 1.0 to 0.4. While from the images in Figure 4.2 (x) the back-to-front interference is present, the proportion of an alpha reduction.

Table 4.1 presents an analysis of the variation of the opacity coefficient α versus the co-occurrence probability matrix for the Red channel.

Table 4.1: Output proposed filter for the Red channel.

α	**kernel-Gauss**	**Red**	**Green**	**Blue**	**kernel-Blur**	**threshold**	**kernel-Bil**	**P(b|b)**	**P(b|f)**	**P(f|f)**	**P(f|b)**
0.1	1x1	154	233	253	3x3	126	1x1	96.49%	3.51%	100.00%	0.00%
0.2	1x1	154	233	253	3x3	126	1x1	96.93%	3.07%	100.00%	0.00%
0.3	1x1	154	233	253	3x3	126	1x1	97.66%	2.34%	100.00%	0.00%
0.4	1x1	154	233	253	3x3	126	1x1	99.60%	0.40%	100.00%	0.00%
0.5	1x1	154	233	253	3x3	126	1x1	99.87%	0.13%	100.00%	0.00%
0.6	1x1	154	233	253	3x3	126	1x1	99.91%	0.09%	100.00%	0.00%
0.7	1x1	154	233	253	3x3	126	1x1	99.94%	0.06%	100.00%	0.00%
0.8	1x1	154	233	253	3x3	126	1x1	99.97%	0.03%	100.00%	0.00%
0.9	1x1	154	233	253	3x3	126	1x1	99.99%	0.01%	100.00%	0,00%
1.0	1x1	154	233	253	3x3	126	1x1	100.00%	0.00%	100.00%	0.00%

Analyzing the co-occurrence probability matrix of the data produced by the proposed filter in Table 4.1, where the alpha varied between 1.0 and 0.4, the value of probability of background-background $P(b|b)$ varied between 100.00% and 99.60%, respectively, a error less than 0.40%, considering that the probability of co-occurrence $P(f|f)$ remained constant at 100%. Such results mean that the back-to-front interference was almost completely removed and the textual information was almost fully preserved. The results are consistent with the visual inspection of the images shown in Figure 4.2.



(a) Synthetic Image with $\alpha = 1$

(b) Output Image Red channel with $\alpha = 1$

(c) Binarized Image with $\alpha = 1$

(d) Synthetic Image with $\alpha = 0.9$

(e) Output Image Red channel with $\alpha = 0.9$

(f) Binarized Image with $\alpha = 0.9$

(g) Synthetic Image with $\alpha = 0.8$

(h) Output Image Red channel with $\alpha = 0.8$

(i) Binarized Image with $\alpha = 0.8$

(j) Synthetic Image with $\alpha = 0.7$

(k) Output Image Red channel with $\alpha = 0.7$

(l) Binarized Image with $\alpha = 0.7$

(m) Synthetic Image with $\alpha = 0.6$

(n) Output Image Red channel with $\alpha = 0.6$

(o) Binarized Image with $\alpha = 0.6$

(p) Synthetic Image with $\alpha = 0.5$

(q) Output Image Red channel with $\alpha = 0.5$

(r) Binarized Image with $\alpha = 0.5$

(s) Synthetic Image with $\alpha = 0.4$

(t) Output Image Red channel with $\alpha = 0.4$

(u) Binarized Image with $\alpha = 0.4$

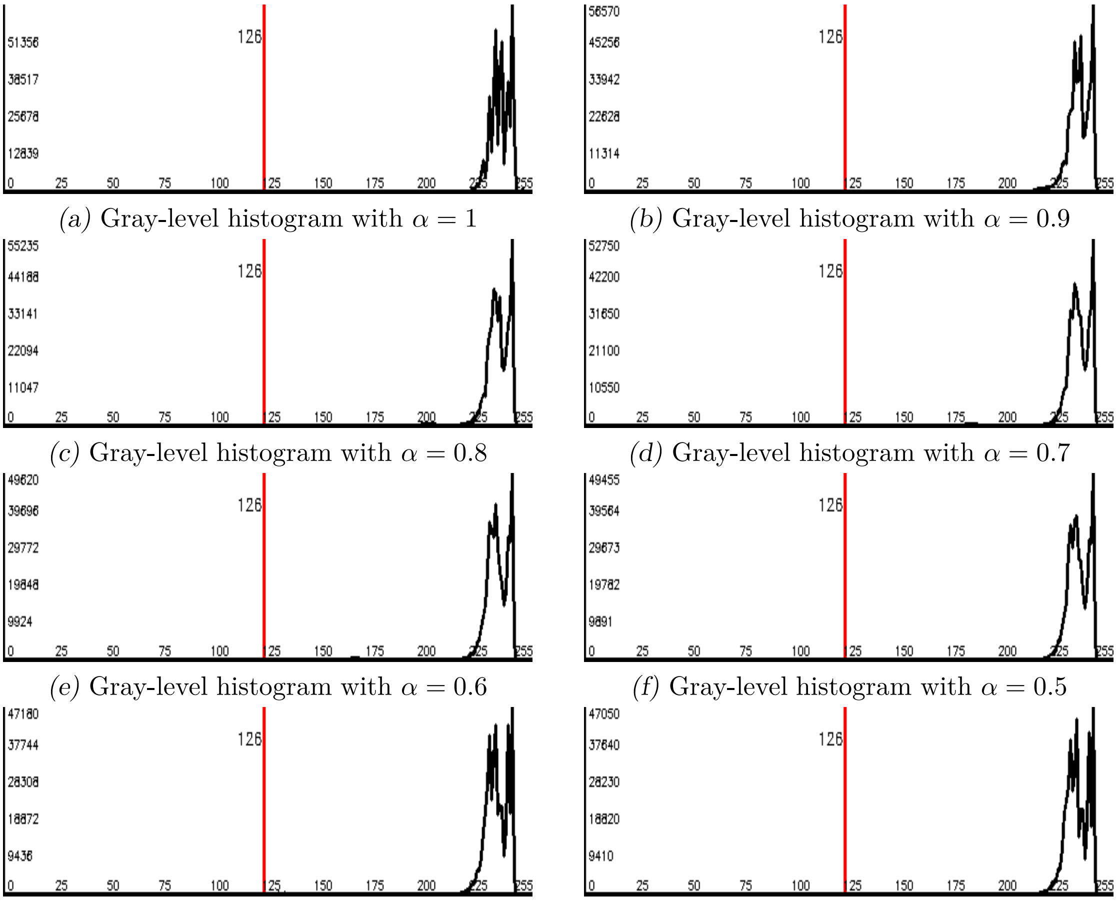
(v) Synthetic Image with $\alpha = 0.3$

(w) Output Image Red channel with $\alpha = 0.3$

(x) Binarized Image with $\alpha = 0.3$

Figure 4.2: Input and Output images of the Red channel using the proposed filter.

Figure 4.3 shows the decision-making block response for the optimal threshold selection t^*. For the Red channel the value that maximizes the background-background probability $P(b|b)$ was chosen, whose value was $t^*_{Red} = 126$, as shown by the grayscale histogram in Figure 4.3.



(a) Gray-level histogram with $\alpha = 1$

(b) Gray-level histogram with $\alpha = 0.9$

(c) Gray-level histogram with $\alpha = 0.8$

(d) Gray-level histogram with $\alpha = 0.7$

(e) Gray-level histogram with $\alpha = 0.6$

(f) Gray-level histogram with $\alpha = 0.5$

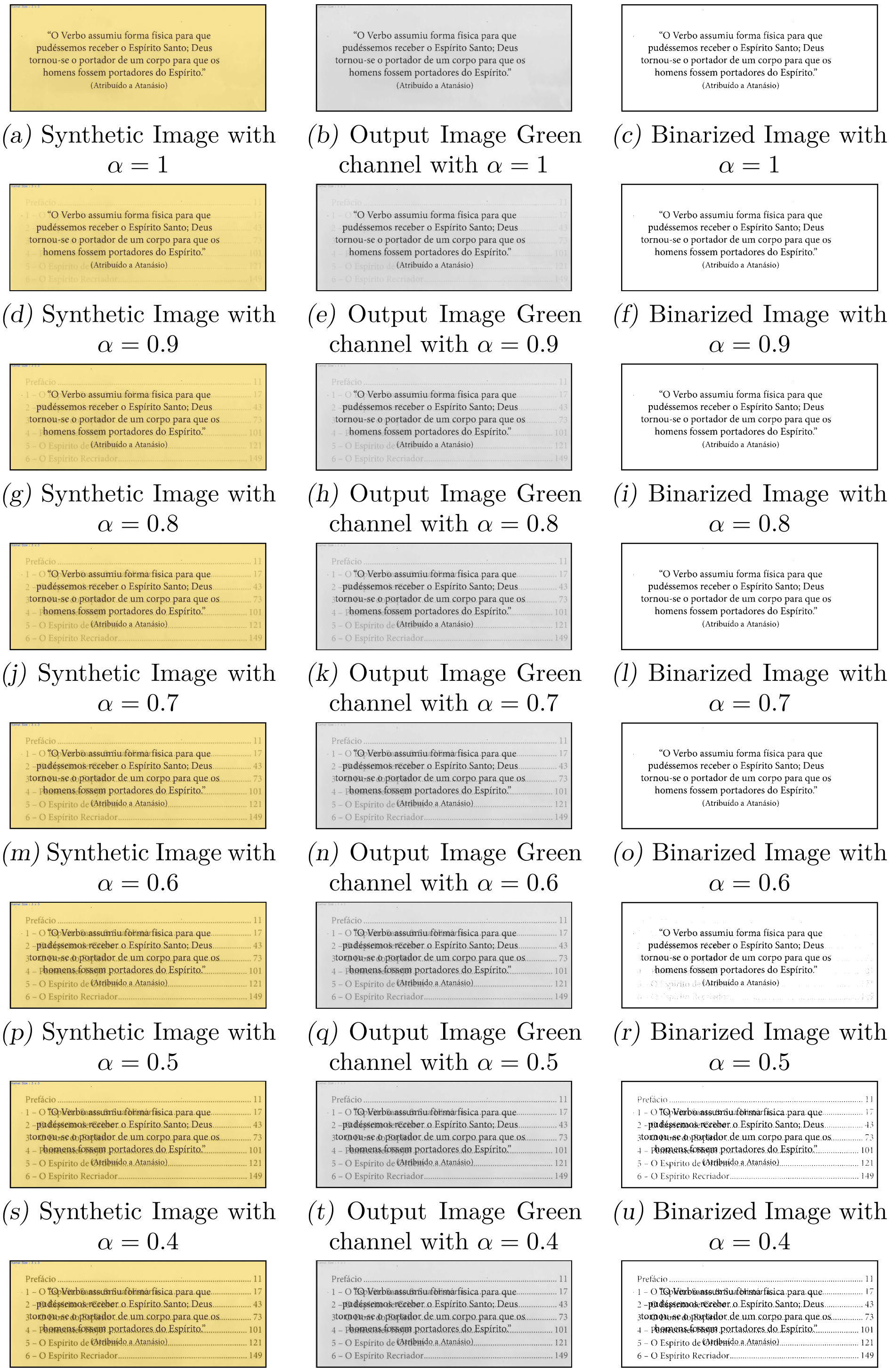
(g) Gray-level histogram with $\alpha = 0.4$

(h) Gray-level histogram with $\alpha = 0.3$

Figure 4.3: Gray-level Images Histogram Red channel from proposed filter.

Figure 4.4 shows the results of the variation of the α for the Green channel.

The examination of the images in Figure 4.4 (c), (f), (i), (l), (o) and (r), allows to observe the removal of the back-to-front interference, which corresponds, to the range of variation of the α between 1.0 and 0.5. The images in Figure 4.4 (u) and (x) shows that the back-to-front interference is present.



(a) Synthetic Image with $\alpha = 1$

(b) Output Image Green channel with $\alpha = 1$

(c) Binarized Image with $\alpha = 1$

(d) Synthetic Image with $\alpha = 0.9$

(e) Output Image Green channel with $\alpha = 0.9$

(f) Binarized Image with $\alpha = 0.9$

(g) Synthetic Image with $\alpha = 0.8$

(h) Output Image Green channel with $\alpha = 0.8$

(i) Binarized Image with $\alpha = 0.8$

(j) Synthetic Image with $\alpha = 0.7$

(k) Output Image Green channel with $\alpha = 0.7$

(l) Binarized Image with $\alpha = 0.7$

(m) Synthetic Image with $\alpha = 0.6$

(n) Output Image Green channel with $\alpha = 0.6$

(o) Binarized Image with $\alpha = 0.6$

(p) Synthetic Image with $\alpha = 0.5$

(q) Output Image Green channel with $\alpha = 0.5$

(r) Binarized Image with $\alpha = 0.5$

(s) Synthetic Image with $\alpha = 0.4$

(t) Output Image Green channel with $\alpha = 0.4$

(u) Binarized Image with $\alpha = 0.4$

(v) Synthetic Image with $\alpha = 0.3$

(w) Output Image Green channel with $\alpha = 0.3$

(x) Binarized Image with $\alpha = 0.3$

Figure 4.4: Input and Output images of the Green channel using the proposed filter.

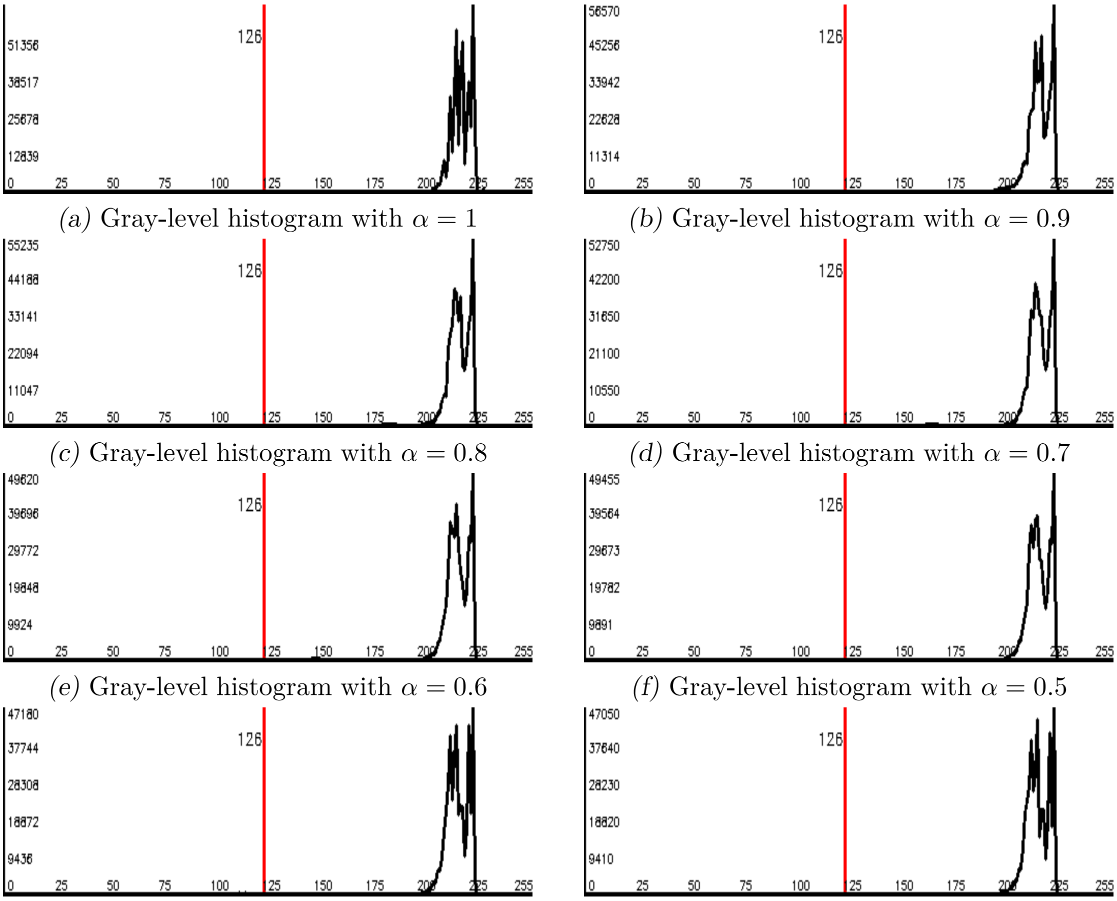
Table 4.2 presents an analysis of the variation of the opacity coefficient α versus the matrix of co-occurrence probability for Green channel.

Table 4.2: Output of the proposed filter for the Green channel.

α	**kernel-Gauss**	**Red**	**Green**	**Blue**	**kernel-Blur**	**threshold**	**kernel-Bil**	**P(bb)**	**P(bf)**	**P(ff)**	**P(fb)**
0.1	1x1	154	233	253	3x3	126	1x1	95.52%	4.48%	100.00%	0.00%
0.2	1x1	154	233	253	3x3	126	1x1	95.87%	4.13%	100.00%	0.00%
0.3	1x1	154	233	253	3x3	126	1x1	96.32%	3.68%	100.00%	0.00%
0.4	1x1	154	233	253	3x3	126	1x1	97.00%	3.00%	100.00%	0.00%
0.5	1x1	154	233	253	3x3	126	1x1	99.10%	0.90%	100.00%	0.00%
0.6	1x1	154	233	253	3x3	126	1x1	99.44%	0.56%	100.00%	0.00%
0.7	1x1	154	233	253	3x3	126	1x1	99.49%	0.51%	100.00%	0.00%
0.8	1x1	154	233	253	3x3	126	1x1	99.52%	0.48%	100.00%	0.00%
0.9	1x1	154	233	253	3x3	126	1x1	99.54%	0.46%	100.00%	0.00%
1.0	1x1	154	233	253	3x3	126	1x1	99.57%	0.43%	100.00%	0.00%

The co-occurrence probability matrix of the data produced by the proposed filter is shown in Table 4.2, in which the value of the α varied between 1.0 and 0.5, the value of the probability of background-background $P(b|b)$ varied between 100.00% and 99.10%, respectively, a error of less than 0.90%, considering that the co-occurrence probability $P(f|f)$ remained constant at 100%, meaning that back-to-front interference was almost completely removed and that the textual information was fully preserved, in this range. Such results are consistent with the visual inspection of images in Figure 4.4.

Figure 4.5 shows the decision-making block response for the optimal threshold selection t^*. For the Green channel the value that maximizes the background-background probability $P(b|b)$, was $t^*_{Green} = 126$, as shown by the grayscale histogram in Figure 4.5.



(a) Gray-level histogram with $\alpha = 1$

(b) Gray-level histogram with $\alpha = 0.9$

(c) Gray-level histogram with $\alpha = 0.8$

(d) Gray-level histogram with $\alpha = 0.7$

(e) Gray-level histogram with $\alpha = 0.6$

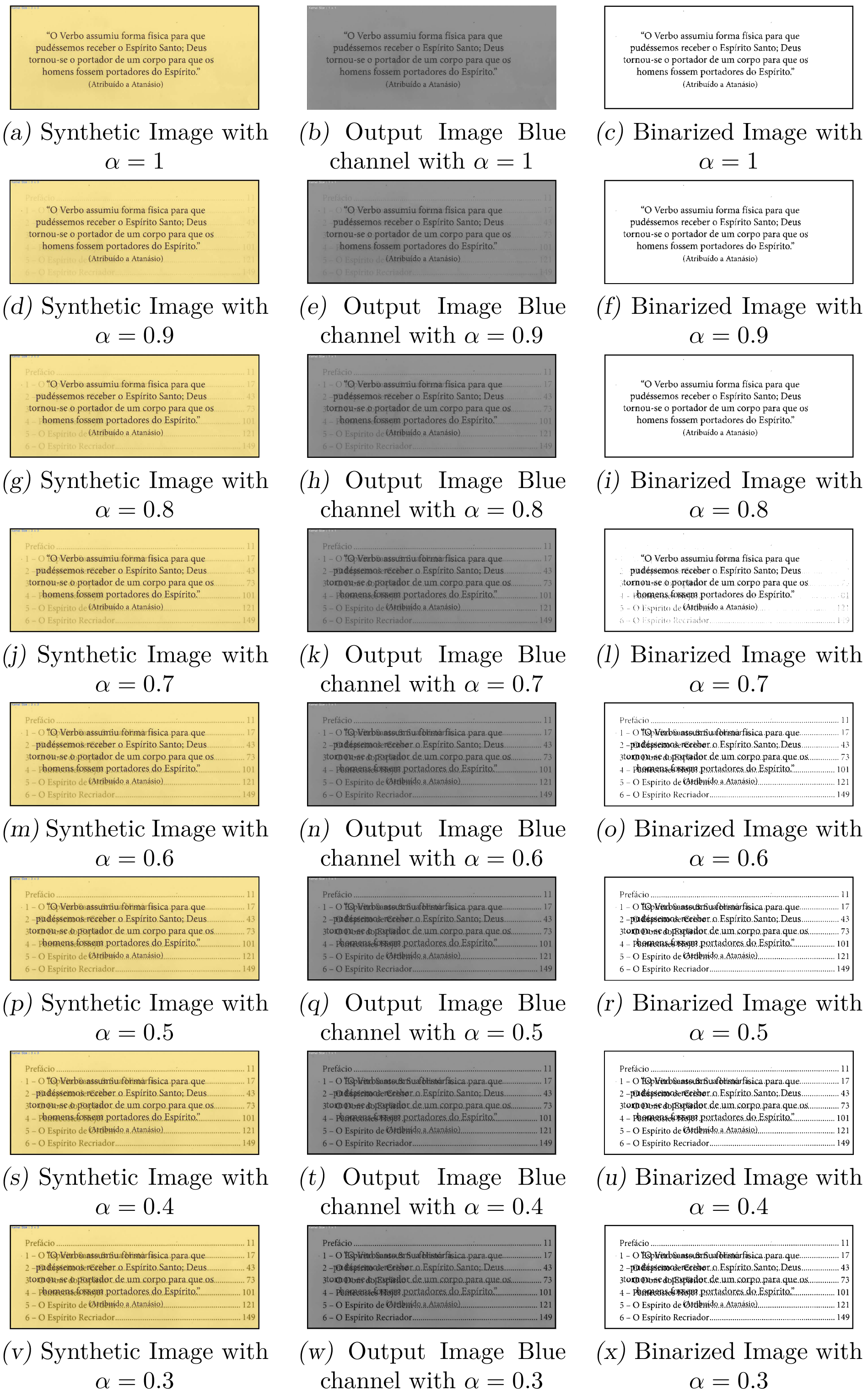
(f) Gray-level histogram with $\alpha = 0.5$

(g) Gray-level histogram with $\alpha = 0.4$

(h) Gray-level histogram with $\alpha = 0.3$

Figure 4.5: Output Images Histogram Green channel from proposed filter.

The output of the binarized images of the Blue channel is shown in Figure 4.6.


"O Verbo assumiu forma física para que
pudéssemos receber o Espírito Santo; Deus
tornou-se o portador de um corpo para que os
homens fossem portadores do Espírito."
(Atribuído a Atanásio)
Prefácio 11
1 – O Verbo 17
2 – 43
3 – 73
4 – 101
5 – O Espírito de 121
6 – O Espírito Recriador 149

(a) Synthetic Image with $\alpha = 1$

(b) Output Image Blue channel with $\alpha = 1$

(c) Binarized Image with $\alpha = 1$

(d) Synthetic Image with $\alpha = 0.9$

(e) Output Image Blue channel with $\alpha = 0.9$

(f) Binarized Image with $\alpha = 0.9$

(g) Synthetic Image with $\alpha = 0.8$

(h) Output Image Blue channel with $\alpha = 0.8$

(i) Binarized Image with $\alpha = 0.8$

(j) Synthetic Image with $\alpha = 0.7$

(k) Output Image Blue channel with $\alpha = 0.7$

(l) Binarized Image with $\alpha = 0.7$

(m) Synthetic Image with $\alpha = 0.6$

(n) Output Image Blue channel with $\alpha = 0.6$

(o) Binarized Image with $\alpha = 0.6$

(p) Synthetic Image with $\alpha = 0.5$

(q) Output Image Blue channel with $\alpha = 0.5$

(r) Binarized Image with $\alpha = 0.5$

(s) Synthetic Image with $\alpha = 0.4$

(t) Output Image Blue channel with $\alpha = 0.4$

(u) Binarized Image with $\alpha = 0.4$

(v) Synthetic Image with $\alpha = 0.3$

(w) Output Image Blue channel with $\alpha = 0.3$

(x) Binarized Image with $\alpha = 0.3$

Figure 4.6: Input and Output images of the Blue channel using the proposed filter.

A closer look at the binarized images in Figure 4.6, allows to observe that the images of the Blue channel did not produce good quality binary images. Therefore, the visual inspection of such images do not recommend the use of the Blue components in removing the front-to back interference in documents.

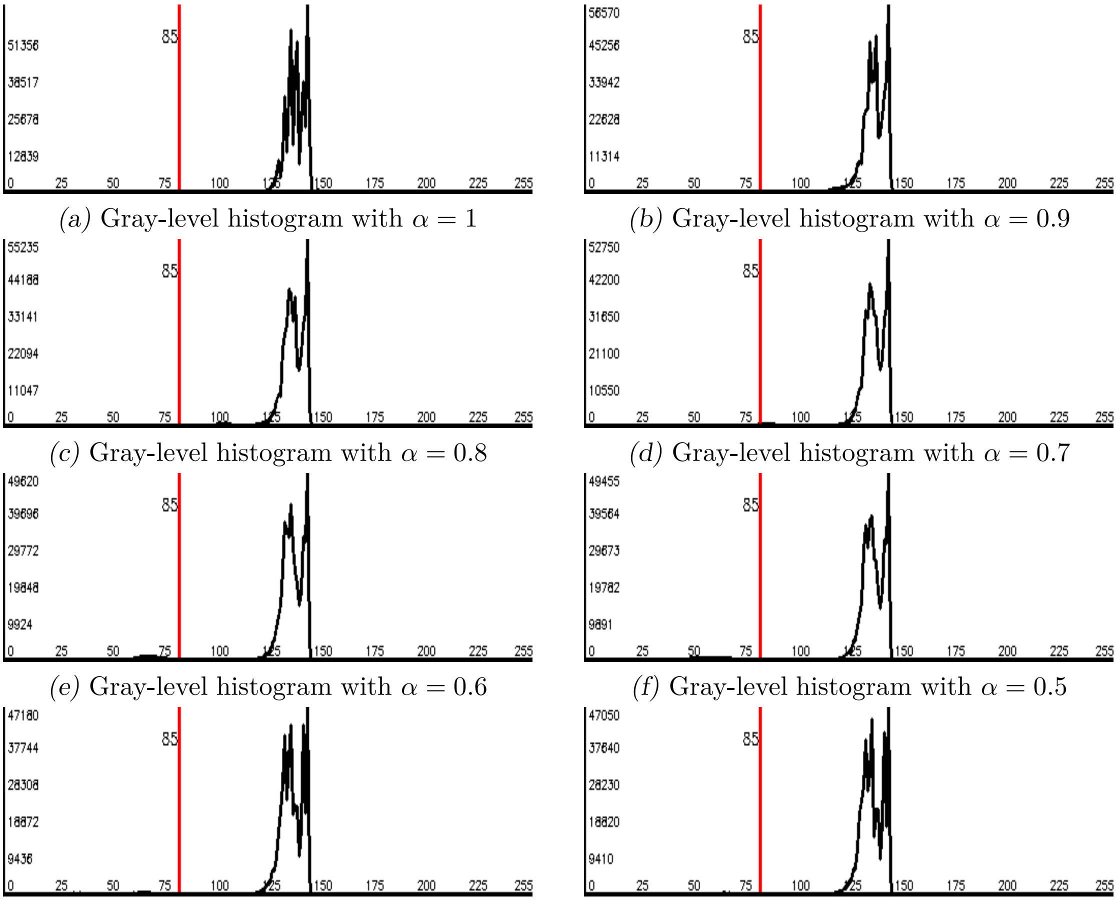
Table 4.3 presents an analysis of the variation of the opacity coefficient α versus the matrix of co-occurrence probability for the Blue channel.

Table 4.3: Output proposed filter for the Blue channel.

α	**kernel-Gauss**	**Red**	**Green**	**Blue**	**kernel-Blur**	**threshold**	**kernel-Bil**	**P(bb)**	**P(bf)**	**P(ff)**	**P(fb)**
0.1	1x1	154	233	254	3x3	85	1x1	96.87%	5.13%	90.32%	9.68%
0.2	1x1	154	233	254	3x3	85	1x1	94.76%	5.24%	89.59%	10.41%
0.3	1x1	154	233	254	3x3	85	1x1	94.10%	5.90%	87.05%	12.95%
0.4	1x1	154	233	254	3x3	85	1x1	94.39%	5.61%	86.68%	13.32%
0.5	1x1	154	233	254	3x3	85	1x1	94.83%	5.17%	86.42%	13.58%
0.6	1x1	154	233	254	3x3	85	1x1	95.53%	4.47%	86.32%	13.68%
0.7	1x1	154	233	254	3x3	85	1x1	97.64%	2.36%	86.42%	13.58%
0.8	1x1	154	233	254	3x3	85	1x1	98.56%	1.44%	86.45%	13.55%
0.9	1x1	154	233	254	3x3	85	1x1	98.62%	1.38%	86.47%	13.53%
1.0	1x1	154	233	254	3x3	85	1x1	98.66%	1.34%	86.51%	12.49%

The minimum value of the background-background probability P(b|b) presented in Table 4.3 was 94.10%, considering that the foreground-foreground probability P(f|f) varied between 90.32% and 86.32%, generating a loss of text information. Thus, the presented results show that the images that correspond to the Blue channel are not recommended for removing the front-to back interference, for any value of α.

Figure 4.7 shows the decision-making block response for the optimal threshold selection t^*. For the Blue channel the value that maximizes the background-background probability $P(b|b)$ was $t^*_{Blue} = 85$, as shown by the grayscale histogram in Figure 4.7.



(a) Gray-level histogram with $\alpha = 1$

(b) Gray-level histogram with $\alpha = 0.9$

(c) Gray-level histogram with $\alpha = 0.8$

(d) Gray-level histogram with $\alpha = 0.7$

(e) Gray-level histogram with $\alpha = 0.6$

(f) Gray-level histogram with $\alpha = 0.5$

(g) Gray-level histogram with $\alpha = 0.4$

(h) Gray-level histogram with $\alpha = 0.3$

Figure 4.7: Output Images Histogram for the Blue channel using the proposed filter.

4.6 Threshold Calculated using Regression Model

Based on the characteristics of the several historical documents, samples can provide enough statistical data to model the automatic calculation of the optimal threshold, using a regression analysis. From the set of data simulating with 160 documents synthesized it generated a threshold estimation model of the proposed method.

A sample of the historical document shown in Section 3.2 was used to measure the mean and variance values of the RGB texture measured using ImageJ. Table 4.4 shows the results of this evaluation, the value of the mean, the standard deviation, the mode and the grayscale intensity of the RGB components, all such information is needed to calculate the filtering threshold. The GrayScale intensity is calculated using in Equation 6.0.11

$$GrayScale - intensity = 0.299R + 0.587G + 0.114B \tag{4.6.1}$$

Table 4.4: Texture information of historical document.

Label	Mean	Standard Deviation	Mode	Min	Max
Red	252.740	2.506	255	226	255
Green	233.209	6.252	234	193	248
Blue	153.654	4.907	155	126	167
GrayScale Intensity	229.99	4.65	231	196	241

For the estimated value of the threshold $\hat{t}$ is given by the regression in Equation 6.0.12:

$$\begin{aligned}\hat{t} \;=\;& 1.80 + 8.57(\text{Red}) + 15.08(\text{Green}) + 3.00(\text{Blue}) - 25.44(\text{GrayScale-Intensity}) \\ & +7.78(\text{GrayScale-Standard-Deviation}) + 0.78(\text{Mode}) \\ & -6.73(\text{kernell aging}) \qquad (4.6.2)\end{aligned}$$

The analysis of Tables 4.5 and 4.6 guarantee the choice of the predictors and also explain the prediction model. Analyzing Table 4.5 we find that the p-value of all predictors are less than ≤ 0.05, except for the predicted Mode. If the p-value is less than or equal to α, one may conclude that the effect is statistically significant. The R^2 value describes the amount of variation in the observed response values that is explained by the predictors in Table 4.6. The R^2 (predict) value indicates that the model explains 72.13% of the variation in threshold when we use it for prediction.

Table 4.5: Analysis of Variance of Threshold.

Source Regression	DF	f-value	p-value
Predictors	7	56.19	0.000
kernell aging	1	31.08	0.000
Red	1	82.60	0.000
Green	1	84.58	0.000
Blue	1	69.16	0.000
Intensity mean	1	74.37	0.000
Intensity standard deviation	1	46.55	0.000
Mode	1	1.40	0.239

Table 4.6: Model Summary.

S	R^2	R^2 (adjusted)	R^2 (predict)
8.53	72.13%	70.84%	69.59%

The 160 synthesized documents with several aging textures in addition to the alpha values ranging from 0.1 to 1.0 in steps of 0.1, were used for generating a database of images whose sample is shown in Table 4.7. For each synthesized historical document, the prediction of threshold $\hat{t}$ was calculated. The "Error" column represents the relative error. The optimal threshold is calculated between the best threshold of R, G and B components.

Table 4.7: Result of the Simulation with Synthetic Documents using the Proposed Filter.

Red	**Green**	**Blue**	**kernell (aging)**	**Grayscale**	**Standard**	**Mode**	**Threshold**	$\hat{t}$	**Error**
252	233	153	5	230	5	231	126	121,41	3,64%
180	165	123	11	165	13	165	105	102,41	2,47%
253	238	158	3	233	3	233	130	126,94	2,35%
241	200	136	5	205	4	206	116	112,76	2,79%
245	204	141	5	209	4	211	120	116,7	2,75%
201	181	134	7	182	7	184	78	89,6	-14,87%
224	174	113	5	182	4	182	98	109,83	-12,07%
254	247	208	5	244	4	246	121	125,57	-3,78%
253	232	152	3	229	4	229	127	131,12	-3,24%
154	128	89	7	131	7	132	100	90,57	9,43%
218	167	113	5	177	6	180	92	97,17	-5,62%
253	223	156	5	224	6	227	127	138,26	-8,87%
248	197	132	7	205	6	207	124	116,83	5,78%
212	192	141	7	192	7	193	87	109,33	-25,67%
242	243	247	5	243	5	246	116	112,63	2,91%
226	219	207	3	220	4	219	100	105,45	-5,45%

The average Error, considering the training database of Table 4.7 is 6.98%.

4.7 Classification of Binarized Images Including the Proposed Filter

From these 16 documents shown in Figures 3.5 and 3.6, 160 images were generated based on synthetic documents similar to historical documents. With these images were investigated the nine methods to remove back-to-front interference: Isodata, Pun, Kapur-Sahoo-Wong, Johannsen-Bille, Yen-Chang-Chang, Otsu, Mello-Lins, Roe-Mello, Silva-Lins-Rocha and Wu-Lu. The 1,760 binarized images and 1,760 matrices of co-occurrence probability were evaluated, also considering the visual inspection of images, and discusses the quality assessment of images binary.

From the generated database, with information from $1,760$ matrices of co-occurrence probability, one can find the best filter for different textures of historical documents with front-to-back interference. As an example, for the condition of 1.00% pixel loss in the text and 1.00% front-to-back interference tolerance, one has $P(f|f) = 99.00\%$ and $P(b|b) = 99.00\%$. For each historical document, one can select a filter with the optimal threshold. The result is shown in tables 4.8 and 4.9:

Table 4.8: Result of the Proposed Filter with Synthetic Documents.

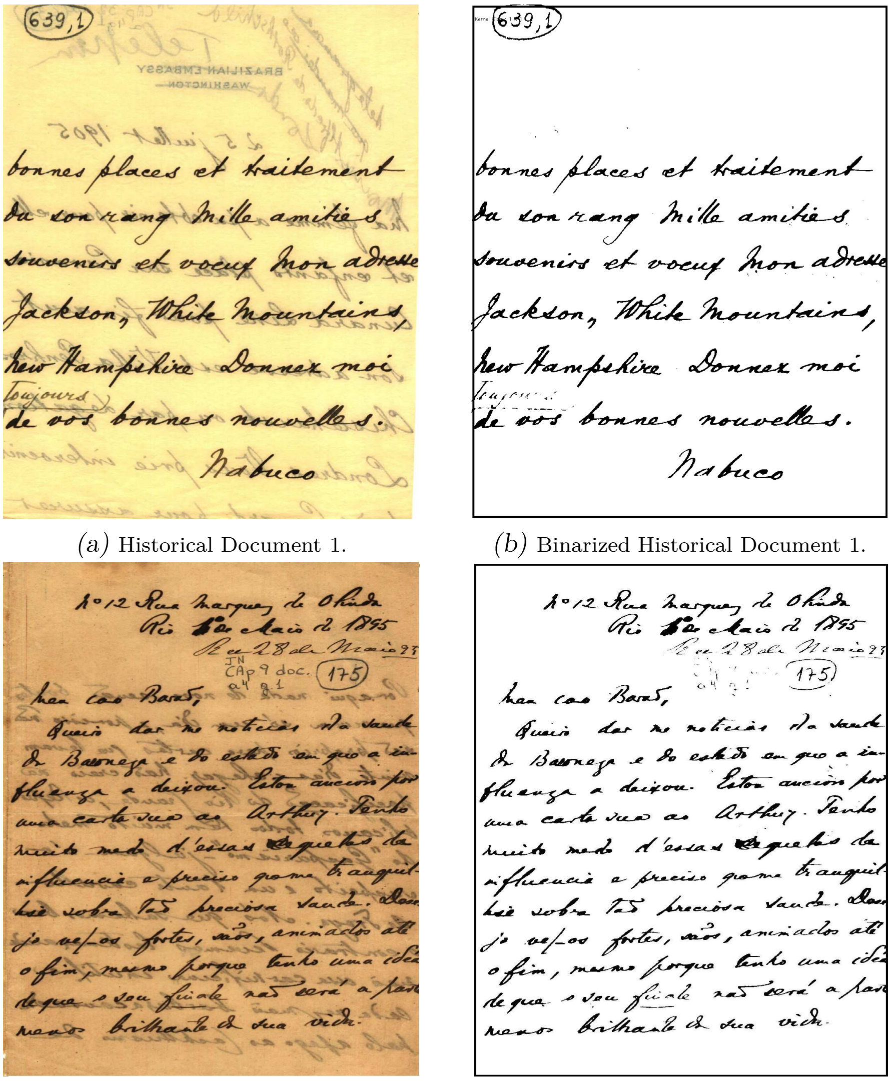
Red	Green	Blue	Gs	σ	Mode	σ-aging	alpha	p(b|b)	p(f|f)	Filter	Th	Channel
252	233	153	230	5	231	3	0.4	99.60	100.00	Proposed	126	Red
252	233	153	230	5	231	3	0.5	99.87	100.00	Proposed	126	Red
252	233	153	230	5	231	3	0.6	99.91	100.00	Proposed	126	Red
252	233	153	230	5	231	3	0.7	99.94	100.00	Proposed	126	Red
252	233	153	230	5	231	3	0.8	99.97	100.00	Proposed	126	Red
252	233	153	230	5	231	3	0.9	99.99	100.00	Proposed	126	Red
252	233	153	230	5	231	3	1.0	100.00	100.00	Proposed	126	Red
252	233	153	230	5	231	3	0.8	99.94	99.23	IsoData	137	
252	233	153	230	5	231	3	0.9	99.98	99.20	IsoData	137	
252	233	153	230	5	231	3	1.0	100.00	99.56	IsoData	138	
252	233	153	230	5	231	3	0.7	99.25	100.00	Kapur-Sahoo-Wong	147	
252	233	153	230	5	231	3	0.7	99.87	99.54	Otsu	138	
252	233	153	230	5	231	3	0.8	99.94	99.56	Otsu	138	
252	233	153	230	5	231	3	0.9	99.97	99.53	Otsu	138	
252	233	153	230	5	231	3	1.0	99.95	99.56	Otsu	140	
252	233	153	230	5	231	3	0.8	98.98	100.00	Silva-Lins-Rocha	161	
252	233	153	230	5	231	3	0.9	99.07	100.00	Silva-Lins-Rocha	167	
252	233	153	230	5	231	3	1.0	99.26	100.00	Silva-Lins-Rocha	165	
252	233	153	230	5	231	3	0.9	99.19	100.00	Mello-Lins	165	
253	238	158	233	3	233	3	0.4	99.24	100.00	Proposed	130	Red
253	238	158	233	3	233	3	0.5	99.75	100.00	Proposed	130	Red
253	238	158	233	3	233	3	0.6	99.80	100.00	Proposed	130	Red
253	238	158	233	3	233	3	0.7	99.82	100.00	Proposed	130	Red
253	238	158	233	3	233	3	0.8	99.85	100.00	Proposed	130	Red
253	238	158	233	3	233	3	0.9	99.87	100.00	Proposed	130	Red
253	238	158	233	3	233	3	1.0	99.89	100.00	Proposed	130	Red
253	238	158	233	3	233	3	0.7	99.75	99.45	Otsu	141	
253	238	158	233	3	233	3	0.8	99.92	99.16	Otsu	140	
253	238	158	233	3	233	3	0.9	99.97	99.14	Otsu	140	
253	238	158	233	3	233	3	0.9	99.04	100.00	Silva-Lins-Rocha	171	
253	238	158	233	3	233	3	0.9	99.29	100.00	Mello-Lins	165	
241	200	136	205	4	206	5	0.4	99.50	100.00	Proposed	116	Red
241	200	136	205	4	206	5	0.5	99.84	100.00	Proposed	116	Red
241	200	136	205	4	206	5	0.6	99.88	100.00	Proposed	116	Red
241	200	136	205	4	206	5	0.7	99.91	100.00	Proposed	116	Red
241	200	136	205	4	206	5	0.8	99.94	100.00	Proposed	116	Red
241	200	136	205	4	206	5	0.9	99.96	100.00	Proposed	116	Red
241	200	136	205	4	206	5	1.0	99.97	100.00	Proposed	116	Red
241	200	136	205	4	206	5	0.7	99.19	99.53	IsoData	122	
241	200	136	205	4	206	5	0.8	99.84	99.55	IsoData	120	
241	200	136	205	4	206	5	0.9	99.87	99.55	IsoData	121	
241	200	136	205	4	206	5	0.8	99.79	99.55	Otsu	122	
241	200	136	205	4	206	5	0.9	99.81	99.55	Otsu	123	
241	200	136	205	4	206	5	0.9	99.04	100.00	Silva-Lins-Rocha	147	
245	204	141	209	4	211	5	0.4	99.50	100.00	Proposed	120	
245	204	141	209	4	211	5	0.5	99.84	100.00	Proposed	120	
245	204	141	209	4	211	5	0.6	99.88	100.00	Proposed	120	
245	204	141	209	4	211	5	0.7	99.91	100.00	Proposed	120	
245	204	141	209	4	211	5	0.8	99.94	100.00	Proposed	120	
245	204	141	209	4	211	5	0.9	99.96	100.00	Proposed	120	
245	204	141	209	4	211	5	1.0	99.97	100.00	Proposed	120	
245	204	141	209	4	211	5	0.7	99.59	99.60	IsoData	124	
245	204	141	209	4	211	5	0.8	99.87	99.65	IsoData	124	
245	204	141	209	4	211	5	0.9	99.92	99.64	IsoData	124	
245	204	141	209	4	211	5	0.7	99.19	99.60	Otsu	127	
245	204	141	209	4	211	5	0.8	99.84	99.65	Otsu	125	
245	204	141	209	4	211	5	0.9	99.87	99.64	Otsu	126	
245	204	141	209	4	211	5	0.9	99.04	100.00	Silva-Lins-Rocha	152	
201	181	134	182	7	184	7	0.4	99.50	99.41	Proposed	78	Red
201	181	134	182	7	184	7	0.5	99.84	99.40	Proposed	78	Red
201	181	134	182	7	184	7	0.6	99.88	99.41	Proposed	78	Red
201	181	134	182	7	184	7	0.7	99.91	99.42	Proposed	78	Red
201	181	134	182	7	184	7	0.8	99.94	99.42	Proposed	78	Red
201	181	134	182	7	184	7	0.9	99.96	99.42	Proposed	78	Red
201	181	134	182	7	184	7	1.0	99.97	99.42	Proposed	78	Red
201	181	134	182	7	184	7	0.9	99.09	100.00	Silva-Lins-Rocha	127	
224	174	113	182	4	182	5	0.4	99.60	100.00	Proposed	98	Red
224	174	113	182	4	182	5	0.5	99.87	100.00	Proposed	98	Red
224	174	113	182	4	182	5	0.6	99.91	100.00	Proposed	98	Red
224	174	113	182	4	182	5	0.7	99.94	100.00	Proposed	98	Red
224	174	113	182	4	182	5	0.8	99.97	100.00	Proposed	98	Red
224	174	113	182	4	182	5	0.9	99.99	100.00	Proposed	98	Red
224	174	113	182	4	182	5	1.0	100.00	100.00	Proposed	98	Red
254	247	208	244	4	246	5	0.4	99.60	100.00	Proposed	121	Green
254	247	208	244	4	246	5	0.5	99.87	100.00	Proposed	121	Green
254	247	208	244	4	246	5	0.6	99.91	100.00	Proposed	121	Green
254	247	208	244	4	246	5	0.7	99.94	100.00	Proposed	121	Green
254	247	208	244	4	246	5	0.8	99.97	100.00	Proposed	121	Green
254	247	208	244	4	246	5	0.9	99.99	100.00	Proposed	121	Green
254	247	208	244	4	246	5	1.0	100.00	100.00	Proposed	121	Green
254	247	208	244	4	246	5	0.7	99.48	100.00	Kapur-Sahoo-Wong	164	
254	247	208	244	4	246	5	0.8	99.17	100.00	Kapur-Sahoo-Wong	180	
254	247	208	244	4	246	5	0.9	99.04	100.00	Silva-Lins-Rocha	126	
254	247	208	244	4	246	5	1.0	99.27	100.00	Silva-Lins-Rocha	123	
254	247	208	244	4	246	5	0.8	99.67	100.00	Mello-Lins	168	
254	247	208	244	4	246	5	0.9	99.84	100.00	Mello-Lins	164	
254	247	208	244	4	246	5	1.0	99.45	100.00	Mello-Lins	179	

Table 4.9: *Continuation of Table 4.8.*

Red	Green	Blue	Gs	σ	Mode	σ-aging	alpha	p(b|b)	p(f|f)	Filter	Th	Channel
253	232	152	229	4	229	3	0.4	99.60	100.00	Proposed	127	Red
253	232	152	229	4	229	3	0.5	99.87	100.00	Proposed	127	Red
253	232	152	229	4	229	3	0.6	99.91	100.00	Proposed	127	Red
253	232	152	229	4	229	3	0.7	99.94	100.00	Proposed	127	Red
253	232	152	229	4	229	3	0.8	99.97	100.00	Proposed	127	Red
253	232	152	229	4	229	3	0.9	99.99	100.00	Proposed	127	Red
253	232	152	229	4	229	3	1.0	100.00	100.00	Proposed	127	Red
253	232	152	229	4	229	3	0.7	99.78	99.21	IsoData	136	
253	232	152	229	4	229	3	0.8	99.92	99.21	IsoData	136	
253	232	152	229	4	229	3	0.9	99.97	99.20	IsoData	136	
253	232	152	229	4	229	3	1.0	100.00	99.51	IsoData	137	
253	232	152	229	4	229	3	0.8	99.07	100.00	Kapur-Sahoo-Wong	157	
253	232	152	229	4	229	3	0.7	99.59	99.53	Otsu	139	
253	232	152	229	4	229	3	0.8	99.87	99.54	Otsu	139	
253	232	152	229	4	229	3	0.9	99.92	99.52	Otsu	139	
253	232	152	229	4	229	3	1.0	99.95	99.51	Otsu	139	
253	232	152	229	4	229	3	0.9	99.04	100.00	Silva-Lins-Rocha	167	
253	232	152	229	4	229	3	1.0	99.27	100.00	Silva-Lins-Rocha	164	
253	232	152	229	4	229	3	0.9	99.14	100.00	Mello-Lins	165	
218	167	113	177	6	180	5	0.4	99.60	100.00	Proposed	92	Red
218	167	113	177	6	180	5	0.5	99.87	100.00	Proposed	92	Red
218	167	113	177	6	180	5	0.6	99.91	100.00	Proposed	92	Red
218	167	113	177	6	180	5	0.7	99.94	100.00	Proposed	92	Red
218	167	113	177	6	180	5	0.8	99.97	100.00	Proposed	92	Red
218	167	113	177	6	180	5	0.9	99.99	100.00	Proposed	92	Red
218	167	113	177	6	180	5	1.0	100.00	100.00	Proposed	92	Red
253	223	156	224	6	227	5	0.4	99.60	100.00	Proposed	127	Red
253	223	156	224	6	227	5	0.5	99.87	100.00	Proposed	127	Red
253	223	156	224	6	227	5	0.6	99.91	100.00	Proposed	127	Red
253	223	156	224	6	227	5	0.7	99.94	100.00	Proposed	127	Red
253	223	156	224	6	227	5	0.8	99.97	100.00	Proposed	127	Red
253	223	156	224	6	227	5	0.9	99.99	100.00	Proposed	127	Red
253	223	156	224	6	227	5	1.0	100.00	100.00	Proposed	127	Red
253	223	156	224	6	227	5	1.0	100.00	99.09	IsoData	135	
253	223	156	224	6	227	5	0.8	99.07	100.00	Kapur-Sahoo-Wong	156	
253	223	156	224	6	227	5	0.7	99.75	99.45	Otsu	136	
253	223	156	224	6	227	5	0.8	99.92	99.14	Otsu	136	
253	223	156	224	6	227	5	0.9	99.97	99.46	Otsu	136	
253	223	156	224	6	227	5	1.0	99.95	99.48	Otsu	138	
253	223	156	224	6	227	5	0.9	99.04	100.00	Silva-Lins-Rocha	166	
253	223	156	224	6	227	5	1.0	99.27	100.00	Silva-Lins-Rocha	163	
253	223	156	224	6	227	5	0.9	99.09	100.00	Mello-Lins	165	
248	197	132	205	6	207	7	0.4	99.60	100.00	Proposed	124	Red
248	197	132	205	6	207	7	0.5	99.87	100.00	Proposed	124	Red
248	197	132	205	6	207	7	0.6	99.91	100.00	Proposed	124	Red
248	197	132	205	6	207	7	0.7	99.94	100.00	Proposed	124	Red
248	197	132	205	6	207	7	0.8	99.97	100.00	Proposed	124	Red
248	197	132	205	6	207	7	0.9	99.99	100.00	Proposed	124	Red
248	197	132	205	6	207	7	1.0	100.00	100.00	Proposed	124	Red
248	197	132	205	6	207	7	1.0	100.00	99.09	IsoData	124	
248	197	132	205	6	207	7	0.8	99.07	100.00	Kapur-Sahoo-Wong	140	
248	197	132	205	6	207	7	0.7	99.75	99.45	Otsu	132	
248	197	132	205	6	207	7	0.8	99.92	99.14	Otsu	125	
248	197	132	205	6	207	7	0.9	99.97	99.46	Otsu	126	
248	197	132	205	6	207	7	1.0	99.95	99.48	Otsu	126	
248	197	132	205	6	207	7	0.9	99.04	100.00	Silva-Lins-Rocha	149	
248	197	132	205	6	207	7	1.0	99.27	100.00	Silva-Lins-Rocha	146	
248	197	132	205	6	207	7	0.9	99.09	100.00	Mello-Lins	165	
212	192	141	192	7	193	7	0.4	99.66	99.49	Proposed	87	Red
212	192	141	192	7	193	7	0.5	99.87	99.49	Proposed	87	Red
212	192	141	192	7	193	7	0.6	99.91	99.49	Proposed	87	Red
212	192	141	192	7	193	7	0.7	99.94	99.49	Proposed	87	Red
212	192	141	192	7	193	7	0.8	99.97	99.49	Proposed	87	Red
212	192	141	192	7	193	7	0.9	99.99	99.49	Proposed	87	Red
212	192	141	192	7	193	7	1.0	100.00	99.43	Proposed	87	Red
212	192	141	192	7	193	7	0.7	99.34	99.55	IsoData	121	
212	192	141	192	7	193	7	1.0	99.91	99.58	IsoData	112	
212	192	141	192	7	193	7	0.8	99.79	99.56	Otsu	113	
212	192	141	192	7	193	7	0.9	99.84	99.58	Otsu	113	
212	192	141	192	7	193	7	1.0	99.88	99.58	Otsu	113	
212	192	141	192	7	193	7	0.9	99.09	100.00	Silva-Lins-Rocha	137	
212	192	141	192	7	193	7	1.0	99.30	100.00	Silva-Lins-Rocha	134	
242	243	247	243	5	246	5	0.4	99.60	100.00	Proposed	116	Red
242	243	247	243	5	246	5	0.5	99.87	100.00	Proposed	116	Red
242	243	247	243	5	246	5	0.6	99.91	100.00	Proposed	116	Red
242	243	247	243	5	246	5	0.7	99.94	100.00	Proposed	116	Red
242	243	247	243	5	246	5	0.8	99.97	100.00	Proposed	116	Red
242	243	247	243	5	246	5	0.9	99.99	100.00	Proposed	116	Red
242	243	247	243	5	246	5	1.0	100.00	100.00	Proposed	116	Red
242	243	247	243	5	246	5	0.7	99.59	100.00	Kapur-Sahoo-Wong	171	
242	243	247	243	5	246	5	0.8	99.17	100.00	Kapur-Sahoo-Wong	188	
242	243	247	243	5	246	5	0.9	99.04	100.00	Silva-Lins-Rocha	199	
242	243	247	243	5	246	5	1.0	99.27	100.00	Silva-Lins-Rocha	196	
242	243	247	243	5	246	5	0.7	99.34	100.00	Mello-Lins	173	
242	243	247	243	5	246	5	0.8	99.89	100.00	Mello-Lins	170	
242	243	247	243	5	246	5	1.0	99.67	100.00	Mello-Lins	182	

4.8 Results of the Proposed Method

In the previous section the results for the synthesized images were presented. Now, in the Figures 4.8 and 4.9 the results applied to historical documents.



639,1

bonnes places et traitement
du son rang Mille amitiés
souvenirs et voeux Mon adresse
Jackson, White Mountains,
New Hampshire Donnez moi
toujours
de vos bonnes nouvelles.
Nabuco

(a) Historical Document 1.

639,1

bonnes places et traitement
du son rang Mille amitiés
souvenirs et voeux Mon adresse
Jackson, White Mountains,
New Hampshire Donnez moi
toujours
de vos bonnes nouvelles.
Nabuco

(b) Binarized Historical Document 1.

N.º 12 Rua Marquez de Olinda
Rio 1.º de Maio de 1895
Rec. 28 de Maio 95
JN
CAp 9 doc.
a4 g1
175

Meu caro Barão,
Queria dar-lhe noticias da saude
da Baroneza e do estado em que a in-
fluenza a deixou. Estou ancioso por
uma carta sua ao Arthur. Tenho
muito medo d'essas sequelas da
influencia e preciso que me tranquil-
lise sobre tão preciosa saude. Dese-
jo vel-os fortes, sãos, animados até
o fim, mesmo porque tenho uma idéa
de que o seu finale não será a parte
menos brilhante de sua vida.

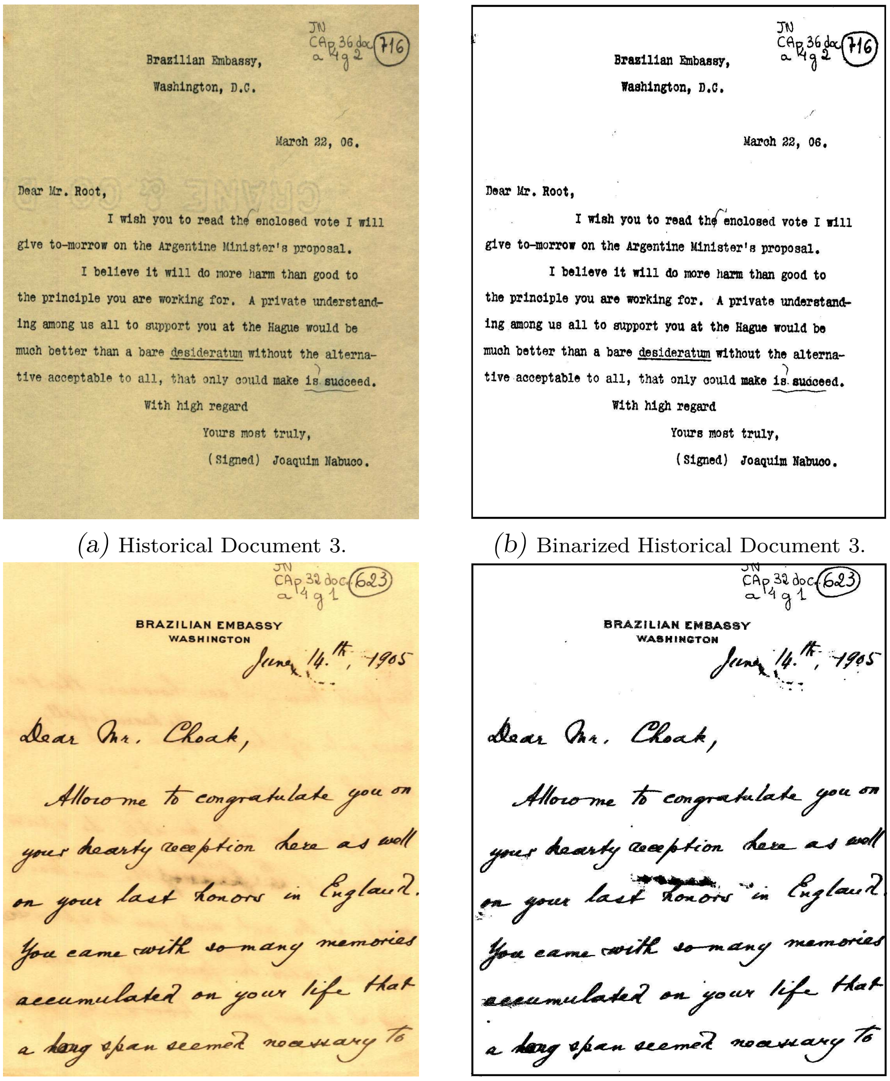
(c) Historical Document 2.

N.º 12 Rua Marquez de Olinda
Rio 1.º de Maio de 1895
Rec. 28 de Maio 95
175

Meu caro Barão,
Queria dar-lhe noticias da saude
da Baroneza e do estado em que a in-
fluenza a deixou. Estou ancioso por
uma carta sua ao Arthur. Tenho
muito medo d'essas sequelas da
influencia e preciso que me tranquil-
lise sobre tão preciosa saude. Dese-
jo vel-os fortes, sãos, animados até
o fim, mesmo porque tenho uma idéa
de que o seu finale não será a parte
menos brilhante de sua vida.

(d) Binarized Historical Document 2.

Figure 4.8: Result of the Proposed Filter with Historical Documents.



JN
CAp 36 doc 716
a 4 g 2

Brazilian Embassy,
Washington, D.C.

March 22, 06.

Dear Mr. Root,

I wish you to read the enclosed vote I will give to-morrow on the Argentine Minister's proposal.

I believe it will do more harm than good to the principle you are working for. A private understanding among us all to support you at the Hague would be much better than a bare desideratum without the alternative acceptable to all, that only could make is succeed.

With high regard
Yours most truly,
(Signed) Joaquim Nabuco.

(*a*) Historical Document 3.

JN
CAp 36 doc 716
a 4 g 2

Brazilian Embassy,
Washington, D.C.

March 22, 06.

Dear Mr. Root,

I wish you to read the enclosed vote I will give to-morrow on the Argentine Minister's proposal.

I believe it will do more harm than good to the principle you are working for. A private understanding among us all to support you at the Hague would be much better than a bare desideratum without the alternative acceptable to all, that only could make is succeed.

With high regard
Yours most truly,
(Signed) Joaquim Nabuco.

(*b*) Binarized Historical Document 3.

JN
CAp 32 doc 623
a 4 g 1

BRAZILIAN EMBASSY
WASHINGTON

June 14th, 1905

Dear Mr. Choate,

Allow me to congratulate you on your hearty reception here as well on your last honors in England. You came with so many memories accumulated on your life that a long span seemed necessary to

(*c*) Historical Document 4.

JN
CAp 32 doc 623
a 4 g 1

BRAZILIAN EMBASSY
WASHINGTON

June 14th, 1905

Dear Mr. Choate,

Allow me to congratulate you on your hearty reception here as well on your last honors in England. You came with so many memories accumulated on your life that a long span seemed necessary to

(*d*) Binarized Historical Document 4.

Figure 4.9: Result of the Proposed Filter with Historical Documents.

4.9 Conclusions

The filter proposed for back-to-front interference removal when tested with the 160 synthetic documents, yielded binary images that meet the imposed conditions of the co-occurrence probabilities $P(b|b) \geq 99.00\%$ and P $(f|f) \geq 99.00\%$ shown as an example in Section 4.7. The visual inspection of the binarized images studied confirmed the results obtained for the matrix co-occurrence probability.

Two of the test documents, the ones with texture (R=224, G=174, B=113) and (R=218, G=167, B=113), only provided reasonable results when binarized by the proposed filter, given the efficiency requirements imposed that the error is less of 1.00% In each binarized document one has one or more filters selected with the optimum threshold and corresponding co-occurrence probabilities $P(b|b)$ and $(f|f)$, providing a measure of the quality of the binarized images. The remaining two historical documents the filters were not able to meet the above conditions.

The filters that met the requirements of co-occurrence probabilities in the given example were: the Proposed Filter, Isodata, Otsu, Kapur-Sahoo-Wong, Silva-Lins-Rocha and the Mello-Lins. The opacity factor value α varied from 0.4 to 1.0, because of the inclusion of the proposed filter in the inclusion to remove back-to-front interference and the binarization processes.

The results presented for this algorithm show that for all the images in the chosen test set this algorithm performed better that its predecessors, exhibiting a steady "behavior" with the variation of the fading coefficient α. It is important to remark that this and the IsoData algorithms claim far more computational resources than the other algorithms assessed.

5 BINARIZING COMPLEX SCANNED DOCUMENTS

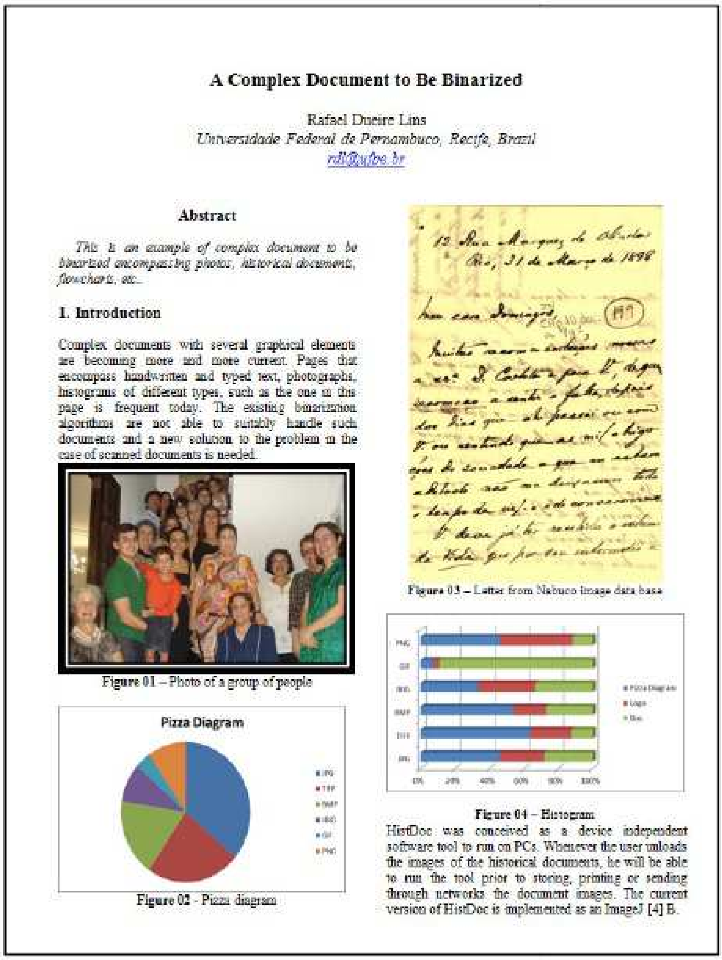
Most binarization algorithms are suitable for scanned text documents and do not work adequately with complex documents that encompass photos, charts and text simultaneously. This chapter presents a new binarization algorithm that works on complex documents. Each of the elements in the image is processed depending on its nature, thus their binarization takes that into account to preserve the original content.

5.1 Introduction

Complex documents encompassing not only text but also several graphic elements such as photos, histograms, pie (or pizza) diagrams, etc. are becoming of widespread use. Figure 5.1 shows an example of such a document. The binarization algorithms found in the literature are not suitable for such documents, as most of them use global threshold techniques.

The direct binarization of Figure 5.1 using (Otsu, 1979) algorithm provides the image shown in Figure 5.2, in which one may observe that the information provided by the pictorial elements was completely lost and that even the historic letter shown with back-to-front interference lost its legibility after binarization.

In 2007, the authors of this paper in the LiveMemory project (Lins, 2010a) took the challenge of building a digital library with the ten previous years of the proceedings of the Brazilian Telecommunications Society and distributing it with all the participants of the event in a DVD. In 2010, they had finished the library encompassing 26 years of the events. For that task 12 years of proceedings had to be scanned, filtered, and indexed. Pages were originally scanned in true color, 300 d.p.i. resolution. Back-to-front interference (Monte da Silva *et al.*, 2008) (bleeding) was observed in many pages and had to be filtered out. As the number of images to be included in the DVD was far too big, images had to be binarized. Several pages included graphical elements as photos, graphs, bar histograms, pie diagrams, etc. The direct binarization of such pages yielded the complete loss of the information of the graphical elements, in a similar situation to the one shown in Figure 5.2. The solution adopted then was to disassemble such pages,



A Complex Document to Be Binarized

Rafael Dueire Lins
Universidade Federal de Pernambuco, Recife, Brazil
rdl@ufpe.br

Abstract

This is an example of complex document to be binarized encompassing photos, historical documents, flowcharts, etc..

I. Introduction

Complex documents with several graphical elements are becoming more and more current. Pages that encompass handwritten and typed text, photographs, histograms of different types, such as the one in this page is frequent today. The existing binarization algorithms are not able to suitably handle such documents and a new solution to the problem in the case of scanned documents is needed.

Figure 01 – Photo of a group of people

Pizza Diagram

Figure 02 - Pizza diagram

Figure 03 – Letter from Nabuco image data base

Figure 04 – Histogram

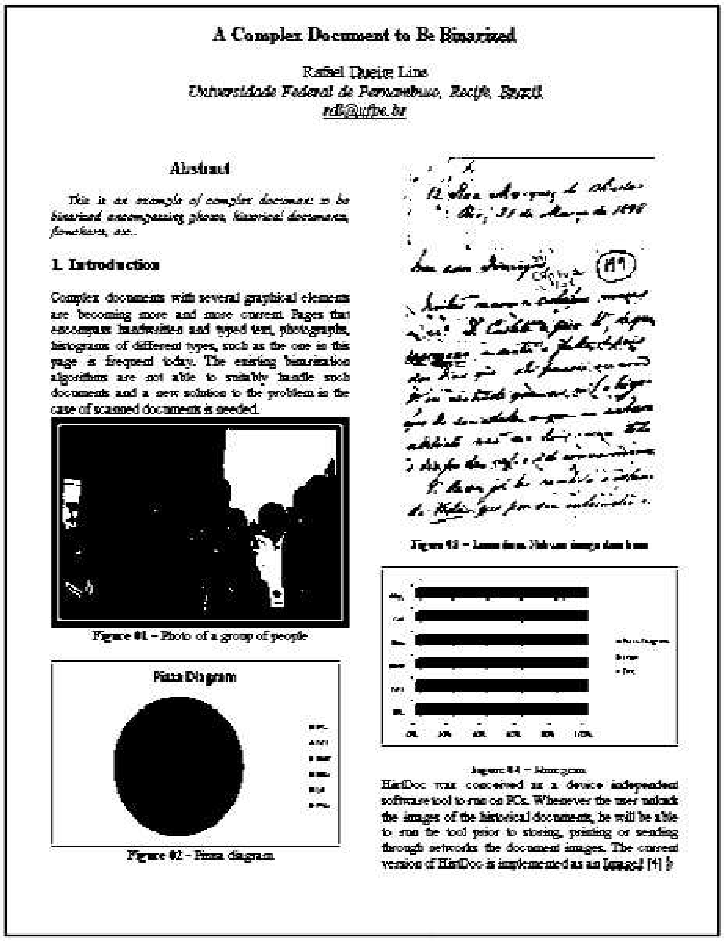
HistDoc was conceived as a device independent software tool to run on PCs. Whenever the user unloads the images of the historical documents, he will be able to run the tool prior to storing, printing or sending through networks the document images. The current version of HistDoc is implemented as an ImageJ [4] B.

Figure 5.1: Example of complex document encompassing text and several graphic elements.

binarize the text area and re-assemble the page with the graphical elements in hues of grey. Such pages claimed far more storage space than the "real" binary ones, but such trick of binarizing the text area and background provided the reader a "continuity effect" as such pages did not differ much from the text-only (binary) pages. Figure 5.3 shows the page shown in Figure 5.1 processed using the LiveMemory scheme (Lins, 2010a).

The data presented in Table 5.1 shows that the LiveMemory processing scheme provides a good image quality and size tradeoff. As the text area is black and all the paper background area is plain white the compressed version of the document is much smaller than the gray-scale equivalent image. The size difference between the LiveMemory synthetic image and the binary (monochromatic) one is a factor of at least 6 times for GIF and 11 times for PNG.

This work presents a binarization scheme suitable for treating scanned complex doc-



A Complex Document to Be Binarized

Rafael Dueire Lins
Universidade Federal de Pernambuco, Recife, Brazil
rdl@ufpe.br

Abstract

This is an example of complex document to be binarized encompassing photos, historical documents, flowcharts, etc..

1. Introduction

Complex documents with several graphical elements are becoming more and more current. Pages that encompass handwritten and typed text, photographs, histograms of different types, such as the one in this page is frequent today. The existing binarization algorithms are not able to suitably handle such documents and a new solution to the problem in the case of scanned documents is needed.

Figure #1 – Photo of a group of people

Pizza Diagram

Figure #2 – Pizza diagram

Figure #3 – [illegible]

Figure #4 – Histogram

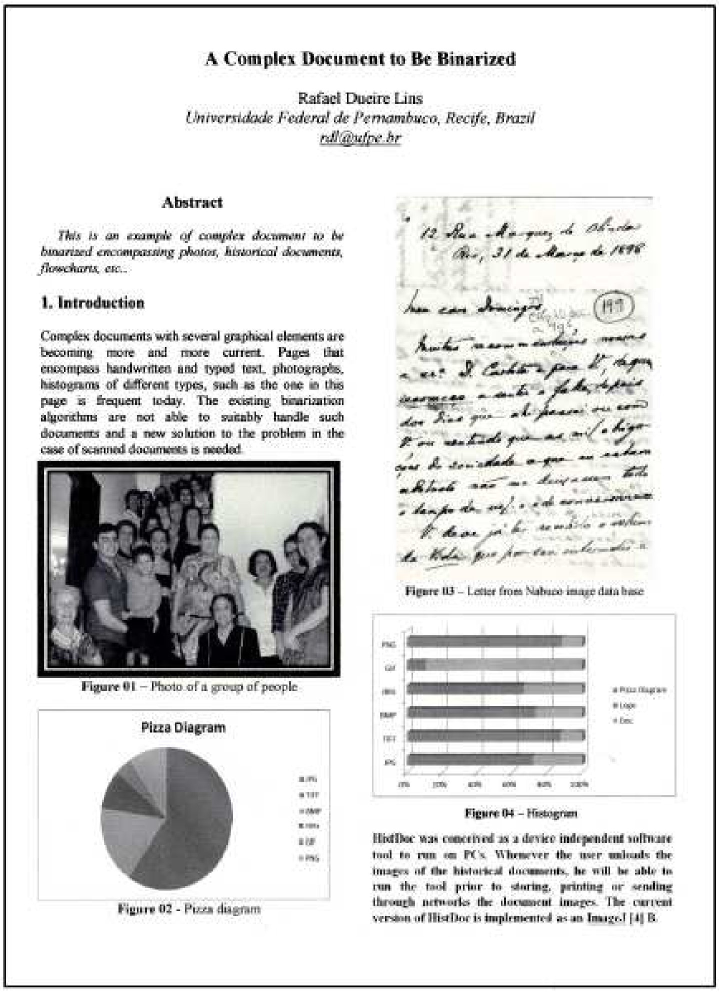
HistDoc was conceived as a device independent software tool to run on PCs. Whenever the user unlock the images of the historical documents, he will be able to run the tool prior to storing, printing or sending through networks the document images. The current version of HistDoc is implemented as an ImageJ [4] p

Figure 5.2: Result of the direct binarization of the document presented in Figure 5.2 using Otsu global binarization algorithm.

Table 5.1: The size in Kbytes of the image in Figure 1.

Kbytes	Bitmap	PNG	TIFF	GIF
True-color	25,509	7,704	8,997	
Gray-scale	8,514	3,461	8,503	903
LiveMemory		1,221	1,552	858
B&W	1,07	104	1,066	130

uments. The scheme automatically decomposes a document image and identifies each of the graphical elements to provide a suitable binarization for each of them independently. In the binarization phase of all the elements, the document image is re-assembled to yield the binary original document. Camera acquired documents have a higher degree of complexity (Silva & Lin, 2007) and are not addressed in this image. The extra complexity is due to uneven illumination that most cameras store images in JPEG file format, a scheme that introduces losses and the JPEG artifact assumes that the document to be processed was scanned in 300 d.p.i resolution and stored in a lossless file format, with no sort of



A Complex Document to Be Binarized

Rafael Dueire Lins
Universidade Federal de Pernambuco, Recife, Brazil
rdl@ufpe.br

Abstract

This is an example of complex document to be binarized encompassing photos, historical documents, flowcharts, etc..

1. Introduction

Complex documents with several graphical elements are becoming more and more current. Pages that encompass handwritten and typed text, photographs, histograms of different types, such as the one in this page is frequent today. The existing binarization algorithms are not able to suitably handle such documents and a new solution to the problem in the case of scanned documents is needed.

Figure 01 – Photo of a group of people

Figure 02 - Pizza diagram

12 Rua Marquez de Olinda
Rio, 31 de Março de 1898

Figure 03 – Letter from Nabuco image data base

Figure 04 – Histogram

HistDoc was conceived as a device independent software tool to run on PC's. Whenever the user unloads the images of the historical documents, he will be able to run the tool prior to storing, printing or sending through networks the document images. The current version of HistDoc is implemented as an ImageJ [4] B.

Figure 5.3: LiveMemory processing scheme used in the document shown in Figure 5.1.

noises (physical, digitalization, storage, etc.) as described reference (Lins, 2009b).

5.2 The New Binarization Scheme

Image de-skew was performed here using the algorithm presented in reference Avila & Lins (2005) and its purpose is to remove skew that may have been introduced during document scanning. Projection profile is a standard technique for spotting image blocks in such kind of document. As projection profile is sensitive to any document skew, the previous step is fundamental for the accuracy of the results of block determination. Block classification is the next step, which is detailed later on. Then each block is independently binarized. Finally, the document is re-assembled with the results of the binary blocks. The operations enumerated above are schematically depicted in Figure 5.4. It is important to say that, although Figure 5.4 presents as if the projection profile and block spotting were performed on the true color version of the image, it is really performed on top of the binary version of documents, obtained by a straightforward binarization process using a threshold algorithm such as Otsu, producing black blocks as the ones presented in Figure 5.2.

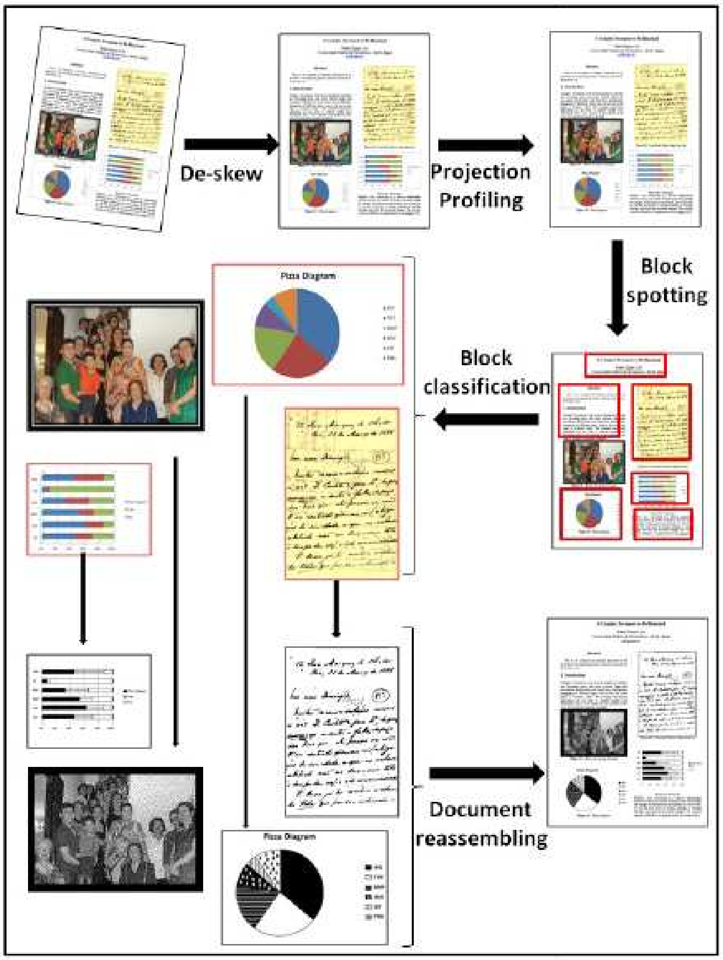


Figure 5.4: Binarization scheme proposed.

The main principle of the scheme proposed here is the recognition that each of the blocks in the complex document has a different nature, and no binarization algorithm that does not take such information into account has the slightest chance of succeeding in keeping the fundamental information of the original document.

5.3 The Block Classifier

The block classifier developed here was based on the one reported in reference Lins (2009a), for improving the quality of image printing of Hewlett-Packard printers, which on its turn is based on the binary classification approach originally described in Simske (2005). It assumes a Gaussian distribution for each of the features, and its performance degrades in proportion to the non-Gaussian nature of the data. The kinds of graphical elements analyzed in this classification level were:

1. Printed text (PT)
2. Cursive text (CT)
3. Photo (Ph)
4. Logo (Lg)
5. Pie or Pizza diagrams (Pie)
6. Histograms or bar diagrams (Hist)
7. Compound complex block/Don't know (Cpd)

The basis of the binarization scheme proposed here is to split the document to be binarized into different blocks, analyze and classify each of them and binarize each of them depending of their nature. The binarization scheme performs thus the following steps:

1. Image de-skew;
2. Projection profile;
3. Block spotting;
4. Block classification;
5. Block binarization;
6. Document re-assembling.

classification:

- Palette (true-color/grayscale).
- Gamut.
- Conversion into Grayscale (if RGB).
- Gamut in Grayscale (if RGB).
- Conversion into Binary (Otsu).
- Mean edge value.
- Number of black pixels in binary image.
- (#Black pixels/Total # pixels)*100%.
- (Gamut/Palette)*100% (true/grayscale).

The classifier works as a set of cascaded random forest binary classifiers as shown in Figure 5.5.

The classifier was trained and tested with images of each of the seven types totaling $30,300$ images (Cpd= 300 images others= $5,000$). The confusion matrix obtained using cross-validation with k-folders= 10 is shown in Table 5.2.

Table 5.2: Confusion Matrix for the random forest cascaded classifier.

	Printed text	Cursive text	Photo	Logo	Pie	Histograms	Compound complex
Printed text	4,966	29	0	4	0	0	1
Cursive text	37	4,958	0	0	0	2	3
Photo	0	0	4,961	36	1	0	2
Logo	0	0	48	4,909	16	15	12
Pie	0	0	3	20	4,971	4	2
Histograms	0	0	0	31	18	4,95	1
Compound complex	0	0	8	2	1	0	289

The classification figures presented in the confusion matrix presented in Table 5.2 shows that the classifier proposed provides good results, yielding a precision of 0.99 for most images and drops down to 0.93 in the case of compound/complex ones Figure 5.6 presents an example of a Compound complex Block, which corresponds to a Brazilian Identification Card, encompassing a photo, a background printed area, stamp, fingerprint on white background, and signature. The direct binarization of such an image, similarly to the image of the document page shown in Figure 5.2 represents a complete loss of the original information. The binarization of such complex block image is left out of the scope of this paper.

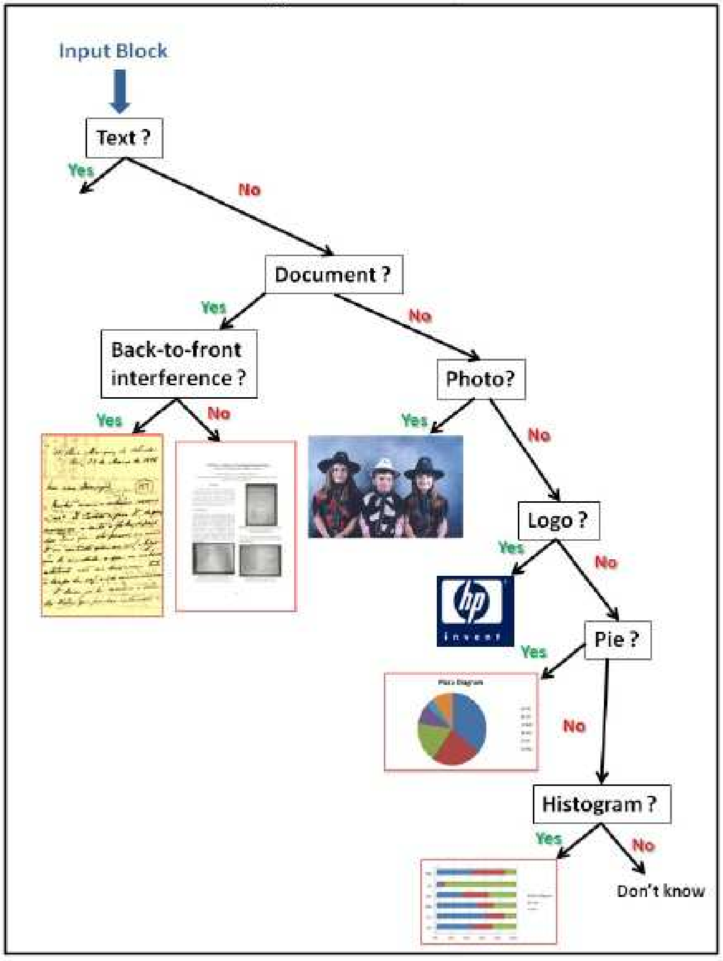


Figure 5.5: Binarization scheme proposed.

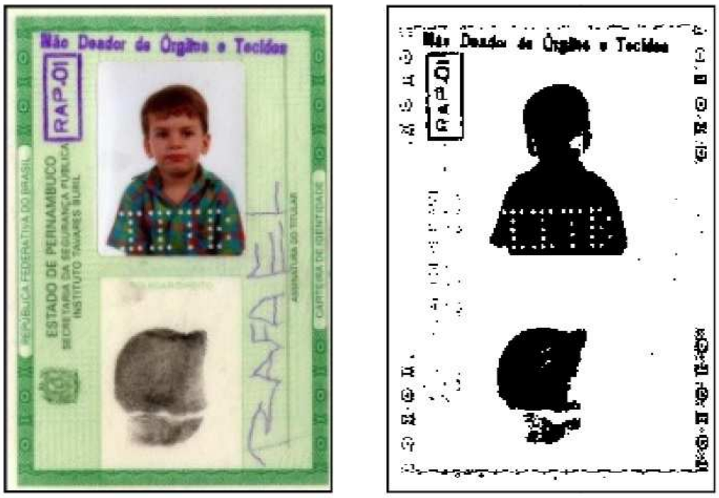


Figure 5.6: Compound complex block encompassing different graphical elements and its binarization using Otsu algorithm.

5.4 Binarizing Photos

As already presented, the direct binarization of photos does not yield an image capable of providing the minimum possible recognition. Image dithering offers a reasonable solution to the problem. Helland (2014) provides a brief explanation and source code for 11 image dithering algorithms. Floyd & Steinberg (1976) dithering is easily the most well-known error diffusion algorithm. It provides reasonably good quality, while only requiring a single forward array (a one-dimensional array with the width of the image, which stores the error values pushed to the next row). Additionally, because its divisor is 16, bit-shifting can be used in place of division making it quite fast, even on old hardware. Figure 5.7 zooms into the photo presented in Figure 5.1. The direct application of Floyd-Steinberg algorithm to it yielded the monochromatic image presented in Figure 5.8 in which one may observe that most of the information of the original content was preserved.

The binary photo presented in Figure 5.7 has over 75% of its area of relevant information, exhibiting very little in background. Most of times, there is no relevant information in the background that is converted into noise in the binary image. The goal is to find a new "intelligent" dithering algorithm that takes into account the content of the photo to be binarized. In such a way, the photo will be classified as people, landscape, object, and complex. Whenever a photo is recognized as being of an object the background is completely removed using an algorithm similar to the one described in reference Lins & Silva (2013) prior to dithering.

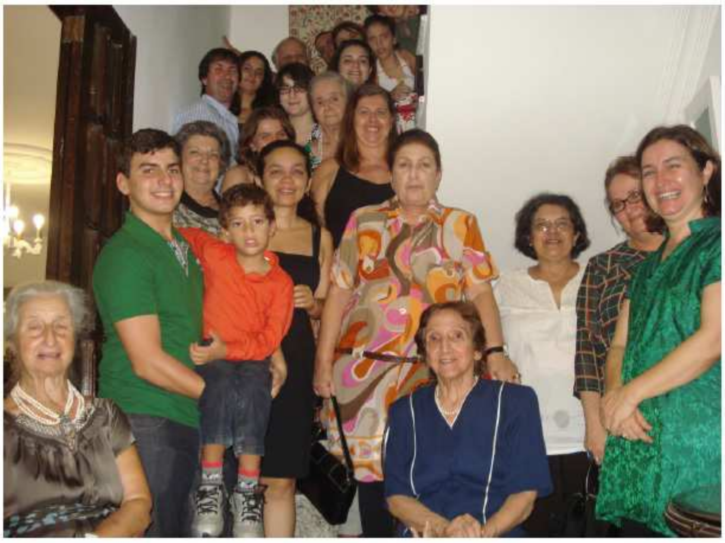


Figure 5.7: Photo of a group of people presented in the document page in Figure 5.1.

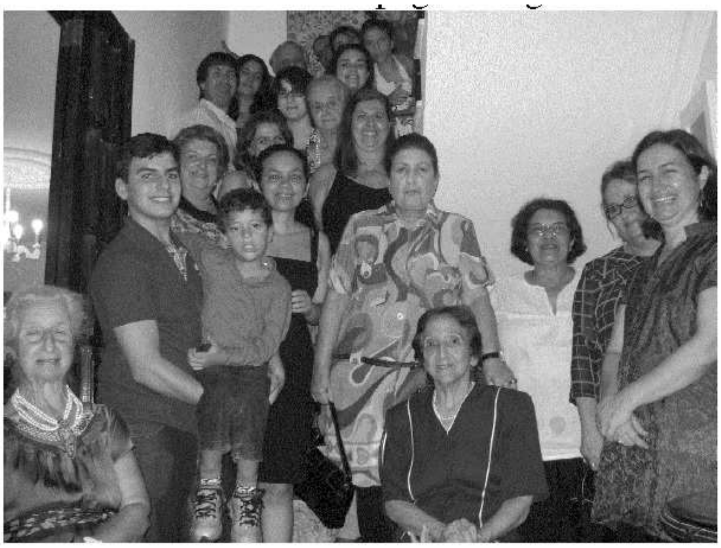


Figure 5.8: Photo of a group of people presented in the document page in Figure 5.1.

5.5 Binarizing Logos

Logos tend to use a few plain colors. The binarization of logos may be a straightforward process in the case of single color logos or may claim for a more complex strategy in the case of using multiple colors. Figure 5.9 presents some logos and their direct binarization using Otsu algorithm. The five logos presented in Figure 5.9 show that the direct binarization using Otsu algorithm yields results that preserve the original information in most cases (the logo of “BRASIL” has un unusual number of colors). It is important to remark that

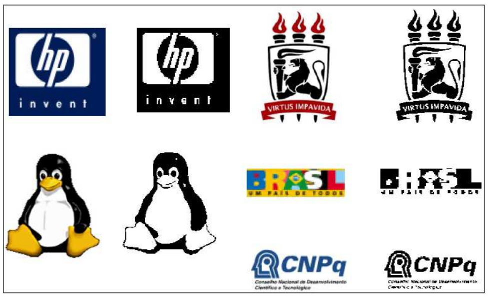


Figure 5.9: Five different logos and their binarization using Otsu algorithm.

the block splitting strategy used here was important to provide the satisfactory results shown. The use of a global threshold binarization algorithm if applied to the whole document page may provide a different threshold value that may destroy part of the information.

5.6 Binarizing Documents

The binarization of text documents either printed or handwritten has a long tradition in document engineering. Document aging, back-to-front interference (bleeding) and some physical noises Lins (2009b) may bring an extra degree of difficulty to this task. In this work, the algorithm described in reference Monte da Silva *et al.* (2008) was used to process the handwritten historical document shown. The authors are currently working on a document classifier that refines the different kinds of documents, making possible to better select and tune the binarization algorithm for documents.

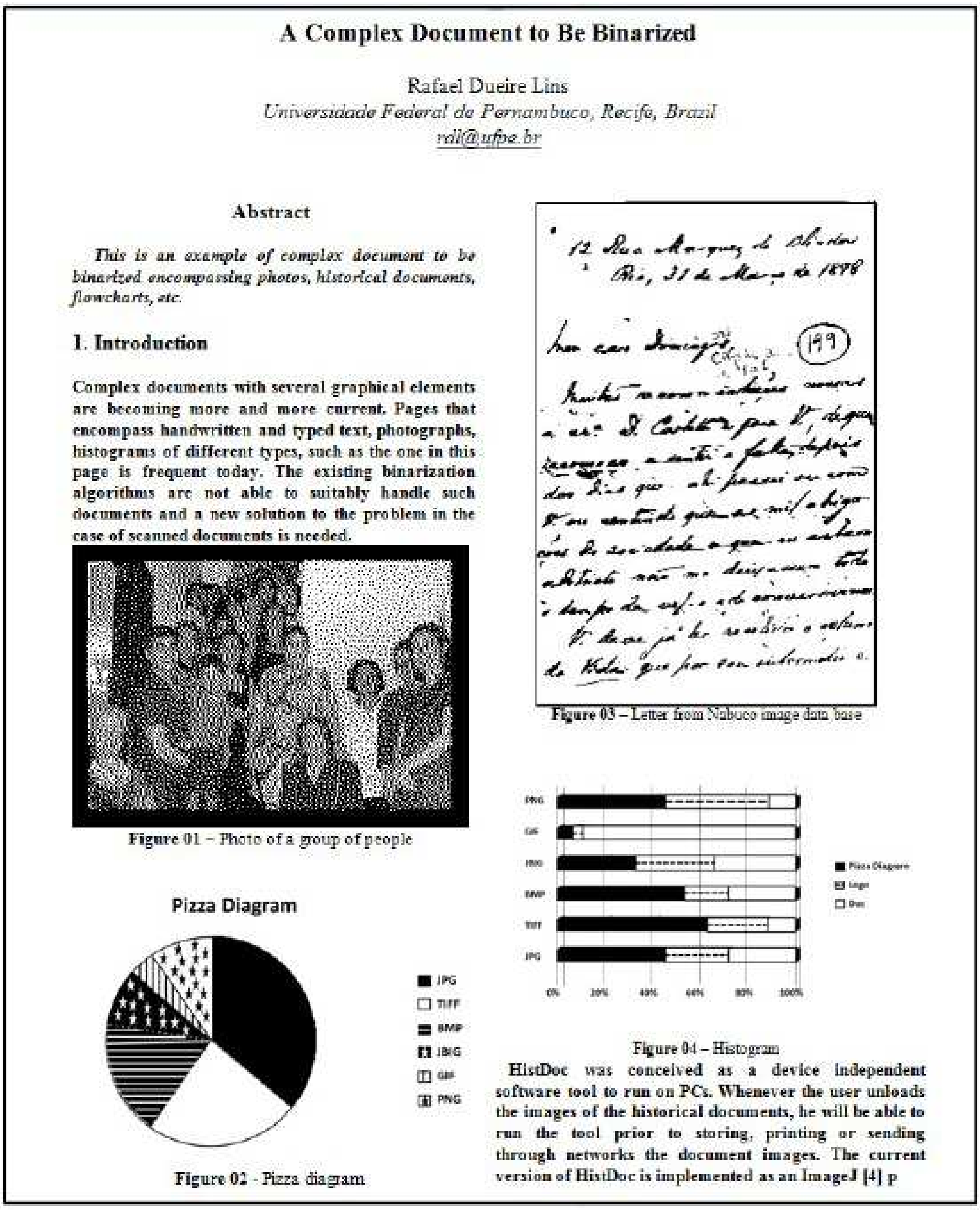
5.7 Binarizing Pie Diagrams and Histograms

Pie (or Pizza) diagrams and Histograms tend to be automatically generated by spreadsheet tools such as Microsoft Excel. As already shown, the direct binarization of such diagrams yield a complete loss of the original information. The only alternative is recognizing the different elements in the image and generating a synthetic monochromatic image that conveys the equivalent information. Different textures are used to represent the different colors in the diagram. In both kinds of diagrams the contour must be drawn first, together with the color separation areas. The largest slice in the pie is painted black, the second largest is painted white, the third largest is replaced by a texture, etc.

5.8 Conclusions

The widespread use of electronic editing tools in the last three decades has yielded documents with several graphical elements incorporated. On the other hand, storage capacity in most organizations cannot meet the demand created and the binarization of such documents is still an option to save and distribute their information. No binarization algorithm is good enough to work suitably in all sorts of documents. This work proposed a new strategy to improve the binarization of complex documents that splits it into different blocks and classify them individually. Once the image blocks were suitably recognized and binarized using the most adequate strategy to the block nature, the image is reassembled yielding the monochromatic version of the original document.

The use of the scheme presented here applied in the document image shown in Figure 5.1 yielded the image shown in Figure 5.10 (the pie and histogram image blocks were enhanced manually because page scaling-down would not allow to distinguish the texture of the different areas). The size of the block assembled image is (in Kbytes): $1,858$ bmp, 132 tiff, 122 png, 58 gif. Notice that the new monochromatic image in gif is less than half of the size than the original image binarized directly using Otsu algorithm, but keeps the original information elements.



A Complex Document to Be Binarized

Rafael Dueire Lins
Universidade Federal de Pernambuco, Recife, Brazil
rdl@ufpe.br

Abstract

This is an example of complex document to be binarized encompassing photos, historical documents, flowcharts, etc.

1. Introduction

Complex documents with several graphical elements are becoming more and more current. Pages that encompass handwritten and typed text, photographs, histograms of different types, such as the one in this page is frequent today. The existing binarization algorithms are not able to suitably handle such documents and a new solution to the problem in the case of scanned documents is needed.

Figure 01 – Photo of a group of people

Figure 03 – Letter from Nabuco image data base

Pizza Diagram

Figure 02 - Pizza diagram

Figure 04 – Histogram

HistDoc was conceived as a device independent software tool to run on PCs. Whenever the user unloads the images of the historical documents, he will be able to run the tool prior to storing, printing or sending through networks the document images. The current version of HistDoc is implemented as an ImageJ [4] p

Figure 5.10: Final Block binarized image.

6 CONCLUSIONS AND LINES FOR FURTHER WORK

The contributions of this thesis are outlined here together with some research directions that may be followed as other unfolding of the results obtained.

Chapter 2 analyzed the consistency of the assessment for senior researchers in Computer Science of CNPq, the Brazilian Research Council. Several statistical classification strategies were considered using international public databases and were compared with the CNPq classification.

The most relevant results in such analysis listed in this thesis are:

- ArnetMiner and Google Scholar are the most representative databases for the senior researchers in Computer Science in Brazil, as they convey the largest number of publications per researcher.

- Taking as reference the ArnetMiner database, the average volume of publications by the researchers in the categories $1A$ and $1C$ from the Northeast is higher than the one in the other regions of Brazil.

- Using the discriminant analysis approach, and making a comparison between the various databases, we found that the data from the Google Scholar database has the highest proportion of correct hits as 85.2%, for the researchers in classifications $1A$, $1B$ and $1C$, although the number of samples is small.

- The analysis of the Lattes and ArnetMiner databases provides equivalent results of about 60.0%. The worst hit rate is obtained with the Web of Science database, 43.6%. The highest proportion of classification similarity is using the data provided by the Google Scholar database in the case of the researchers in the $1A$ group, 100.0%. The worst classification results is obtained with the Web of Science database in group $1C$, 20.6%.

- The discriminant analysis method was effective in identifying the appropriate database for the classification and ordering of researchers for $1A$, $1B$ and $1C$ levels, given the correct percentage of correlation to the CNPq classification. In this case, the ArnetMiner database represents the production of CNPq researchers correctly as well.

Considering a set of all databases, the method remained robust with satisfactory results.

- When taking into consideration the geographic distribution of the researchers using the CNPq classification is uneven, with the Southeast region of Brazil with about 70.00% of all senior researchers, while the Notheast region represents only 7.00% of the total number of researchers. An important result obtained is that, the analysis of the scientific production, allows to say that overall there was **no discrimination** against the researchers in the regions with less senior researchers, and that one may even, contrary to the general belief, that some of the researchers from the less represented regions suffered *positive discrimination*, being ranked above their peers in the Southeast.

- The criteria presented by CNPq does not differ much from the results presented by the methods described, i.e., discriminant analysis and k-mean.

- With respect to the ArnetMiner database, it showed a consistency in the set of given variables, considering the volume of data as well as the representative variables able to generate a clustering and reclassification models.

As an overall conclusion, one way say that the CNPq rank is fair in general terms as the researcher classification distortions are small.

The Chapter 3 analyzed nine types of filters to binarize document images with different degrees of strength of the back-to-front interference, totaling 160 synthetic images as input. The output images generated a database of 1,440 binary documents. The matrix of co-occurrence probabilities and visual inspection were used to evaluate the quality of the resulting images, from which one may draw the following conclusions:

- Of the 16 historical documents analyzed, 12 could have some treatment by the studied filters, under the imposed conditions of the co-occurrence probabilities $P(b|b) \geq 99.00\%$ and P $(f|f) \geq 99.00\%$ shown as an example in Section 3.4. In each binarized document one has one or more filters selected with the optimum threshold and corresponding co-occurrence probabilities $P(b|b)$ and $(f|f)$, providing a measure of the quality of the binarized images. Four historical documents were not able to meet the above conditions.

- The Pun method was not efficient in filtering the 160 images generated synthetically, because the co-occurrence probability $P(b|f)$ or maximum error was above 32.00%. The visual inspection of the binarized documents also corroborate to this result.

- With the visual inspection of the binarized images studied, as well as the matrix co-occurrence probability, one finds that the remaining algorithms are good thresholding methods to remove back-to-front interference in documents, depending on the texture of the historical document and the intensity of the back-to-front interference.

- The filters that met the requirements of co-occurrence probabilities in the given example were: Isodata, Otsu, Kapur-Sahoo-Wong, Silva-Lins-Rocha and Mello-Lins. The opacity factor value α varied from 0.7 to 1.0.

A new filter for back-to-front interference removal is proposed in Chapter 4. Tested with the 160 synthetic documents described in Chapter 3, it produced good-quality binary images that meet the imposed conditions of the co-occurrence probabilities $P(b|b) \geq$ 99.00% and P $(f|f) \geq$ 99.00%. The visual inspection of the binarized images studied confirmed the results obtained for the matrix co-occurrence probability.

Assessing the newly proposed filter together with the ones in Chapter 3 one may draw the following conclusions:

- Two of the test documents, the ones with texture (R=224, G=174, B=113) and (R=218, G=167, B=113), only provided reasonable results when binarized by the proposed filter, given the efficiency requirements imposed that the error is less of 1.00%

- For each binarized document one has one or more filters selected with the optimum threshold and corresponding co-occurrence probabilities $P(b|b)$ and $(f|f)$, providing a measure of the quality of the binarized images. The remaining two historical documents the filters were not able to meet the above conditions.

- The filters that met the requirements of co-occurrence probabilities in the given example were: the Proposed Filter, Isodata, Otsu, Kapur-Sahoo-Wong, Silva-Lins-Rocha and the Mello-Lins. The opacity factor value α varied from 0.4 to 1.0, because of the inclusion of the proposed filter in the inclusion to remove back-to-front interference and the binarization processes.

- The proposed filter performed better than all nine filters analyzed with respect to the alpha range, in some historical documents reaching the range of 0.4 to 1.0, when the co-occurrence probability P(b|b) and P(f|f) are above of 99.00%.

Chapter 5 proposes the "intelligent" binarization of complex documents, a document which encompasses several different graphical elements besides text, such as photos, logos, diagrams, pie charts, etc. This work proposes a new strategy to improve the binarization of complex documents that splits it into different blocks and classify them individually.

The basis of the *semantic binarization* scheme proposed here is to split the document to be binarized into different blocks, analyze and classify each of them and binarize each of them depending of their nature.

Some of the lines for further work along the lines of this thesis are:

- Refining the degree of the back-to-front intensity (α) in the critical regions to better "understand" the behavior of the different filters.

- In this work, the opacity factor α ranged from 0.1 to 1.00 in steps of 0.1. One alternative would be to do in smaller steps, especially around values where the of co-occurrence probability are more tolerable.

- Including new filters in the assessment.

- Consider other kinds of textures of historical documents.

- Developing several new algorithms for semantic binarization.

References

Abburu, S.; Golla, S. B. Satellite Image Classification Methods and Techniques: A Review. *International Journal of Computer Applications*, v. 119, n. 8, p. 20–25, 2015.

Abramson, N. *Information Theory and Coding*. Mc Graw-Hill Book Company, 1963.

Academic, Microsoft. *Academic Social Networks*. http://academic.research.microsoft.com/, accessed: December the 3rd, 2013.

Academic, Microsoft. *Academic Social Networks*. http://academic.research.microsoft.com/, accessed: November the 15th, 2016.

Al-Hinnawi, A. R.; Daer, M. Assessment of bilateral filter on low NEX open MRI views. *SIViP*, 9–17, 2015.

Anderson, T. W. *An Introduction to Multivariate Analysis*. 2. ed. error, 1984. v. 1.

ArnetMiner. *Academic Social Networks*. https://aminer.org, accessed: November the 18th, 2013.

ArnetMiner. *Academic Social Networks*. https://aminer.org, accessed: November the 15th, 2016.

Aurich, V.; Weule, J. Non-linear Gaussian Filters Performing Edge Preserving Diffusion. *Proceedings of the DAGM symposium*, 1995.

Avila, B. T.; Lins, R. D. A Fast Orientation and Skew Detection Algorithm for M. Document Images. *ACM DocEng*, 118–126, 2005.

Ball, G. H.; Hall, D. J. A Novel Method of Data Analysis and Pattern Classification. *Information Science Branch Office Naval Research - Technical Report*, 1965.

BARATA, R. B.; GOLDBAUM, M. A Profile of Researchers in Public Health with Productivity Grants from the Brazilian National Research Council. *CNPq*, v. 19, n. 6, p. 1863–1876, 2003.

COSTAS, R.; BORDONS, M. Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, v. 77, n. 2, p. 267–288, 2008.

CRAMMER, J. S. *Logit Models from Economics and other Fields.* Cambridge University Press, 2003.

DAWOUD, A.; KAMEL, M. S. Iterative Model-Based Binarization Algorithm for Cheque Images. *International Journal on Document Analysis and Recognition*, v. 5, p. 28–38, 2002.

DAWOUD, A.; KAMEL, M. S. Iterative Multimodel Subimage Binarization for Handwritten Character Segmentation. *IEEE Transactions on Image Processing*, v. 13, n. 9, p. 1223–1230, 2004.

EGGHE, L. Theory and practise of the g-index. *Scintometrics*, v. 69, n. 1, p. 131–152, 2006.

EGGHE, L. An Improvement of the h-Index: The g-Index, 2007.

FLOYD, R. W.; STEINBERG, L. An Adaptative Algorithm for Spatial Greyscale. *Proceddings SID*, v. 17, n. 2, p. 75–77, 1976.

FUNDAJ. *Joaquim Nabuco Fundation.* http://www.fundaj.gov.br, accessed: february the 3rd, 2016.

GUPTA, M. R.; JACOBSON, N. P.; GARCIA, E. K. OCR binarization and image preprocessing for searching historical documents. *The Journal of the Pattern Recognition Society*, v. 40, p. 389–397, 2007.

HAIR, JOSEPH F. *Multivariate Data Analysis.* 7. ed. error, 2009. v. 1.

HAMPSON, G. *Especialista avalia mudanças na publicação de revistas científicas.* http://agencia.fapesp.br/especialista avalia mudancas na publicacao de revistas cientificas/24548/, accessed: January the 5th, 2017.

HELLAND, T. *Image Dithering.* http://www.tannerhelland.com/4660/dithering-eleven-algorithms-source-code/, december, 23, 2014.

HIRSCH, J. E. An Index to Quantify an Individual's Scientific Research Output. *Physics Society*, September, 2005.

JOHANNSEN, G; BILLE, J. A Threshold Selection Method Using Information Measure. *ICPR'82 - Proceeding 6th International Conference on Pattern Recognition*, 140–143, 1982.

JOHNSON, RICHARD A. ; WICHERN, DEAN W. *Applied Multivariate Statistical Analysis.* 6th. ed. Pretince Hall, 2007. v. 1.

KAPUR, J. N.; SAHOO, P. K.; WONG, A. K. C. A New Method for Gray-Level Picture Thersholding Using the Entropy of the Histogram. *Computer Vision Graphics and Image Processing*, v. 29, p. 273–285, 1985.

KASTURI, R.; O'GORMAN, L.; GOVINDARAJU V. Document image analysis: A primer. *Sadhana*, 3–22, 2002.

KAVALLIERATOU, E.; ANTONOPOULOU, H. Cleaning and Enhancing Historical Document Images. *Intelligent Vision Systems*, 681–688, 2005.

LABBÉ, C. Ike Antkare, One of the Great Stars in the Scientific Firmament. *International Society for Scientometrics and Informetrics Newsletter*, v. 6, n. 2, p. 48–52, 2010.

LATTES. *Academic Social Networks.* http://lattes.cnpq.br/, accessed: November the 20th, 2013.

LATTES. *Academic Social Networks.* http://lattes.cnpq.br/, accessed: November the 15th, 2016.

LEEDHAM, G.; VARMA S.; PATANKAR A.; GOVINDARAJU V. Separating Text and Background in Degraded Document Images - A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding. *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR02)*, 244–249, 2002.

LINS, R. D.; SILVA, G. F.; SIMSKE S. J.; FAN J.; SHAW M.; SA P.; THIELO M. Image Classifacation to Improve Printing Quality od Mixed-Typed Documents. *ICDAR 2009*, 1106–1110, 2009a.

LINS, R. D.; SILVA, G. P.; TORREÃO G.; ALVES N. F. Efficiently Generating Digital Libraries of Proceedings with The LiveMemory Platform. *The 7th International Telecommunications Symposium (ITS 2010)*, 2010a.

LINS, R. D.; SILVA, G.; SILVA J. M. Assessing Strategies to Remove Back-to-Front Interference in Color Documents. In: RESEARCHGATE (Ed.), INTERNATIONAL TELECOMMUNICATIONS SYMPOSIUM, Septemberb, 2010.

LINS, R. D. A Taxonomy for Noise Detection in Images of Papers Documents-The Physical Noises. *Spring Verlag*, v. 4627, p. 844–854, 2009b.

LINS, R. D.; SILVA, G.F. Removing Shade and Specular Noise by using Multiplr Images od Objects and Documents. *Springer Verlag*, 70–79, 2013.

LINS, R. D. ET AL. An Environment for Processing Images of Historical Documents. *Microproc. and Microprogramming*, 111–121, 1995.

LIU, Y.; SRIHARI, S. N. Document Image Binarization Based on Texture Features. *IEEE Transaction on Pattern Analysis and Machine Inteligence*, v. 19, n. 5, p. 540–544, 1997.

MELLO, C. A. B.; SANCHEZ, A.; OLIVEIRA A. L. I. Image Thresholding of Historical Documents: Application to the Joaquim Nabuco's File. *Eva Vienna*, 115–122, 2006.

MELLO, C. A. B.; LINS, R. D. Image segmentation of historical documents. *Visual 2000*, 2000.

MELLO, C. A. B.; LINS, R. D. Generation of Images of Historical Documents by Composition. *Proceedings of the 2002 ACM symposium on Document engineering*, 127–133, 2002.

MEMARSADEGHI, N.; MOUNT, D. M.; NETANYAHU N. S.; MOIGNE J. A Fast Implementation of the IsoData Clustering Algorithm. *International Journal of Computational Geometry and Applications*, 71–103, 2007.

Monte da Silva, J. M.; Lins, R. D.; Martins, F. M. J.; Wachenchauzer, R. A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference. *Journal of Universal Computer Science*, v. 14, n. 2, p. 293–313, 2008.

of Science, Web. *Academic Social Networks.* https://webofknowledge.com, accessed: Novembr the 20th, 2013.

of Science, Web. *Academic Social Networks.* https://webofknowledge.com, accessed: Novembr the 15th, 2016.

Oha, H.; Limb, K.; Chienc S. An improved binarization algorithm based on a water flow model for document image with inhomogeneous background. *Pattern Recognition*, 2612–2625, 2005.

Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transaction on Systems, Man and Cybernetics*, v. SMC-9, n. 1, p. 62–66, 1979.

Paris, S.; Durand, F. A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach. *International Journal of Computer Vision*, v. 1, n. 81, p. 24–52, 2009.

Pun, T. Entropic Thresholding, A New Approach. *Computer Vision Graphics and Image Processing*, 210–239, 1981.

Richards, J. A. *Remote Sensing Digital Image Analysis: An Introduction.* second. ed., 1993.

Roe, Edward; Mello, Carlos A. B. Binarization of Color Historical Document Images Using Local Image Equalization and XDoG. *12th International Conference on Document Analysis and Recognition*, August, p. 205–209, 2013.

Scholar, Google. *Academic Social Networks.* https://scholar.google.com.br, accessed: December the 2nd, 2013.

Scholar, Google. *Academic Social Networks.* https://scholar.google.com.br, accessed: November the 15th, 2016.

SCOPUS. *Academic Social Networks*. https://www.elsevier.com/solutions/scopus, accessed: November the 20th, 2013.

SCOPUS. *Academic Social Networks*. https://www.elsevier.com/solutions/scopus, accessed: November the 15th, 2016.

SEZGIN, M.; SANKUR, B. A Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. *Journal of Electronc Imaging*, v. 1, n. 13, p. 146–165, 2004.

SHARMA, G. Show-trough cancellation in scans of duplex printed documents. *IEEE Transaction Image Processing*, v. 10, n. 5, p. 736–754, 2001.

SILVA, G. P.; LIN, R. D. PhotoDoc: A Toolbox for Processing Document Images Acquired Using Portable Digital Cameras. *CBDAR 2007*, 107–114, 2007.

SILVA, J. M. M.; LINS, RAFAEL D.; ROCHA, VALDEMAR C. Binarizing and Filtering Historical Documents with Back-to-Front Interference. *Proceedings of the 2006 ACM symposium on Applied Computing*, 853–858, 2006.

SIMSKE, S. J. Low Resolution Photo/drawing Classification: metrics, method and archiving optimization. *IEEE ICIP*, 534–537, 2005.

TAN, C. L.; CAO, R.; SHEN P. Restoration of archival documents using a wavelet technique. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, v. 24, n. 10, p. 1399–1404, 2002.

TOMASI, C.; MANDUCHI, R. Bilateral Filtering for Gray and Color Images. *Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India*, 1998.

VALIZADEH, M.; KABIR, E. Binarization of degraded document image based on feature space partitioning and classification. *IJDAR*, v. 15, p. 57–69, 2012.

WANG, Q.; TAN, C. L. Matching of Double-Sided Document Images to Remove Interference. *IEEE CVPR 2001*, 1084–1089, 2001.

WESZKA, J. S.; ROSENFELD, A. Histogram Modification for Threshold Selection. *IEEE Transaction on Systems, Man and Cybernetics*, v. SMC-9, n. 1, p. 38–52, 1979.

Winnemöller, H. XDoG: advanced image stylization with eXtended Difference of Gaussians. *Proceedings of the ACM SIGGRAPH-Eurographics Symposium on Non-Photorealistic Animation and Rendering*, 147–156, 2011.

Wu, L. U.; Songde, A.; Haqing L. U. An Effective Entropic Thresholding for Ultrasonic Imaging. *International Conference Pattern Recognition*, 1522–1524, 1998.

Yadav, J.; Sharma, M. A Review of K-mean Algorithm. *International Journal of Engineering Trends and Technology (IJETT)*, v. 4, n. 7, p. 2972–2976, 2013.

Yen, J. C.; Chang, F. J.; Chang S. A New Criterion for Automatic Multilevel Thresholding. *IEEE Transaction Image Process IP-4*, 370–378, 1995.

APPENDIX A

Bilateral Filtering for Gray and Color Images

The bilateral filter was first introduced by Aurich & Weule (1995) under the name “nonlinear Gaussian filter”. It was rediscovered later by Tomasi & Manduchi (1998) who called it the “bilateral filter” which is now the most commonly used name according to Paris & Durand (2009). The bilateral filter is technique to smooth images while preserving edges. The filter output at each pixel is a weighted average of its neighbors. The weight assigned to each neighbor decreases with both the distance values among pixels of the image plane (the spatial domain S) and the distance on the intensity axis (the range domain R). The filter applies spatial weighted averaging without smoothing edges. The mathematical expression, Equation 6.0.7 is achieved by combining two Gaussian filters; one filter works in the spatial domain, the other filter works in the intensity domain. Therefore, not only the spatial distance but also the intensity distance is important for the determination of weights. At a pixel location (x, y), the output of a bilateral filter can be formulated as follows:

Bilateral filter combines two stages of filtering. These are the geometric closeness (i.e., filter domain) and the photometric similarity (i.e., filter range) among pixels in an $N \times N$ window size. The bilateral filter mathematics is described in Equation 6.0.7.

$$I_{BF}(x,y) = \frac{1}{K} \sum_{x,y=(\hat{x},\hat{y})-N}^{(\hat{x},\hat{y})+N} \exp - \frac{||x-\hat{x}||^2 + ||y-\hat{y}||^2}{2\sigma_d^2} \exp - \frac{(I(x,y) - I(\hat{x},\hat{y}))^2}{2\sigma_r^2}, \quad (6.0.1)$$

where $I(x, y)$ is the pixel intensity in the image before applying the bilateral filter, $I_{BF}(x, y)$ is the resulting pixel intensity after applying the bilateral filter, $(\hat{x}, \hat{y})$ is the coordinates of the pixels encompassed in the bilateral filter window, K is a normalization constant given by Equation 6.0.8:

$$K = \sum_{x,y=(\hat{x},\hat{y})-N}^{(\hat{x},\hat{y})+N} \exp - \frac{||x-\hat{x}||^2 + ||y-\hat{y}||^2}{2\sigma_d^2} \exp - \frac{(I(x,y) - I(\hat{x},\hat{y}))^2}{2\sigma_r^2}. \tag{6.0.2}$$

These Equations show that the bilateral filter has three parameters. The first parameter is the σ_d which defines the filter domain, as illustrated in equation 6.0.9, whereas equation 6.0.10 states that the filter range is defined by the second parameter denoted as σ_r. The third parameter is the bilateral filter window size $[N \times N]$.

$$\exp - \frac{||x-\hat{x}||^2 + ||y-\hat{y}||^2}{2\sigma_d^2} \tag{6.0.3}$$

$$\exp - \frac{(I(x,y) - I(\hat{x},\hat{y}))^2}{2\sigma_r^2} \tag{6.0.4}$$

The geometric spread of the bilateral filter is controlled by σ_d. As σ_d is increased, more neighbors are combined for diffusion resulting in higher smoothening, while σ_r representing the photometric spread of the bilateral. Only pixels with a percentage difference of less than σ_r are processed, while those higher than σ_r are not, according Al-Hinnawi & Daer (2015).

Otsu (1979) considered the pixels of a given picture be represented in L gray levels $[1, 2, \ldots, L]$. The number of pixels at level i is denoted by n_i and the total number of pixels by $N = n_1 + n_2 + \cdots + n_L$. In order to simplify the discussion, the gray-level histogram is normalized and regarded as a probability distribution:

$$p_i = \frac{n_i}{N}, \qquad p_i \ \geq 0, \qquad \sum_{i=1}^{L} p_i = 1. \tag{6.0.5}$$

Otsu (1979) dichotomized the pixels into two classes C_0 and C_1 (background and objects, or vice versa) by a threshold at level k; C_0 denotes pixels with levels $[1, 2, \ldots, k]$, and C_1 denotes pixels with levels $[k+1, \ldots, L]$. Then the probabilities of class occurrence and the class mean levels, respectively, are given by

$$\omega_0 = Pr(C_0) = \sum_{i=1}^{k} p_i = \omega_k, \tag{6.0.6}$$

$$\omega_1 = Pr(C_1) = \sum_{i=k+1}^{L} p_i = 1 - \omega_k, \tag{6.0.7}$$

and

$$\mu_0 = \sum_{i=1}^{k} iPr(i|C_0) = \sum_{i=1}^{k} \frac{ip_i}{\omega_0} = \frac{\mu(k)}{\omega(k)}, \tag{6.0.8}$$

$$\mu_1 = \sum_{i=k+1}^{L} iPr(i|C_1) = \sum_{i=k+1}^{L} \frac{ip_i}{\omega_1} = \frac{\mu(\tau) - \mu(k)}{1 - \omega(k)}, \tag{6.0.9}$$

where

$$\omega_k = \sum_{i=1}^{k} p_i, \tag{6.0.10}$$

and

$$\mu_k = \sum_{i=1}^{k} ip_i. \tag{6.0.11}$$

The Equations 6.0.10 and 6.0.11 are the $zero^{th}$ and the first-order cumulative moments of the histogram up to the k^{th} level, respectively, and

$$\mu_\tau = \mu_L = \sum_{i=1}^{L} ip_i \tag{6.0.12}$$

is the total mean level of the original picture. For any choice of k, we have:

$$\omega_0\mu_0 + \omega_1\mu_1 = \mu_\tau, \qquad \omega_0 + \omega_1 = 1. \tag{6.0.13}$$

The class variances are given by

$$\sigma_0^2 = \sum_{i=1}^{k} (i - \mu_0)^2 Pr(i|C_0) = \sum_{i=1}^{k} (i - \mu_0)^2 \frac{p_i}{\omega_0}, \tag{6.0.14}$$

$$\sigma_1^2 = \sum_{i=k+1}^{L} (i - \mu_1)^2 Pr(i|C_1) = \sum_{i=k+1}^{L} (i - \mu_1)^2 \frac{p_i}{\omega_1}. \tag{6.0.15}$$

These require second-order cumulative moments.

Otsu also introduced the following discriminant criterion measures used in the discriminant analysis:

$$\lambda = \frac{\sigma_B^2}{\sigma_W^2}, \qquad \kappa = \frac{\sigma_\tau^2}{\sigma_W^2}, \qquad \eta = \frac{\sigma_B^2}{\sigma_\tau^2}, \tag{6.0.16}$$

where

$$\sigma_W^2 = \omega_0\sigma_0^2 + \omega_1\sigma_1^2 \tag{6.0.17}$$

$$\sigma_B^2 = \omega_0(\mu_0 - \mu_\tau)^2 + \omega_1(\mu_1 - \mu_\tau)^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2. \tag{6.0.18}$$

Due to Equation 6.0.13, this becomes the function object goal (the criterion measure) of an optimization problem, and

$$\sigma_\tau^2 = \sum_{i=1}^{L}(1 - \mu_\tau)^2 p_i. \tag{6.0.19}$$

This standpoint is motivated by a conjecture that well-thresholded classes would be separated in gray levels, and conversely, a threshold giving the best separation of classes into gray levels would be the best threshold.

$$\sigma_W^2 + \sigma_B^2 = \sigma_\tau^2. \tag{6.0.20}$$

Both σ_W^2 and σ_B^2 are functions of threshold level k, but σ_τ^2 is independent of k. It is also noted that σ_W^2 is based on the second-order statistics (class variances), while σ_B^2 is based on the first-order statistics (class means). Therefore, η is the simplest measure with respect to k. Thus we adopt η as the criterion measure to evaluate the "goodness" (or separability) of the threshold at level k. The optimal threshold k^* that maximizes η, or equivalently maximizes σ_B^2 is selected in the following sequential search by using the simple cumulative quantities 6.0.10 and 6.0.11, or explicitly using the difference of equations 6.0.6 - 6.0.9:

$$\eta(k) = \frac{\sigma_B^2}{\sigma_\tau^2} \tag{6.0.21}$$

$$\sigma^2(k) = \frac{[\mu_\tau \omega(k) - \mu(k)]^2}{\omega(k)[1-\omega(k)]} \tag{6.0.22}$$

and the optimal threshold k^* is:

$$\sigma_B^2(k^*) = \underset{\{1 \leq k < L\}}{Max} \sigma_B^2(k) \tag{6.0.23}$$

The maximum value $\eta(k^*)$, denoted simply by η^*, can be used as a measure to evaluate the separability of classes (or ease of thresholding) for the original picture or the bimodality of the histogram.

The effective range of the gray-level histogram is

$$S^* = \{k; \omega_0\omega_1 = \omega(k)[1-\omega(k)]\} > 0, \qquad or \qquad 0 < \omega_k < 1\}. \tag{6.0.24}$$

From the definition in 6.0.18, the criterion measure σ_B^2 (or η) takes a minimum value of zero for such k as $k \in S - S^* = \{k; \omega(k) = 0 \quad or \quad 1\}$ (i.e., making all pixels either C_1 or C_0, which is, of course, not our concern) and takes a positive and bounded value for $k \in S^*$. It is, therefore, obvious that the maximum always exists. All these results are classic according according to Otsu (1979).

Gupta *et al.* (2007) presented a local version of the Otsu method which looks at blocks of pixels at several resolution levels when determining the threshold. The goal is to adapt to changing backgrounds and differing font sizes. The smallest block size is adaptively chosen to be twice the dominant line height h (see Appendix for an algorithm to calculate the dominant line height automatically). This fundamental block size $2h \times 2h$ was designed so that it is large enough to contain intact letters (when located in a text region), but small enough to adapt to background changes. Several larger block sizes are also used $4h \times 4h$, $8h \times 8h$, $16h \times 16h$, $32h \times 32h$, $64h \times 64h$, and the entire image. For each block size, the image is considered to be tiled with adjacent non-overlapping blocks. This means that a $2h \times 2h$ block is not centered in its containing $4h \times 4h$ block, but this speeds up the algorithm significantly over using centered multiresolutional blocks.

The binarization consists of the following steps (note that Steps (2) to (5) can be implemented in parallel over the $2h \times 2h$ blocks, and Step (7) is also a parallel operation).

- Step (1) - The image is completely divided up into nonoverlapping adjacent blocks of size $2h \times 2h$.
- Step (2) - For each block, an Otsu threshold is calculated based on the pixels in that block.
- Step (3) - The pixels in each block are binarized.
- Step (4) - If the ratio of binarized white pixels to binarized black pixels is less than two, then Steps $(2) to (4)$ are repeated for the next larger block which contains the given block.
- Step (5) - Each $2h \times 2h$ block is assigned the last Otsu threshold calculated for that block.
- Step (6) - Thresholds t_i for each pixel are formed by bilinear interpolation of the thresholds in Step (5).
- Step (7) - Each pixel x_i of the original image is compared to the corresponding threshold t_i to form the binarized pixel b_i .

Thresholding is a special case of pattern classification in which a one-dimensional feature space is used, the feature being the gray level of the pixel.

The threshold is a "hyper-plane" decision surface (i.e., a point) in this one-dimensional space according to Weszka (1979). If we knew the distribution of gray levels in the given ensemble of images represented here by a mixture of two Gaussian populations with given means and standard deviations then we could determine analytically the threshold that minimizes the classification error.

In the absence of such knowledge, we can approach the problem of threshold selection by performing cluster analysis on the feature space. Suppose that we construct the gray-level histogram of the image (or ensemble); this is a plot showing how often each gray level occurs. A cluster of feature values is then nothing more than a peak on the histogram, corresponding to a densely populated range of gray levels. If we find two peaks on the histogram, it is reasonable to choose a threshold that separates these peaks, at the bottom of the valley between them, since this threshold appears to separate the gray-level population into two distinctive subpopulations.

Several methods have been proposed that produce a transformed gray-level histogram in which the valley is deeper, or is converted into a peak, and is thus easier to detect. Weszka (1979) proposed a standard approach to threshold selection for image segmentation based on locating valleys in the image's gray-level histogram.

Kapur *et al.* (1985) used a method for gray-level picture thresholding using the entropy of the histogram, maximizing the function $\phi(s) = \{H_b + H_w\}$, to obtain the maximum information between the object and background distribution in the picture.

Model-based methods have been proposed by Dawoud & Kamel (2004) and Liu & Srihari (1997).

These methods are efficient for the images in which the gray levels of the text and background pixels are separable. If the histogram of the text overlaps with that of the background, they result in improper binary images, (Valizadeh & Kabir, 2012).

Initially, a category of image was proposed where the image is divided into subimages: $\{Image(1), Image(2) \ldots, Image(M)\}$, with the levels of background noise independent and unknown. The objective of the approach is to find an optimal threshold for each subimage that would eliminate background noise, while preserving as much handwritten stroke data as possible. Suppose that threshold is given by $T(x), x \in \{1, 2, \ldots, M\}$, where M is the total number of subimages. Dawoud & Kamel (2004) showed how to optimize the binarization of a subimage (b) using information from other subimage (a).

The challenge is to show how to optimize the binarization of Image (y) using information from other another subimage, Image (x). Suppose that all subimages are binarized at a CT_i (Candidate to Threshold-i). If the Image (x) is noise free, then its gray-level statistics of the extracted pixels (pixels with gray-level lower than CT_i only) can be used to estimate the parameters of a Gaussian distribution $N_x \sim (\mu_x, \sigma_x)$. An empirical study done by Dawoud & Kamel (2002) showed that the Gaussian distribution is the best among other distributions in representing the handwriting gray-level population. If the binarized Image (y) is noise free also, then we expect its gray-level statistics of its extracted pixels

(μ_y) to be similar to N_x. When fails to eliminate Image (y) noise, these statistics will become different from N_x. Experimentally, we found this expectation to be true, regardless of the noise's pattern or characteristics. So, by testing the following null and alternative hypotheses, we can infer whether or not a eliminated background noise from Image (y):

$$H_0 : \mu_a - \mu_b = 0, \tag{6.0.25}$$

$$H_1 : \mu_a - \mu_b \neq 0. \tag{6.0.26}$$

In hypothesis testing the decision making about the state of nature is based on data. The Table 6.1 represents the true state of nature. Thus, there are two possibilities of error:

Table 6.1: Summary of Statistical Decisions.

Statistical Decision	True state of null hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

1. Type I: rejecting the null hypothesis when the null hypothesis is true.

2. Type II: failing to reject the null hypothesis when the null hypothesis is false.

Where the probabilities of these errors, shown in Figure 6.1, are defined as follows:

- α = P(Type I error) = P(rejecting $H_0|H_0$ is true)

- β = P(Type II error) = P(failing to reject $H_0|H_0$ is false)

- Power = $1 - \beta$ = P(rejecting $H_0|H_0$ is false)

The hypothesis test is conducted by calculating the gray-level that corresponds to 5% of subimage (x) noise-free Gaussian distribution. A Gaussian distribution was used to model the gray-level statistics. To corroborate that this distribution is an appropriate model, the Kolmogorov-Smirnov goodness-of-fit test was used by Dawoud & Kamel (2002).

Then, as shown in Figure 6.2, was calculated the number of pixels above this gray-level was calculated as a percentage of the total number of extracted pixels in the subimage (b). This percentage is interpreted as $\alpha(xy)$, the probability of error associated with accepting the hypothesis that the gray-level statistics of the subimage (b) obey the Gaussian model obtained from the subimage (a), where $a, b \in K$ and $K = \{1, 2, \ldots, M\}$. As long as CT_i eliminates subimage (b) background noise, this error will be small (around 5%), given that subimage (a) is noise free also. The gray-level feature for all subimages is showed in Equation 6.0.27:

$$\begin{bmatrix} - & \alpha(12) & \ldots & \alpha(1M) \\ \alpha(21) & - & \ldots & \alpha(2M) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha(M1) & \alpha(M2) & \ldots & - \end{bmatrix} \tag{6.0.27}$$

The decision to accept or reject H_0 is reached by comparing with a predetermined parameter called the gray-level rejection criterion (GRC). If error exceeds GRC, is rejected; otherwise, it is accepted. When noise interferes in submage (z) at a certain CT_i, then subimage (z) can no longer be used as a valid model to evaluate other subimages. Therefore, we remove it from the list of valid subimages used to calculate at the following iterations.

When the noise interference starts at the same CT_i for two subimages: subimage (a) and subimage (b), there is a chance that neither $\alpha(yx)$ nor $\alpha(xy)$ will increase to reflect

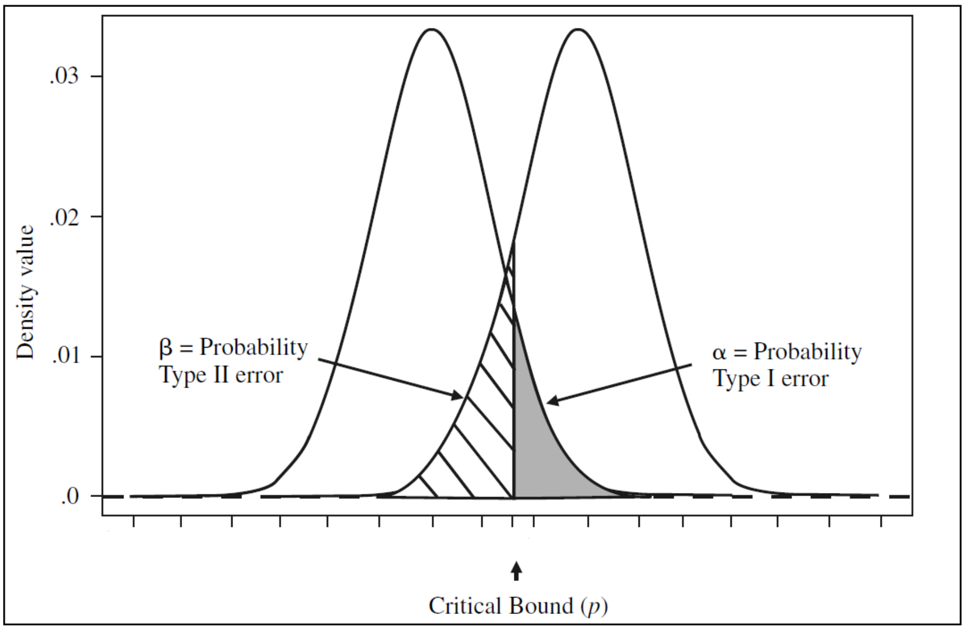


Figure 6.1: Critical bound.

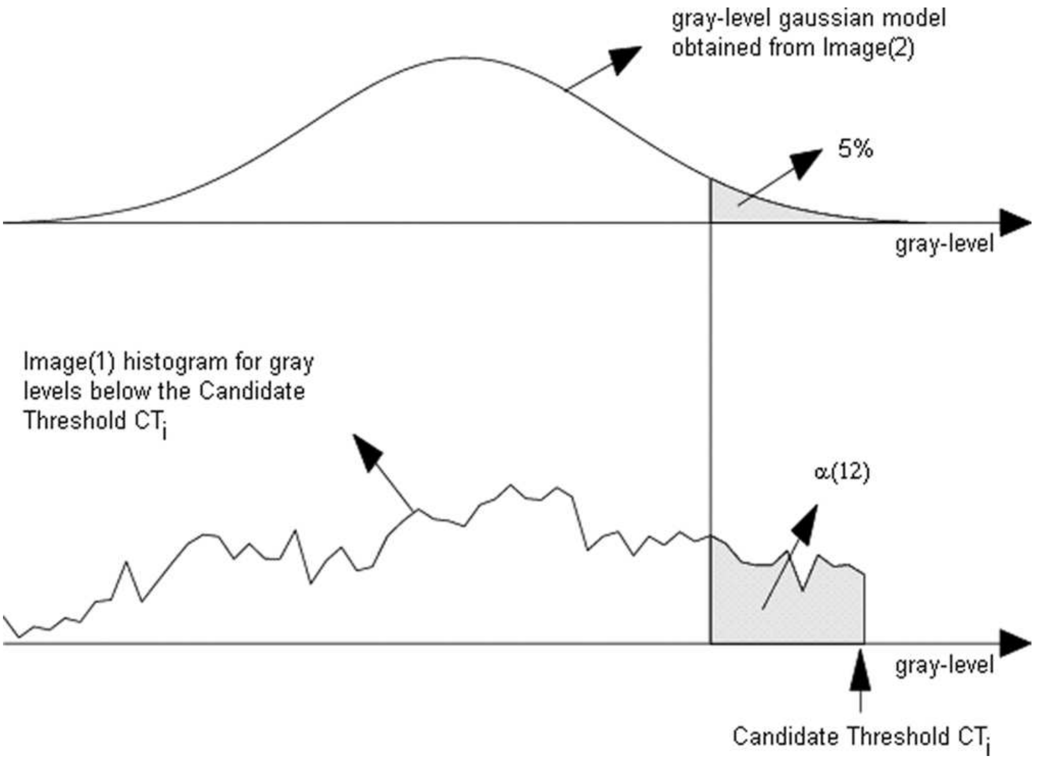

gray-level gaussian model obtained from Image(2)
5%
gray-level
Image(1) histogram for gray levels below the Candidate Threshold CT_i
α(12)
gray-level
Candidate Threshold CT_i

Figure 6.2: Hypothesis test (Dawoud & Kamel, 2004).

this noise interference. In this case, this feature may fail to detect the interference.

The IsoData Method

IsoData is one of the most popular and widely used clustering methods in geoscience applications, but it can run slowly, particularly with large data sets (Memarsadeghi, 2007). The attribute set includes properties as follows:

- Don't need to know the number of clusters.
- Algorithm splits and merges clusters.
- User defines threshold values for parameters.
- Computer runs algorithm through many iterations until threshold is reached.
- Clusters associated with fewer than the user-specified minimum number of pixels are eliminated.
- Isolated pixels are either put back in the pool for reclassification, or ignored as "unclassifiable".

Unsupervised classification techniques use clustering mechanisms to group image pixels into unlabeled classes/clusters. Later on, a human analyst assigns meaningful labels to the clusters and produces a suitably classified image.

The Isodata method works as follows:

1. Cluster centers are randomly placed and pixels are assigned based on the shortest distance to the center method.
2. The standard deviation within each cluster, and the distance between cluster centers is calculated
   - Clusters are split if one or more standard deviation is greater than the user-defined threshold.
   - Clusters are merged if the distance between them is less than the user-defined threshold.
3. A second iteration is performed with the new cluster centers.
4. Further iterations are performed until:
   - the average inter-center distance falls below the user-defined threshold,
   - the average change in the inter-center distance between iterations is less than a threshold, or
   - the maximum number of iterations is reached

Pun Method

Let t be the value of the threshold and define two a posteriori entropies, where H_b and H_w are:

$$H_b = -\sum_{i=0}^{t} p(i)\log_e(p(i)), \tag{6.0.28}$$

$$H_w = -\sum_{i=t+1}^{255} p(i)\log_e(p(i)), \tag{6.0.29}$$

where H_b and H_w can be regarded, respectively, as measure of the a posteriori information associated with the black and white pixels after thresholding.

Knowing the a priori entropy of the gray level histogram, Pun proposed an algorithm to determine the optimal threshold the upper bound of the posteriori entropy:

$$H = H_b + H_w. \tag{6.0.30}$$

Pun (1981)has shown the maximizing H is equivalent to maximizing the evaluation function with respect to t:

$$f(t) = \frac{H_t}{H_T} \frac{\log_e(P(t))}{\log_e[max(p_0, \ldots, p_t)]} + (1 - \frac{H_t}{H_T}) \frac{\log_e[1 - P(t)]}{\log_e[max(p_{t+1}, \ldots, p_{255})]}, \tag{6.0.31}$$

where,

$$H_t = -\sum_{i=0}^{t} p(i) \log_e(p(i)), \tag{6.0.32}$$

$$H_T = -\sum_{i=0}^{255} p(i) \log_e(p(i)), \tag{6.0.33}$$

$$P_t = \sum_{i=0}^{t} p_i. \tag{6.0.34}$$

In his second algorithm, Pun proposed the use of an anisotropic coefficient β in thresholding, where,

$$\beta = \frac{H_t}{H_T} = \frac{\sum_{i=0}^{m} p_i \log_e p_i}{\sum_{i=0}^{255} p_i \log_e p_i}, \tag{6.0.35}$$

and m is the smallest integer such that:

$$\sum_{i=0}^{m} p_i \geq 0.5. \tag{6.0.36}$$

The threshold t^* is chosen such that:

$$\sum_{i=0}^{t^*} p_i = \begin{cases} 1 - \beta & \text{if} \quad \beta \leq 0.5, \\ \beta & \text{if} \quad \beta > 0.5. \end{cases} \tag{6.0.37}$$

Kapur-Sahoo-Wong Filter

Two probability distributions (e.g., object distribution and background distribution) are derived from the original gray level distribution of the image as follows:

$$\frac{p_0}{P_t}, \frac{p_1}{P_t}, \ldots, \frac{p_t}{P_t}, \tag{6.0.38}$$

and

$$\frac{p_{t+1}}{1-P_t}, \frac{p_{t+2}}{1-P_t}, \ldots, \frac{p_{t-1}}{1-P_t}, \tag{6.0.39}$$

where t is the value of the threshold and $P_t = \sum_{i=0}^{t} p_i$. The values of the entropies H_w and H_b are calculated through Equations 6.0.40 and 6.0.41:

$$H_b = -\sum_{0}^{t} \frac{p_i}{P_t} \log_e(\frac{p_i}{P_t}), \tag{6.0.40}$$

$$H_w = -\sum_{t+1}^{t-1} \frac{p_i}{1-P_t} \log_e(\frac{p_i}{1-P_t}), \tag{6.0.41}$$

Then the optimal threshold t^* is defined as the gray level which maximizes $\phi(s) = \{H_b + H_w\}$, that is:

$$t^* = \arg\max\{H_b + H_w\}. \tag{6.0.42}$$

Johannsen-Bille Method

The Johannsen-Bille method choose the threshold t^* from the relation:

$$t^* = \arg\min\{S(t) + S\bar{(}t)\}, \tag{6.0.43}$$

where the algorithm proposed by Johannsen & Bille (1982) aims at minimizing the function $S(t)$ defined as follows.

$$S(t) = \log_e(\sum_{i=0}^{t} p_i) - \frac{1}{(\sum_{i=0}^{t} p_i)}[p_t log_e p_t + (\sum_{i=0}^{t-1} p_i)\log_e(\sum_{i=0}^{t-1} p_i)], \tag{6.0.44}$$

and

$$\bar{S}(t) = \log_e(\sum_{i=t}^{t} p_{L-1}) - \frac{1}{(\sum_{i=t}^{t} p_{L-1})}[p_t log_e p_t + (\sum_{i=t+1}^{L-1} p_i)\log_e(\sum_{i=t+1}^{L-1} p_i)], \tag{6.0.45}$$

Yen-Chang-Chang Method

The criterion is based on the consideration of two factors. The first one is the discrepancy between the thresholded and original images and the second one is the number of bits required to represent the thresholded image. Based on a new maximum correlation criterion for bilevel thresholding, the discrepancy is defined and then a cost function that takes both factors into account is proposed for multilevel thresholding. By minimizing the cost function, the classification number that the gray-levels should be classified and the threshold value can be determined automatically. Computational analysis indicate that the number of required mathematical operations in the implementation of our algorithm is much less that of maximum entropy criterion, according Yen (1995).

A total entropy is defined as:

$$TE(t) = E_b(t) + E_w(t) = -\log\{\sum_{i=0}^{t}[\frac{p_i}{P_t}]^2\} - \log\{\sum_{i=t+1}^{255}[\frac{p_i}{1-P_t}]^2\} \tag{6.0.46}$$

and the threshold is the argument that maximizes that expression.

The maximum entropy criterion is determined the threshold s^* such that

$$TE(s^*) = \max TE_s. \tag{6.0.47}$$

It is well-known that the thresholded image becomes more similar to the original one as the classification number increases. Hence, the discrepancy between the original and thresholded images decreases as the classification number increases. However, the total number of bits required to represent the thresholded image increases as the number of classes increases. Hence, there must exist a compromise between these two factors. Let k denote the classification number and $D(k)$ the discrepancy between the thresholded and original images. The cost function $C(.)$ that takes into account both factors is proposed as

$$C(k) = \rho[D(k)]^{\frac{1}{2}} + [\log_2(k)]^2, \tag{6.0.48}$$

where ρ is a positive weighting constant.

The first term of $C(k)$ measures the cost incurred by the discrepancy between the thresholded and original images, and the second measures the cost resulted from the number of bits used to represent the thresholded image. The automatic thresholding criterion, according to Yen (1995), is then defined to determine the optimal classification number k^* such that

$$C(k^*) = \min C(k). \tag{6.0.49}$$

Otsu Threshold Method

The mean and variance of the object and background in relation to the threshold t are defined as follows.

$$\mu_b(t) = \sum_{i=0}^{t} ip_i, \tag{6.0.50}$$

$$\mu_w(t) = \sum_{i=t+1}^{255} ip_i. \tag{6.0.51}$$

The class variances are given by

$$\sigma_b^2(t) = \sum_{i=0}^{t}(i - m_b)^2 p_i, \tag{6.0.52}$$

$$\sigma_w^2 = \sum_{i=t+1}^{255} (i - m_t)^2 p_i. \tag{6.0.53}$$

The "optimal" value for this limit is the argument that maximizes the following expression

$$\eta(t) = \frac{P_t(1-P_t)[m_b(t)-m_w(t)]^2}{P_t\sigma_b^2(t)+(1-P_t)\sigma_w^2(t)}. \tag{6.0.54}$$

Mello-Lins Algorithm

The algorithm by Mello & Lins (2002) and Mello & Lins (2000) looks for the most frequent gray level of the image and takes it like initial threshold to evaluate the values H_b, H_w and H by equations 6.0.28, 6.0.29 and 6.0.30. Based on the value of H, three classes of documents were identified, according to Mello (2006). It was defined two multiplicative factors and the entropy H determines the value of weights m_b and m_w, as follows:

- If $H \leq 0.25$, (documents with few parts of text or vary faded ink), then $m_w = 2$ and $m_b = 3$.
- If $0.25 < H < 0.30$, (the most common cases), then $m_w = 1$ and $m_b = 2.6$.
- If $H \geq 0.30$, (documents with many blacks areas), then $m_w = 1$ and $m_b = 1$,

then, the threshold is directly calculated by:

$$t = (m_b H_b + m_w H_w). \tag{6.0.55}$$

Roe-Mello Algorithm

The proposed method makes use of local image equalization based on color constancy, and an extension to the standard difference of Gaussians edge detection operator, XDoG. The binarization is achieved after three main steps: the first step removes undesirable degradation artifacts using a local image equalization and Otsu binarization algorithm. The second step uses global image equalization and XDoG edge detection operator to binarize the text. The final step combines the two previous steps, performing a cleanup to remove remaining degradations artifacts and fix possible missing text or area, to produce the final result.

The main approach is to minimize the document degradation using an equalization schema based on one used to compensate differences in illumination. The equalization is applied twice: the first time locally using a small neighborhood around each pixel and the second time globally over the entire image. After the first equalization, the resulting image is binarized using the Otsu (1979) algorithm and an extension of the Difference of Gaussians, called XDoG (Winnemöller, 2011) is applied to the result of the second equalization. In the next step, the two binarized results are merged together into a new image and cleaned. Finally, in the last step, the cleaned image is analyzed and gaps left by equalization are fixed. The steps, and its sub-steps, are summarized in the following list and further detailed:

1. First binarization: (A) Local image equalization and (B) Binarization using the Otsu algorithm;

2. Second binarization: (C) Global image equalization and (D) Edges detection using XDoG;

3. Cleanup and restoration: (E) Combine the results from steps (B) and (D); (F) Remove noise from image generated in step (E); and Restoration, filling gaps.

Silva-Lins-Rocha Approach

Thus, the entropy of the a priori source is given by

$$H = -\sum_{i=0}^{255} p_i \log_2 p_i, \tag{6.0.56}$$

where p_i is provided by Equation 6.0.5. As the resulting image is binarized, the distribution of its histogram may be seen as a distribution of a binary source (a posteriori source). The entropy of the a posteriori source is given by:

$$H'(t) = h(P_t), \tag{6.0.57}$$

where $h(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ is the binary entropy function, according to Abramson (1963), and P_t is provided by Equation 6.0.6. Next, one makes an extension of a binary source to represent without losses all the 256 symbols of the a priori source. This new binary source is called the a priori binary source. The value of the entropy of this new source is given by

$$H_{\textbf{(a priori binary source)}} = \frac{H}{\log_2(256)} = \frac{H}{8}. \tag{6.0.58}$$

One then looks for a value of t such that the entropy of the a posteriori source is as close as possible to the value of the entropy of the a priori binary source, i.e., one looks for the following equality:

$$H'(t) = H_{\textbf{(a priori binary source)}}. \tag{6.0.59}$$

This argument maps the distribution of the a posteriori source onto the distribution of the a priori binary source. Applying Equations 6.0.57 and 6.0.58 to 6.0.59 it follows that

$$h(P_t) = \frac{H}{8}. \tag{6.0.60}$$

The behavior of the binary entropy function, Figure 6.3, must be taken into account as follows. It must be taken into account that the target images are of documents, with a much higher frequency of background (paper) pixels than object (print or written) ones. Thus, it is reasonable to work with the argument of P_t within the interval $[0.0, 0.5]$. In this interval, the entropy function is injective, thus there is only one value of P_t that satisfies the equation, unless p_i is zero. In such a case it would not matter if the calculated limit were i or $i-1$. The target of the proposed algorithm is to filter out the back-to-front interference in binarization. Due to its features, the interference raises the value of the a priori source's entropy. A loss factor $\gamma H_{\textbf{(a priori binary source)}}$, according to Abramson (1963), experimentally determined, is introduced to reduce the presence of the interference. Thus, the following relation holds:

$$H'(t) = \gamma[H_{\textbf{(a priori binary source)}}][H_{\textbf{(a priori binary source)}}] \tag{6.0.61}$$

Once the bases of the algorithm are presented, its steps are now detailed.

1. Calculate H, the entropy of the image histogram.

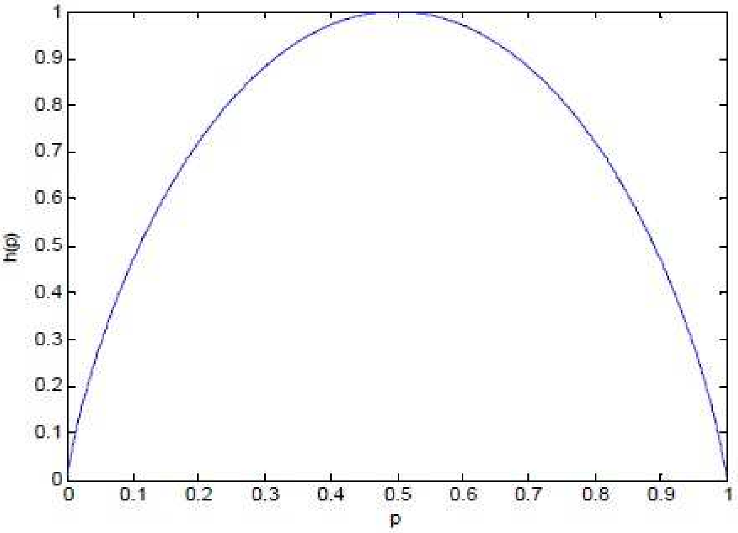


Figure 6.3: Behavior of the binary entropy function h(p).

2. Scan the t levels, calculating of each of them the distributions $(P_t, 1-P_t)$, while $P_t \leq 0.5$, and the entropy associated with that distribution $H'(t) = h(P_t)$;

3. Determine the "optimal" limit that minimizes $|e(t)|$ given as:

$$e(t) = |\frac{H'(t)}{\frac{H}{8}} - \gamma(H/8)|. \tag{6.0.62}$$

Wu-Lu Algorithm

According to Kapur-Sahoo-Wong method, the optimal threshold t^*, as defined in Equation 6.0.42 is defined as the gray level which maximizes $\phi(s) = \{H_b + H_w\}$. As a result, a gray level minimizing the difference between the entropy of the object and that of the background $\phi^{'}(s) = \{H_b - H_w\}$ will be the desired threshold, according Equation 6.0.63 (Wu, 1998).

$$t^* = \arg\min\{H_b - H_w\}. \tag{6.0.63}$$