Assessing Binarization Techniques for Document Images

Rafael Dueire Lins
UFPE/UFRPE, Recife, PE
Brazil
rdl.ufpe@gmail.com

Marcos Martins de Almeida
UFPE, Recife, PE
Brazil
mm.ufpe@gmail.com

Rodrigo Barros Bernardino
UFPE, Recife, PE
Brazil
rbbernardino@gmail.com

Darlisson Jesus
UFPE, Recife, PE
Brazil
dmj.ufpe@gmail.com

José Mário Oliveira
IFPE/UFPE, Recife, PE
Brazil
josealexandre@recife.ifpe.edu.br

ABSTRACT

Image binarization is a technique widely used for documents as monochromatic documents claim for far less space for storage and computer bandwidth for network transmission than their color or even grayscale equivalent. Paper color, texture, aging, translucidity, kind and color of ink used in handwritting, printing process, digitalization process, etc., are some of the factors that affect binarization. No algorithm is good enough to be a winner in the binarization of all kinds of documents. This paper presents a methodology to assess the performance of binarization algorithms for a wide variety of text documents, allowing a judicious quantitative choice of the best algorithms and their parameters.

CCS CONCEPTS

• **Applied computing → Computers in other domains → Publishing**

KEYWORDS

Documents; binarization; back-to-front interference; bleeding; show through; image filtering; big-data.

ACM Reference format:

R.D.Lins, M.M. de Almeida, R.B. Bernardino, D. Jesus, J.M. Oliveira. 2017. Assessing Binarization Techniques for Document Images. In *Proceedings of ACM Symposium on Document Engineering, Valetta, Malta, September 2017, (DocEng' 17)*, 10 pages.
DOI: 10.1145/3103010.3103021

1 INTRODUCTION

Document image binarization is an important step in the document image analysis and recognition pipeline. Monochromatic documents claim for far less storage space and computer bandwidth for network transmission than color or grayscale documents. It is imperative to have a benchmarking dataset along with an objective evaluation methodology to capture the efficiency of current document image binarization algorithms.

The international competitions on binarization algorithms are an evidence of the relevance of this area. The most traditional of such competitions is possibly DIBCO - Document Image Binarization Competition, which was first organized at the ICDAR-International Conference on Document Analysis and Recognition in 2009 and has been repeated yearly ever since. The methodology used by DIBCO is to offer a small set of "real-world" images and their "ground-truth" binary equivalent that were "hand-generated" or "hand-retouched". Figure 1 presents the complete test set of the ten images used at DIBCO 2016, which may be obtained at http://vc.ee.duth.gr/h-dibco2016/benchmark/. As one may observe in Figure 1, the DIBCO test set is formed only by handwritten documents both in grayscale and color. Some documents present stains (1, 3, 4, 10) and aging marks (4, 9, 10). DIBCO provides an evaluation tool that yields as output the F-Measure, pseudo F-Measure, PSNR, DRD, Recall, Precision, pseudo-Recall and pseudo-Precision. Some of those measures are not usual and are explained in reference [1]. DIBCO 2017 intends to include images of typed or printed documents in its dataset, which has not been released so far (https://vc.ee.duth.gr/dibco2017/ last visited on 04th July, 2017).

As one may observe, all document images in DIBCO 2016 test set, but the first one, have the back-to-front interference, that is, 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 image overlap phenomenon was first addressed in the literature by Lins in 1994 [2], who called it back-to-front interference. Much later, other researchers called it bleeding or

DocEng '17, September 04-07, 2017, Valletta, Malta
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4689-4/17/09...$15.00
http://dx.doi.org/10.1145/3103010.3103021

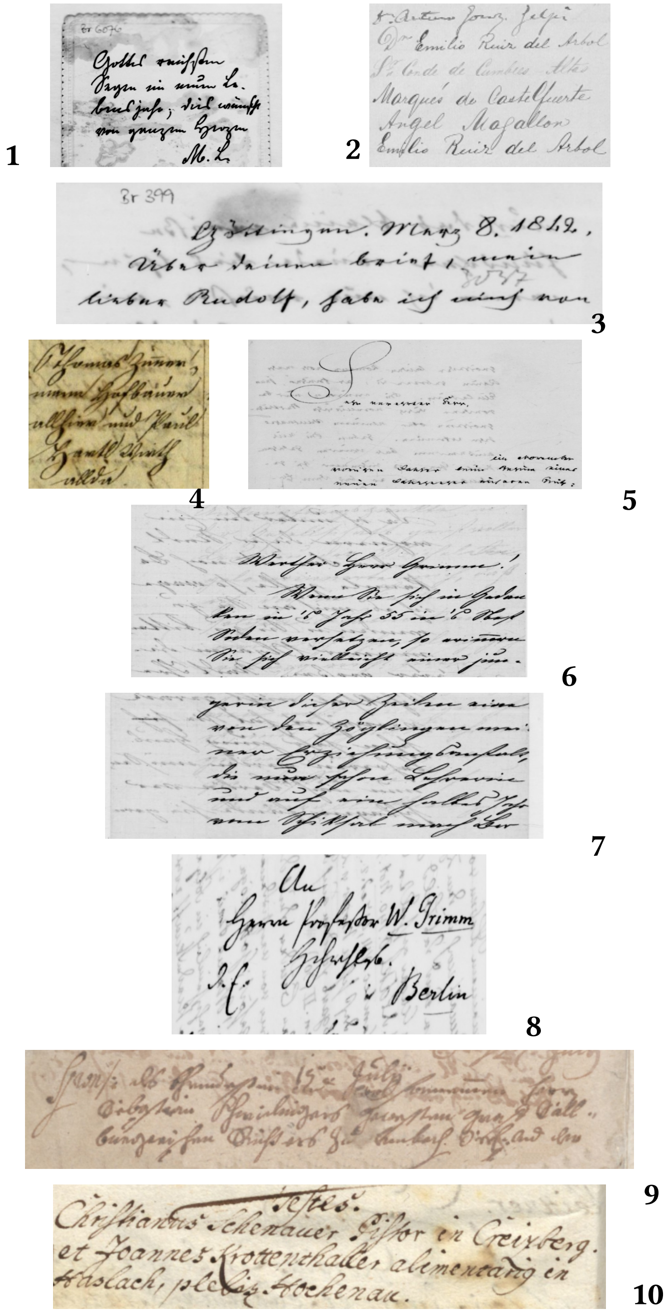


Figure 1: DIBCO 2016 Test images

show-through [3]. The human brain is able to filter out that sort of noise keeping document readability. This is not the case with automatic tools such as OCRs. The direct application of some binarization algorithms such as the one in Jasc Paint Shop Pro TM version 8 (Palette component: Gray values, Reduction component: nearest color, Palette weight: non-weighted), as many other commercial tools, yield a completely unreadable document, as the interfering ink of the backside of the paper overlaps with the binary one in the foreground. Several algorithms were developed specifically to binarize documents with back-to-front interference [7][10][11][13], but depending on the strength of the interference present, which accounts on the opacity of the paper, its permeability, the kind and degree of fluidity of the ink used, the degree of difficulty for obtaining a good segmentation capable of filtering-out such a noise increases enormously, as new set of hues of paper and printing colors appear.

Document image binarization is extremely challenging and there is no chance of a specific algorithm to be an all case winner as many parameters may interfere in the quality of the resulting image. Besides that, a small set of test images will never be able to provide a real quality assessment of binarization algorithms. It is important to be able to have a very large test set of synthetic images representative of the universe of text documents and to know for each of them which algorithms and with which parameters, minimum space and processing time one is able to get the best binarization result. Artificial intelligence and big-data strategies now provide the resources to given a "real-world" document image to be able to decide which kind of document it better matches in such a large database. Known the best-match between the "real-world" document and the synthetic one, the set of suitable binarization algorithms and their parameters becomes known.

This paper explains the methodology used in the generation of such a large controlled database for synthetic images. A quantitative measure of quality is introduced. Some evidence of the effectiveness of the method proposed is also provided.

2 GENERATING SYNTHETIC IMAGES

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. Figure 2 presents the block diagram for the generation of synthetic images.

Two images of documents of different nature (typed, handwritten with different pens, printed, etc.) are taken: F – front and V – verso (back). The verso image is offset by 10, 20 and 30 pixels to make the back image not to coincide with the front one. Then, the offset verso image 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 [20] 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.

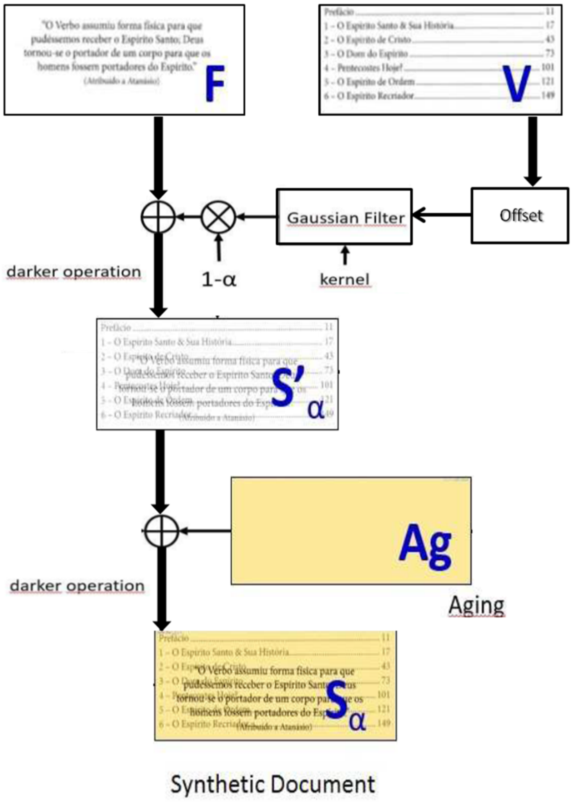
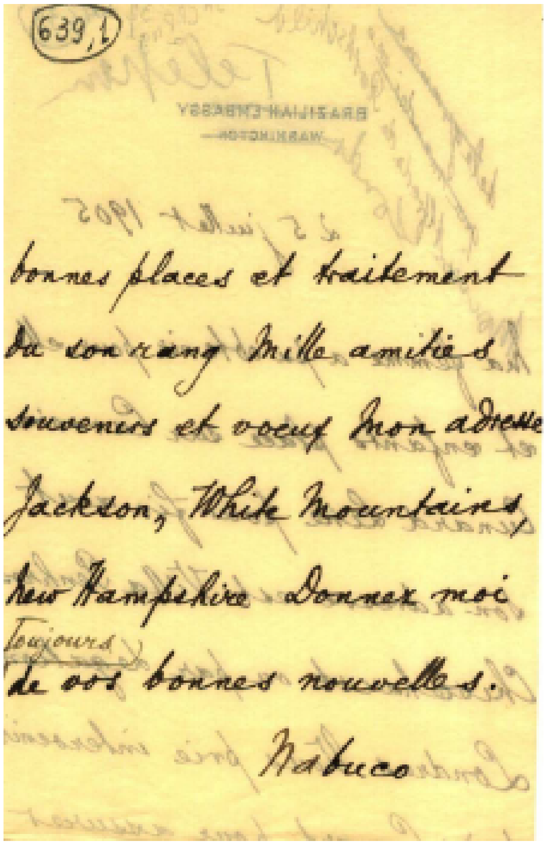

"O Verbo assumiu forma física para que pudéssemos receber o Espírito Santo; Deus tornou-se o portador de um corpo para que os homens fossem portadores do Espírito."
(Atribuído a Atanásio)
F
Prefácio 11
1 – O Espírito Santo & Sua História 17
2 – O Espírito de Cristo 43
3 – O Dom do Espírito 73
4 – Pentecostes Hoje! 101
5 – O Espírito de Ordem 121
6 – O Espírito Recriador 149
V
Offset
Gaussian Filter
kernel
1-α
darker operation
S'α
Ag
Aging
darker operation
Sα
Synthetic Document

Figure 2: Block diagram of the scheme for the generation of synthetic images



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

Figure 3 – Letter from Joaquim Nabuco

2.1 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 2,480 × 3,508 pixels. 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. Salt and pepper noise is removed. Such images correspond to 43 handwritten and 88 printed documents.

The set of ground truth documents of the whole DIBCO series, 61 handwritten and 25 typewritten documents, were also used here. Besides those, 14 documents electronically generated pdf documents are also used as ground-truth. Thus, currently, 231 document images compose the set of ground-truth images in total.

2.2 The Back-to-front blur

As already mentioned, documents with the back-to-front interference are much harder to binarize. Depending on the thickness of the paper, its texture, permeability, age, storage conditions (temperature, humidity, direct exposure to sun light, etc.), kind of ink, printing process or pen in case of handwritten documents, etc., the back ink is seen more or less blurred in the front side of the paper. Such effect has been modeled now as being performed by a Gaussian filter.

Two "light" Gaussian filters 3x3 and 5x5 pixel-kernels were used at the current stage of the generation of the database of synthetic images, presenting "similar" effect as the one in real documents under visual inspection. Current work is being developed to better model this effect in the different kinds of documents. For that, several samples of small windows are being used collecting parts from the foreground and back-to-front interference. The foreground window will be blurred using Gaussian filters having their parameters modified to match the one of the interference. Performing such approximation in several different kinds of documents one will be able to obtain the parameters of the different low-pass filters that better model the bleeding effect, or the back-to-front blur.

2.3 Image Fading

The origin of this project dates back to the early 1990's when the first author of this paper [4] undertook the mission of digitalizing the bequest of historic documents of Joaquim Nabuco, a Brazilian statesman, writer and diplomat, leader in the freedom of black slaves in Brazil. His active correspondence is of paramount importance for understanding the history of the Americas in the late 19th century. That bequest of about 6,500 documents encompassed over 18,000 pages. Those documents were risking of degradation due to problems in the extreme acidity of the paper. A careful analysis of the preservation staff of the Joaquim Nabuco Foundation, the social science research institute in Recife, Brazil, that keeps most of Nabuco's documents, selected about 300 documents as representative of the universe of documents. At that time, for storage restrictions and transmission of documents via FAX-simile devices, binarization was mandatory. That was exactly the first time that the back-to-front interference was reported in the technical literature [2], because about 200 of those documents were written on both sides of translucent paper, with a great variability of strength. Figure 3 presents an example of one of those letters from Nabuco bequest.

The "strength" of the back-to-front interference is modeled by the fading coefficient α. One hundred different levels of fading coefficients were chosen, thus 0<α<1 in steps of 0.01.

2.4 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 3,351 document images were used, of which 1,048 were from Nabuco bequest and the other 2,303 were obtained from five years of the LiveMemory Project, which generated a digital library of all the proceedings the SBrT - Brazilian Telecommunications Symposium. The images were automatically scanned looking for a window of 20x50 pixels such as the purple one shown in Figure 4.

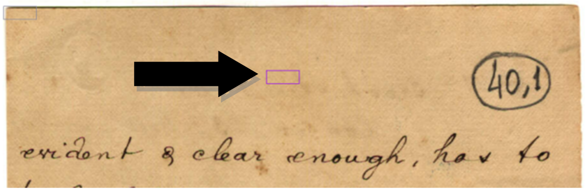


Figure 4 – Sample of the texture of paper background.

The automatic window selection was human checked to guarantee that the area has no ink or other sort of noises. 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 5 images the following 7 statistic measures were taken and placed in a vector: mean, standard deviation, mode, minimum value, maximum value, median, and kurtosis.

The 3,351 vectors were statistically analyzed using the hierarchical clustering method implemented in the scikit-learn library [22]. It uses a bottom up approach, where each observation starts in its own cluster, and clusters are successively merged together, providing 84 cluster distributions of paper texture as shown in Figure 5. The texture in the centroid of each of such clusters was taken as being representative of the whole cluster. The visual inspection made in the 84 clusters showed acceptable texture variation within each cluster.

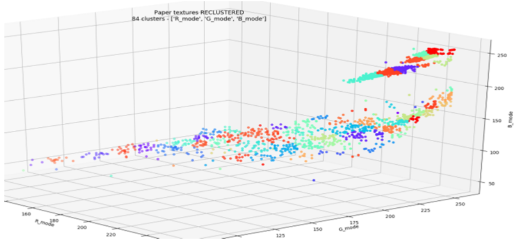
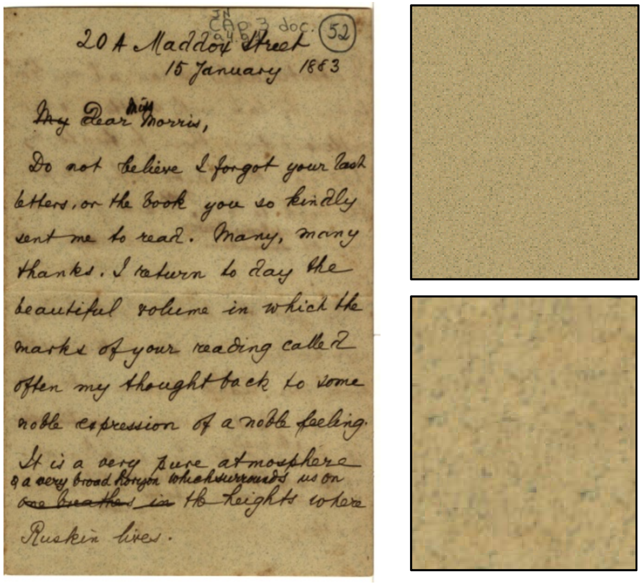


Figure 5 – Distribution of 3,351 paper textures in 84 representative clusters.

Besides those 84 centroid cluster-representative textures, 16 "isolated" textures that were left-out of the clusters were added to the texture set, totaling 100 different textures. Each of those textures is used for generating a "blank" sheet of paper to be used to colorize the synthetic image providing the "aging" effect in the scheme presented in Figure 2. For that, a RGB-image with 2,480 × 3,508 pixels (equivalent to an A4 blank sheet of paper with 300 dpi resolution) is generated. A similar technique is used to generate a 300 dpi texture for the smaller DIBCO ground-truth images. Two different texture generation strategies were adopted. In the first one, the color of each pixel is randomly chosen from the 10,000 pixels in another 100x100 pixels sample of the texture at the center of the texture cluster, providing a 300 dpi image with the same distribution as the original sample. The second technique employs image quilting [17]. Figure 6 presents an example of a texture generated using both techniques, in which the latter more closely resemble the texture of the sample document.

Each image is than added with a "darker" operation [20] generating the set of Sα synthetic images, which will be used to assess the binarization algorithms. Reference [5] proposes a parametric scheme for image compression and generation in which the paper texture is generated through a Gaussian distribution centered on the mean value of the color of the pixels. Both schemes presented here allows more "natural looking" textures that can be efficiently indexed.

The current version of the test set of synthetic images encompasses a total 2,777,000 color images (231 groud-truth x 2 blur x 100 α-fading-coefficients x 3 offsets x 100 textures-patterns x 2 texture generation schemes) and the same number of grayscale equivalent. It is probable that the analysis of the binarization of this set of 5,554,000 images will provide a better assessment of the binarization capability of algorithms than the set of only 10 images in DIBCO 2016.



CAP 3 doc. 52

20 A Maddox Street
15 January 1863

My Dear Miss Morris,

Do not believe I forgot your last letters, or the book you so kindly sent me to read. Many, many thanks. I return to day the beautiful volume in which the marks of your reading called often my thought back to some noble expression of a noble feeling. It is a very pure atmosphere a a very broad horizon which surrounds us on one breathes in the heights where Ruskin lives.

Figure 6 – (Left) Historic document. (Top-right) Texture: random distribution. (Bottom-right) Texture: image quilting.

3 ASSESSING ALGORITHMS

The enormous variety of kinds of text documents makes extremely improbable that one single algorithm is able to satisfactorily binarize all kinds of documents. Most probably, depending on the nature (or degree of complexity) of the image several or no algorithm will be able to provide good results. If binarization is part of an OCR transcription platform, the higher the correct transcription rate the better the algorithm is. It is important to remark that, according to the experiments made by the authors of this paper, OCR transcription and “visual inspection” assessment methods do not provide similar results, even in printed or typed documents. The assessment method proposed here is to provide accurate information about the binarized documents generated by the different algorithms, and the user will choose the most suitable one depending on the target application.

The assessment methodology proposed here is “image centered” instead of the traditional “algorithm centered” approach. This means that the question to be answered here is “Which are the best algorithms and their parameters to binarize image X?” instead of the traditional one “Which is the best algorithm?”. Such a new approach does not provide an answer, but a set of answers. Obviously, humans are not able to handle and analyze such a large set of data, which has to be made “user-friendly” in an automated platform, currently under development by the authors.

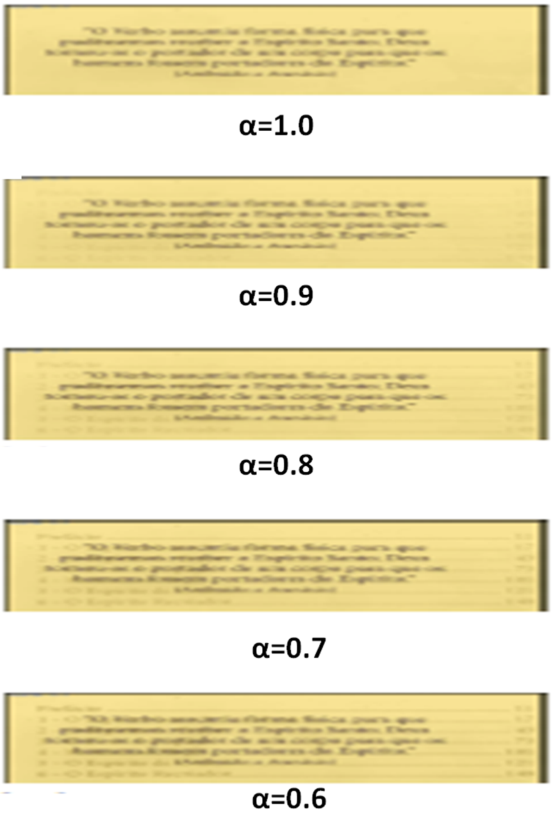
Binarization algorithms, in general, make use of different criteria to find a threshold that splits the mapping of pixels onto white or black. Thresholding algorithms can be classified into global or local algorithms. Global algorithms define a unique threshold value for the complete image. Local algorithms 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 local algorithms potentially provide better results as their parameters are better tuned for each small window, the kind of “tiling” effect of the small blocks tend not to yield acceptable quality results. The assessment methodology presented here works equally well with global and local binarization algorithms.

Sezgin and Sankur [6] presented a comprehensive overview and comparison of the “classical” binarization algorithms, clustering them according to their nature. From the almost forty algorithms presented there, six schemes were chosen to illustrate: Kapur-Sahoo-Wang [7], Otsu [8], Johannsen-Bille [9], Yen-Chang-Chang [10], Wu-Lu [11], and Pun [19] algorithm.

The binarization using the IsoData - Iterative Self Organizing Data Analysis Technique [18] was also tested. It is a method of unsupervised classification, and the computer runs the algorithm through several iterations until the threshold is reached.

Four algorithms specifically developed to filter-out the back-to-front interference were also assessed: Mello-Lins [13], Silva-Lins-Rocha [11], Roe-Mello [7], and Almeida-Lins-Lima [15].

The basic criterion for the choice of the algorithms assessed here was code availability. To illustrate the assessment methodology proposed here, one synthetic document was chosen with $0.1 \leq \alpha \leq$ in steps of 0.1. Samples of some of those documents are presented in Figure 7.



α=1.0

α=0.9

α=0.8

α=0.7

α=0.6

Figure 7 – Synthetic images with 0.6<α<1.

The tables below present: P(b|b) - the percentage of background pixels correctly mapped onto white pixels of the ground-truth image, P(f|f) – the percentage of foreground pixels correctly mapped onto black pixels of the ground-truth image, P(f|b) and P(b|f) are the percentage of mismatches. The column "Threshold" presents the value of the threshold automatically chosen by the algorithm.

The tables below present: P(b|b) - the percentage of background pixels correctly mapped onto white pixels of the ground-truth image, P(f|f) – the percentage of foreground pixels correctly mapped onto black pixels of the ground-truth image, P(f|b) and P(b|f) are the percentage of mismatches. The column "Threshold" presents the value of the threshold automatically chosen by the algorithm.

3.1 The Kapur-Sahoo-Wong Filter

The algorithm by Kapur et al. [7] 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.

Table 1: Kapur-Sahoo-Wong

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	176	90.88	9.12	100.00	0.00
0.2	174	91.50	8.50	100.00	0.00
0.3	174	91.86	8.15	100.00	0.00
0.4	174	92.29	7.71	100.00	0.00
0.5	173	92.98	7.02	100.00	0.00
0.6	174	93.49	6.51	100.00	0.00
0.7	147	99.25	0.75	100.00	0.00
0.8	162	98.87	1.13	100.00	0.00
0.9	175	98.59	1.41	100.00	0.00
1.0	182	98.36	1.64	100.00	0.00

The analysis of the data in Table 1 reveals that there was the partial elimination of the back-to-front interference, for 0.7≤α≤1.0 as the value of background-background 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 1 clearly shows that this algorithm reaches the best performance for the image with α=0.7, with a P(b|f) of 0.75%.

3.2 Otsu threshold method

Otsu [8] 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. The result of this algorithm applied to the synthetic images with different alphas is shown in Table 2.

Although Otsu algorithm was originally developed for ultrasound images, the results above show that it performs well with document images. Table 2 shows that for 0.7≤α≤1.0, the value of background-background correct mapping percentage was 99.87%≤P(b|b) ≤99.95% yielding error less than 0.13%, while the foreground-foreground percentage 99.54%≤P(f|f)≤99.56%, an error less than 0.47%. Comparing the data presenting in Table 1 and 2 one may conclude that Otsu presented better results than Kapur-Sahoo-Wong filter for that specific set of images.

Table 2: Otsu Filter

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	145	94.19	5.81	100.00	0.00
0.2	145	94.57	5.43	100.00	0.00
0.3	145	95.05	4.95	100.00	0.00
0.4	149	95.24	4.76	100.00	0.00
0.5	149	96.00	4.00	100.00	0.00
0.6	146	97.51	2.49	100.00	0.00
0.7	138	99.87	0.13	99.54	0.46
0.8	138	99.94	0.06	99.56	0.44
0.9	138	99.97	0.03	99.53	0.47
1.0	140	99.95	0.05	99.55	0.45

3.3 Johannsen-Bille

This method [9] 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 presents the performance obtained by this filter for the test set. 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 was completely black images. The Johanssen-Bille algorithm presented in some of the cases (α=0.5, 0.9, 1.0) an information loss, as over 10% of the foreground pixels were mapped onto background ones.

Table 3: Johanssen-Bille

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	142	94.49	5.51	99.52	0.48
0.2	149	94.23	5.77	100.00	0.00
0.3	210	0.00	100.00	100.00	0.00
0.4	150	95.15	4.85	100.00	0.00
0.5	100	99.97	0.03	84.63	15.37
0.6	211	0.00	100.00	100.00	0.00
0.7	211	0.00	100.00	100.00	0.00
0.8	211	0.00	100.00	100.00	0.00
0.9	112	100.00	0.00	88.39	11.61
1.0	112	100.00	0.00	88.11	11.89

3.4 Yen-Chang-Chang

The binarization algorithm by Yen-Chang-Chang [10] follows the same ideas as the one by Kapur et al. [7] 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 4.

Table 4: Yen-Chang-Chang

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	210	0.00	100.00	100.00	0.00
0.2	210	0.00	100.00	100.00	0.00
0.3	210	0.00	100.00	100.00	0.00
0.4	210	0.00	100.00	100.00	0.00
0.5	178	92.14	7.86	100.00	0.00
0.6	211	0.00	100.00	100.00	0.00
0.7	211	0.00	100.00	100.00	0.00
0.8	211	0.00	100.00	100.00	0.00
0.9	176	98.47	1.53	100.00	0.00
1.0	183	98.23	1.77	100.00	0.00

The results presented in Table 4 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 black images.

3.5 The Wu-Lu algorithm

The Wu-Lu binarization algorithm [11] 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 5 presents the results obtained in using Wu-Lu algorithm in the binarization of the test set images.

Analyzing the results presented in Table 5, one may see that, although the value of the percentage of background-background mapping P(b|b) did not vary much and is either 100.00% or very close to that value for all the α´s, 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.

Table 5: Wu-Lu

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	75	99.13	0.87	62.81	37.19
0.2	75	99.00	1.00	62.45	37.55
0.3	74	99.96	0.04	61.00	39.00
0.4	73	100.00	0.00	59.72	40.28
0.5	72	100.00	0.00	57.70	42.30
0.6	71	100.00	0.00	55.86	44.14
0.7	70	100.00	0.00	54.23	45.77
0.8	68	100.00	0.00	50.21	49.79
0.9	66	100.00	0.00	45.99	54.01
1.0	62	100.00	0.00	36.61	63.39

3.6 Pun Algorithm

The algorithm proposed by Pun [19] takes as input a gray level 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.

Table 6: Pun

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	195	61.99	38.01	100.00	0.00
0.2	196	57.97	42.03	100.00	0.00
0.3	196	59.15	40.85	100.00	0.00
0.4	196	61.64	38.36	100.00	0.00
0.5	196	65.20	34.80	100.00	0.00
0.6	196	67.16	32.84	100.00	0.00
0.7	198	55.51	44.49	100.00	0.00
0.8	198	58.39	41.61	100.00	0.00
0.9	198	60.52	39.48	100.00	0.00
1.0	199	60.52	39.48	100.00	0.00

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 α's, the P(b|b) was around 60%, reaching 55.51 % for α = 0.7, meaning that are large number of background pixels were mapped onto black pixels of the monochromatic image.

3.7 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) [18] allows the number of clusters to be adjusted automatically during the iteration by merging similar clusters and splitting clusters with large standard deviations. 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. Table 7 presents the result of the binarization of the test set images using the IsoData algorithm.

Table 7: IsoData Clustering

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	142	94.49	5.51	99.52	0.48
0.2	142	94.84	5.16	99.53	0.47
0.3	144	95.14	4.86	100.00	0.00
0.4	146	95.54	4.46	100.00	0.00
0.5	147	96.22	3.78	100.00	0.00
0.6	144	97.85	2.15	100.00	0.00
0.7	136	99.89	0.11	98.87	1.13
0.8	137	99.94	0.06	99.23	0.77
0.9	137	99.98	0.02	99.20	0.80
1.0	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≤α≤1.0, the value of the background-background mapping yielded an error of less than 0.11% as 99.89%<P(b|b)<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 (α=0.1, α=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.

3.8 Mello-Lins Algorithm

The algorithm by Mello and Lins [12] is based on Shannon entropy to calculate a global threshold. 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 8.

Table 8: Mello-Lins

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	174	91.19	8.81	100.00	0.00
0.2	183	89.76	10.24	100.00	0.00
0.3	181	90.58	9.42	100.00	0.00
0.4	180	91.21	8.78	100.00	0.00
0.5	178	92.14	7.86	100.00	0.00
0.6	176	93.14	6.86	100.00	0.00
0.7	174	94.47	5.53	100.00	0.00
0.8	170	97.30	2.70	100.00	0.00
0.9	165	99.19	0.81	100.00	0.00
1.0	181	98.45	1.55	100.00	0.00

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 α=0.9.

3.9 Silva-Lins-Rocha algorithm

The algorithm developed by Silva-Lins-Rocha [13] 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 [13], 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 9.

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 α=0.9, for some reason.

Table 9: Silva-Lins-Rocha

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	89	97.60	2.40	78.73	21.27
0.2	95	97.77	2.23	82.80	17.20
0.3	105	97.94	2.06	86.73	13.27
0.4	115	98.17	1.83	90.60	9.40
0.5	126	98.44	1.56	94.96	5.04
0.6	137	98.80	1.20	99.22	0.74
0.7	150	98.80	1.20	100.00	0.00
0.8	161	98.98	1.02	100.00	0.00
0.9	167	99.07	0.93	100.00	0.00
1.0	165	99.26	0.74	100.00	0.00

3.10 Roe-Mello

The Roe-Mello [14] 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 10.

Table 10: Roe-Mello

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	181	88.16	11.84	39.39	60.61
0.2	181	88.41	11.59	39.10	60.90
0.3	181	88.73	11.27	39.11	61.89
0.4	180	89.23	10.76	36.45	63.55
0.5	181	94.84	5.16	23.70	76.30
0.6	181	95.41	4.59	22.46	77.54
0.7	181	95.55	4.45	22.10	77.90
0.8	181	95.63	4.37	22.04	77.96
0.9	181	95.63	4.37	22.03	77.97
1.0	181	98.58	4.42	22.13	77.87

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

3.11 The Almeida-Lins-Lima algorithm

The algorithm recently proposed by Almeida, Lins and Lima [15] is performed in four steps: filtering the image using a bilateral filter [16], splitting image into the RGB components, decision-making for each RGB channel based on an adaptive binarization method inspired by Otsu's method with a choice of the threshold level, and classification of the binarized images to decide which of the RGB components best preserved the document information in the foreground. It is far more computation intensive than its predecessors and involves training for the Decision-making block. Testing this algorithm with the same set of test images the automatically chosen threshold is equal to 126 and the channel that is chosen for providing the best results in binarizing the images is the Red channel. The results obtained are summarized in Table 11.

Table 11: Almeida-Lins-Lima

α	Threshold	P(b|b)%	P(b|f)%	P(f|f)%	P(f|b)%
0.1	126	96.49	3.51	100.00	0.00
0.2	126	96.93	3.07	100.00	0.00
0.3	126	97.66	2.34	100.00	0.00
0.4	126	99.60	0.40	100.00	0.00
0.5	126	99.87	0.13	100.00	0.00
0.6	126	99.91	0.09	100.00	0.00
0.7	126	99.94	0.06	100.00	0.00
0.8	126	99.97	0.03	100.00	0.00
0.9	126	99.99	0.01	100.00	0.00
1.0	126	100.00	0.00	100.00	0.00

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.

4 GLOBAL RESULTS

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. Further testing was made with a larger set of 1,600 synthetic images with the coefficient α varying between 0 and 1 in steps of 0.01. 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 12. The data presented in Table 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.

Table 12: Overall algorithm classification for 1,600 synthetic images with 0<α<1 in steps of 0.1.

α	*P(b|b)%*	*P(f|f)%*	*Filter*	*Threshold*
1.0	100.00	100.00	Almeida-Lins-Lima	126
1.0	100.00	99.56	IsoData	138
1.0	99.95	99.56	Otsu	140
0.9	99.99	100.00	Almeida-Lins-Lima	126
0.9	99.98	99.20	IsoData	137
0.9	99.97	99.53	Otsu	138
0.9	99.07	100.00	Silva-Lins	167
0.8	99.97	100.00	Almeida-Lins-Lima	126
0.8	99.94	99.23	IsoData	137
0.8	99.94	99.56	Otsu	138
0.8	98.98	100.00	Silva-Lins	161
0.7	99.94	100.00	Almeida-Lins-Lima	126
0.7	99.25	100.00	Kapur SW	147
0.7	99.87	99.54	Otsu	138
0.6	99.91	100.00	Almeida-Lins-Lima	126
0.5	99.87	100.00	Almeida-Lins-Lima	126
0.4	99.60	100.00	Almeida-Lins-Lima	126

5 CONCLUSIONS

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. This paper presents an assessment methodology based on the controlled generation of a large set of synthetic images that allows identifying quality aspects of the binarized images.

Eleven different binarization algorithms presented in this paper were used to binarize the images in the test set database of 1,478,400 binary images that were compared with the 134,400 ground truth images, allowing to know for each of them the percentage and type of matching (P(b|b)% and P(f|f)%) and mismatched (P(b|f)% and P(f|b)%) pixels.

The authors plan to develop an image "matcher" or "classifier" that will be trained with the database developed of synthetic images. The aim of such classifier is that, given a real-world document, the platform will automatically find the closest synthetic document to it. Once that document is found, one knows the set of binarization algorithms that are more likely to provide the best results. One important point is worth remarking here is that binarization assessments tend only to consider the quality of the resulting image "for visual inspection". In the more global assessment methodology presented here, the user will be even able to choose to prioritize to minimize either the P(b|f)% or P(f|b)%) errors, depending on the "sensitiveness" of the target application. For instance, if the resulting binary image will go through an OCR it may be better to have P(f|b)%) < P(b|f)%.

Preliminary tests made in matching the synthetic images with "real world" documents for "visual inspection" provided very good results. The image shown in Figure 8 may witness the good quality of the binary image provided by using the Almeida-Lins-Lima algorithm in the document image presented in Figure 3. The document image in Figure 9 provides another evidence of that, using the same binarization algorithm.

The assessment strategy presented here is a generalization of the platform described in reference [20]. The current version of the assessment environment encompasses 5,554,00 images (231 groud-truth x 2 blur x 100 α-fading-coefficients x 3 offsets x 100 textures-patterns x 2 texture generation schemes x 2 color/grayscale). The authors of this paper consider this image set representative of the universe or "real world" text documents. At present, twenty-five binarization algorithms are being assessed. Another relevant aspect that should be taken into account is that the proposed binarizarion platform accounts now for the time elapsed by each algorithm to binarize each image. This allows the user to choose the lightest algorithm that provides the best results. For instance, the computational cost of Otsu is extremely small if compared with the IsoData or the Almeida-Lins-Lima algoritms. At a later stage, space consumed will also be considered.

It is most relevant to emphasize the computational challenge involved in the task proposed here, as each of the synthetic images is over 10 MB large. If one attempts to store the 5,554,000 images, over 50 TB of storage would be needed, a volume of data unreasonable to be used. Each image is generated a time and then binarized in a pipeline with the 25 filters currently tested against the ground-truth image and the data is collected and stored. A slice of the image that corresponds to central one-fifth of it is being saved as a lossless PNG image to later be used in the training of the image matcher. A cluster with 10 machines is being used in this platform, using the technology described in the BigBatch project [21]. Priority was given to four different values of alpha (α=1 no interference, α=0.8 weak interference, α=0.6 medium interference, α=0.4 strong back-to-front interference). The partial assessment results will be made publically available as they are obtained. The authors would like to remark that even processing in a dedicated cluster with ten nodes, several months of processing are needed. The preliminary version of the DIB-Document Image Binarization platform and website is publically available at www.cin.ufpe.br/~dib.

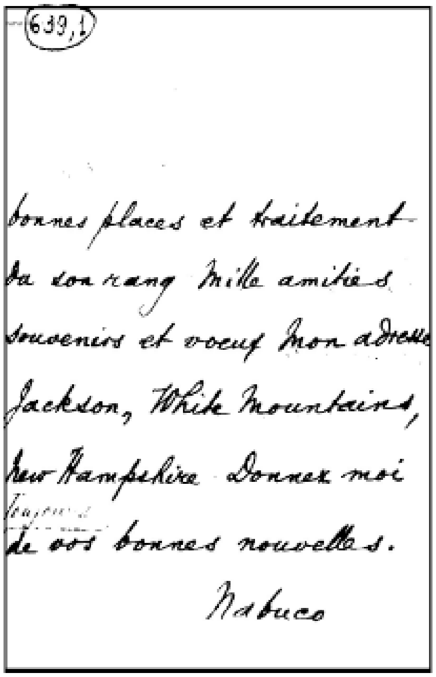


Figure 8 –Binarized version of the document shown in Figure 3 using the Almeida-Lins-Lima algorithm.

ACKNOWLEDGMENTS

The authors of this paper are grateful for those who made the code of their algorithms publically available for testing and performance analysis and to CNPq – Brazilian Government for sponsoring this research.

REFERENCES

[1] K. Ntirogiannis, B. Gatos and I. Pratikakis, Performance Evaluation Methodology for Historical Document Image Binarization, IEEE Trans. Image Proc., vol.22, no.2, pp. 595-609, Feb. 2013..

[2] R. D. Lins et al. An Environment for Processing Images of Historical Documents. Microproc. and Microprogramming, 111–121, 1995.

[3] G. Sharma. Show-trough cancellation in scans of duplex printed documents. IEEE Transaction Image Processing, v. 10, n. 5, p. 736–754, 2001.

[4] R. D. Lins. Nabuco – Two Decades of Processing Historical Documents in Latin America. Journal of Universal Computer Science. , March 2011.

[5] C. A. B. Mello and R. D. Lins. 2002. Generation of Images of Historical Documents by Composition. Symposium on Document Engineering, 127–133. 2002.

[6] M.Sezgin and B.Sankur. A Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. Journal of Electronc Imaging, v. 1, n. 13, p. 146–165, 2004.

[7] J. N. Kapur, P. K. Sahoo, A. K. C. Wong. A New Method for Gray-Level Picture Thersholding Using the Entropy of the Histogram. C. Vision Graphics and Image Processing, v. 29, p. 273–285, 1985.

[8] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. IEEE Transaction on Systems, Man and Cybernetics, v. SMC-9, n. 1, p. 62–66, 1979.

[9] G. Johannsen and J. A. Bille. A Threshold Selection Method Using Information Measure. ICPR'82 - Proceeding 6th International Conference on Pattern Recognition, 140–143. 1982.

[10] J. C. Yen, F. J. Chang, S. Chang. 1995. A New Criterion for Automatic Multilevel Thresholding. IEEE Transaction Image Process IP-4, 370–378.

[11] U. L. Wu, A. Songde, L. U. Haqing. 1998. An Effective Entropic Thresholding for Ultrasonic Imaging. International Conference Pattern Recognition, 1522–1524.

[12] C. A. B. Mello and R. D. Lins. Generation of Images of Historical Documents by Composition. Proceedings of the 2002 ACM symposium on Document engineering, 127–133, 2002.

[13] J. M. M. Silva, R. D. Lins, V. C. Rocha. Binarizing and Filtering Historical Documents with Back-to-Front Interference. ACM Symposium on Applied Computing, 853–858, 2006.

[14] E. Roe and C. A. B. Mello. 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.

[15] M. A. M. de Almeida, R. D. Lins, B. C. Lima, A New Binarization Algorithm for Images with Back-to-Front Interference. Submitted for publication, 2017.

[16] S. Paris, P. Kornprobst, J. Tumblin and F. Durand. Bilateral Filtering: Theory and Applications. Foundations and Trends in Computer Graphics and Vision. Vol. 4, No. 1, 1–73. 2008.

[17] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. SIGGRAPH '01 28th annual conference on Computer graphics and interactive techniques, 341-346. 2001.

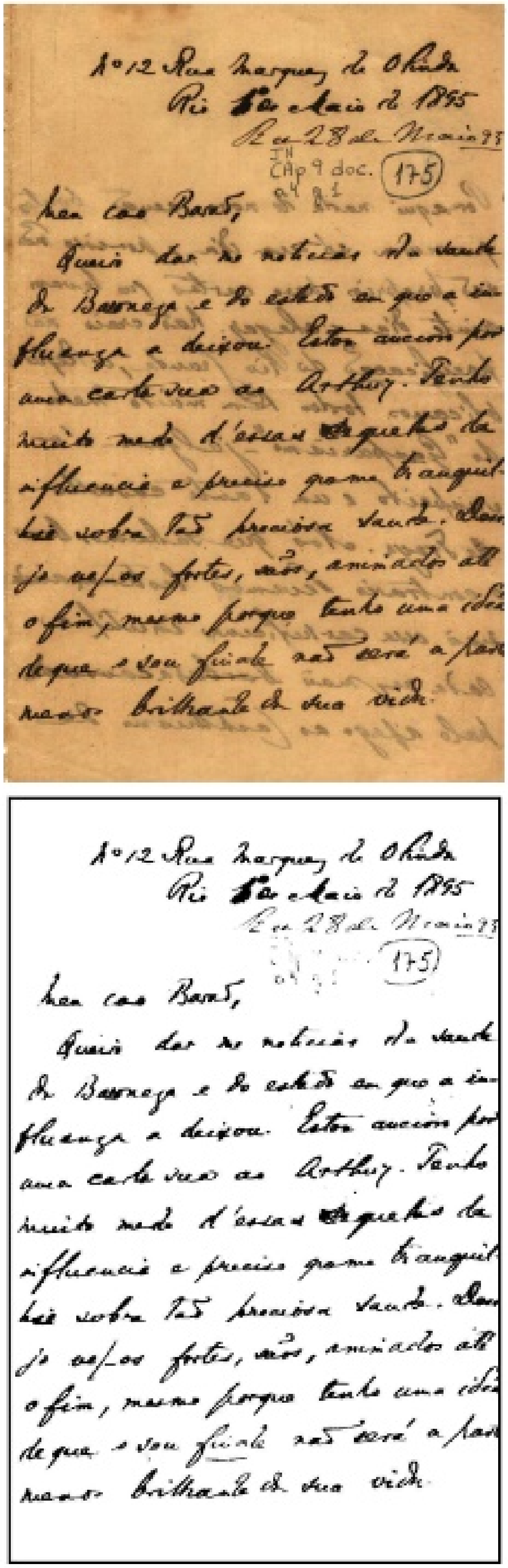
[18] [N. Memarsadeghi, D. M. Mount, N. S. Netanyahu, J. Moigne. 2007. A Fast Implementation of the IsoData Clustering Algorithm. International Journal of Computational Geometry and Applications, 71–103.

[19] T. Pun. Entropic Thresholding, A New Approach. Computer Vision Graphics and Image Processing, 210–239, 1981.

[20] R. D. Lins and G. F. P e Silva. Assessing Strategies to Remove Back-to-Front Interference in Color Documents. IEEE International Telecommunications Symposium, 2010, IEEE Press, p. 1-6, 2010.

[21] G. G.Mattos, A. A. Formiga, R. D. Lins, F. M. J. Martins. BigBatch: a document processing platform for clusters and grids. ACM-SAC 2008. ACM Press, 2008. v. I. p. 434-441.

[22] Scikit-learn. http://scikit-learn.org/stable/ (visited: 31st May 2017)



Nº 12 Rua Marquez de Olinda
Rio 5 de Maio de 1895
Rec. 28 de Maio 95
IN
CAp 9 doc.
175

Meu caro Barão,

Queria dar-te noticias da saude da Baroneza e do estado em que a influenza a deixou. Estou ancioso por uma carta sua ao Arthur. Tenho muito medo d'essas sequelas da influencia e preciso para tranquilisar-me sobre tão preciosa saude. Desejo-lhes fortes, nós, animados até o fim, mesmo porque tenho uma idéa de que o seu final não será a parte menos brilhante da sua vida.

Figure 9 –Historic document and its binary version produced by Almeida-Lins-Lima algorithm.