Journal of *Imaging*

MDPI

Article

A New Binarization Algorithm for Historical Documents

Marcos Almeida [1,*], Rafael Dueire Lins [1,2], Rodrigo Bernardino [1], Darlisson Jesus [1] and Bruno Lima [1]

[1] Federal University of Pernambuco, Recife-PE 50740-560, Brazil; rdl.ufpe@gmail.com (R.D.L.); rbbernardino@gmail.com (R.B.); dmj.ufpe@gmail.com (D.J.); brunocesar182@hotmail.com (B.L.)

[2] Federal Rural University of Pernambuco, Recife-PE 52171-900, Brazil

* Correspondence: mmar@ufpe.br; Tel.: +55-81-2126-7129

Received: 31 October 2017; Accepted: 16 January 2018; Published: 23 January 2018

Abstract: Monochromatic documents claim for much less computer bandwidth for network transmission and storage space than their color or even grayscale equivalent. The binarization of historical documents is far more complex than recent ones as paper aging, color, texture, translucidity, stains, back-to-front interference, kind and color of ink used in handwriting, printing process, digitalization process, etc. are some of the factors that affect binarization. This article presents a new binarization algorithm for historical documents. The new global filter proposed is performed in four steps: filtering the image using a bilateral filter, 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. The quantitative and qualitative assessment made with 23 binarization algorithms in three sets of "real world" documents showed very good results.

Keywords: documents; binarization; back-to-front interference; bleeding

1. Introduction

Document image binarization plays an important role in the document image analysis, compression, transcription, and recognition pipeline [1]. Binary documents claim for far less storage space and computer bandwidth for network transmission than color or grayscale documents. Historical documents drastically increase the degree of difficulty for binarization algorithms. Physical noises [2] such as stains and paper aging affect the performance of binarization algorithms. Besides that, historical documents were often typed, printed or written on both sides of sheets of paper and the opacity of the paper is often such as to allow the back printing or writing to be visualized on the front side. This kind of "noise", first called back-to-front interference [3], was later known as bleeding or show-through [4]. Figure 1 presents three examples of documents with such a noise extracted from the three different datasets used in this paper in the assessment of the proposed algorithm. If the document is exhibited either in true-color or gray-scale, the human brain is able to filter out that sort of noise keeping its readability. The strength of the interference present varies with the opacity of the paper, its permeability, the kind and degree of fluidity of the ink used, its storage, age, etc. Thus, the difficulty for obtaining a good binarization performance capable of filtering-out such a noise increases enormously, as a new set of hues of paper and printing colors appear. The direct application of binarization algorithms may yield a completely unreadable document, as the interfering ink of the backside of the paper overlaps with the binary one in the foreground. Several document image compression schemes for color images are based on "adding color" to a binary image. Such compression strategy is unable to handle documents with back-to-front

interference [5]. Optical Character Recognizers (OCRs) are also unable to work properly for such documents. Several algorithms were developed specifically to binarize documents with back-to-front interference [3,4,6–9]. There is no binarization technique to be an all case winner as many parameters may interfere in the quality of the resulting image [9]. The development of new binarization algorithms is still an important research topic. International competitions on binarization algorithms, such as DIBCO - Document Image Binarization Competition [10], are an evidence of the relevance of this area.

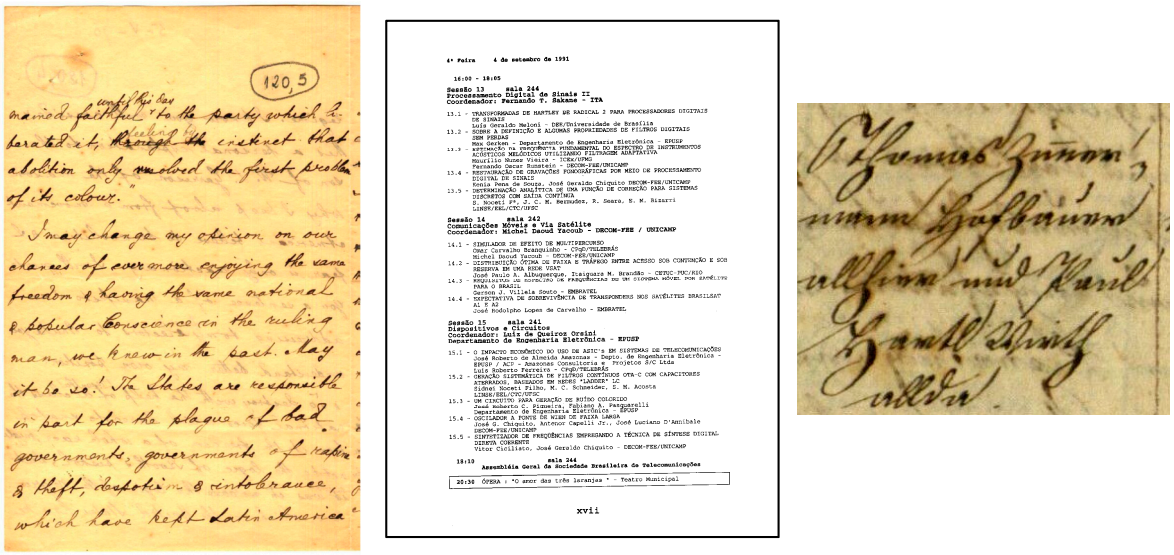

Figure 1. Images with back-to-front interference from the three test sets used in this paper: Nabuco bequest (**left**), LiveMemory (**center**) and DIBCO (**right**).

This paper presents a new global filter [1] to binarize documents, which is able to remove the back-to-front noise in a wide range of documents. Quantitative and qualitative assessments made in a wide variety of documents from three different "real-world" datasets (typed, printed and handwritten, using different kinds of paper, ink, etc.) allow to witness the efficiency of the proposed scheme.

2. The New Algorithm

The algorithm proposed here is performed in four steps: 1. decision-making for finding the vector of parameters of the image to be filtered, 2. filtering the image using a bilateral filter, 3. splitting the image into the RGB components, and performing their binarization using a method inspired by Otsu's algorithm for each RGB channel, and 4. choice of which of the RGB components best preserved the document information in the foreground, which is considered the final output of the algorithm. Figure 2 presents the block diagram of the proposed algorithm. The functionality of each block is detailed as follows.

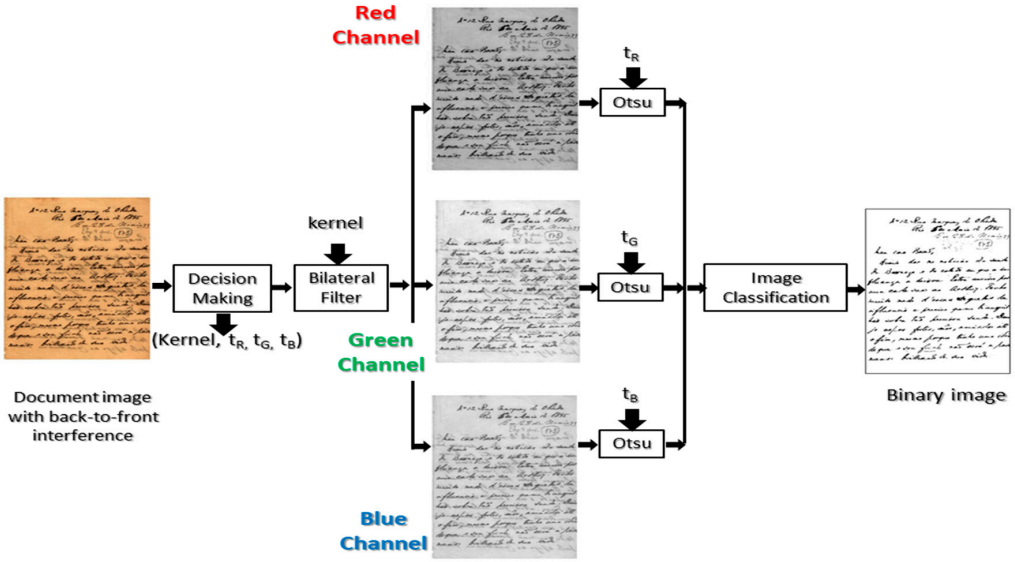

Figure 2. Block diagram of the proposed algorithm.

2.1. The Decision Making Block

The decision making block takes as input the image to be binarized and outputs a vector with four parameters: the value of the kernel (*kernel*) for the bilateral filter and three threshold values (t_R, t_G, t_B) that will be later used in the modified Otsu filtering.

The training of the binarization process proposed here is made with synthetic images which were generated as explained in Section 2.2. After filtering, the matrix of co-occurrence probabilities between the original image and of the binary image was calculated for each of the images in the document training set, whose generation is explained below.

The probabilistic structure applied in the analysis to each of the images in the training set is similar to the transmission of binary data in a Binary Asymmetric Channel, as shown in Figure 3. The probabilities P(f/b) and P(b/f) represent an additive noise in communication channels in information theory, here it represents the inability of the algorithm to correct the back-to-front interference of the image tested in the binarization process. The probabilities P(b/b) and P(f/f) are calculated from the pixel-to-pixel comparison of the binarized image generated by the proposed algorithm with the ground-truth image.

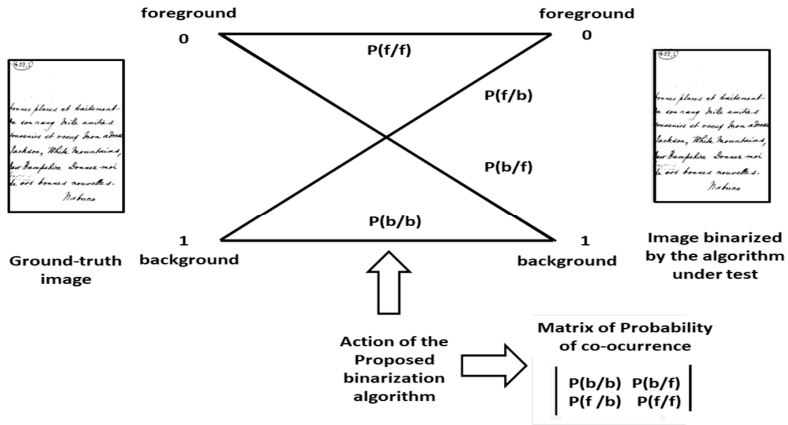

Figure 3. Generation of the co-occurrence matrix for each of the images in the training set.

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:

$$P(b/b) = f(\alpha,R,G,B). \quad (1)$$

The optimal threshold t_c* for each channel is calculated in the decision-making block, the index c can be R, G or B, maximizing P(b/b):

$$t_c^* = MaxP(b/b), \quad (2)$$

subject to a given criterion $P(f/f) \geq M$. The criterion used here was M = 97%, that is at most 3% of the foreground pixels may be incorrectly mapped. During the training phase, the best t_c* will be chosen from the three channels, which best maximizes the P(b/b) for each of the images in the training set. The matrix of co-occurrence probability is calculated and the decision maker chooses the best binary image. The decision-making block was trained with 32,000 synthetic images in such a way to, given a real image to be binarized, it finds the optimal threshold parameters.

2.2. Generating Synthetic Images

The Decision-Making Block needs training to "learn" about the optimal threshold parameters and the value of the kernel to be used in the bilateral filter. Such training must be done using controlled images which are synthesized to mimic the different degrees of back-to-front interference, paper aging, paper translucidity, etc. Figure 4 presents the block diagram for the generation of synthetic images. Two binary images of documents of different nature (typed, handwritten with different pens, printed, etc.) are taken: F—front and V—verso (back). The front image is blurred with a weak Gaussian filter to simulate the digitalization noise [1], the hues that appear in after document scanning.

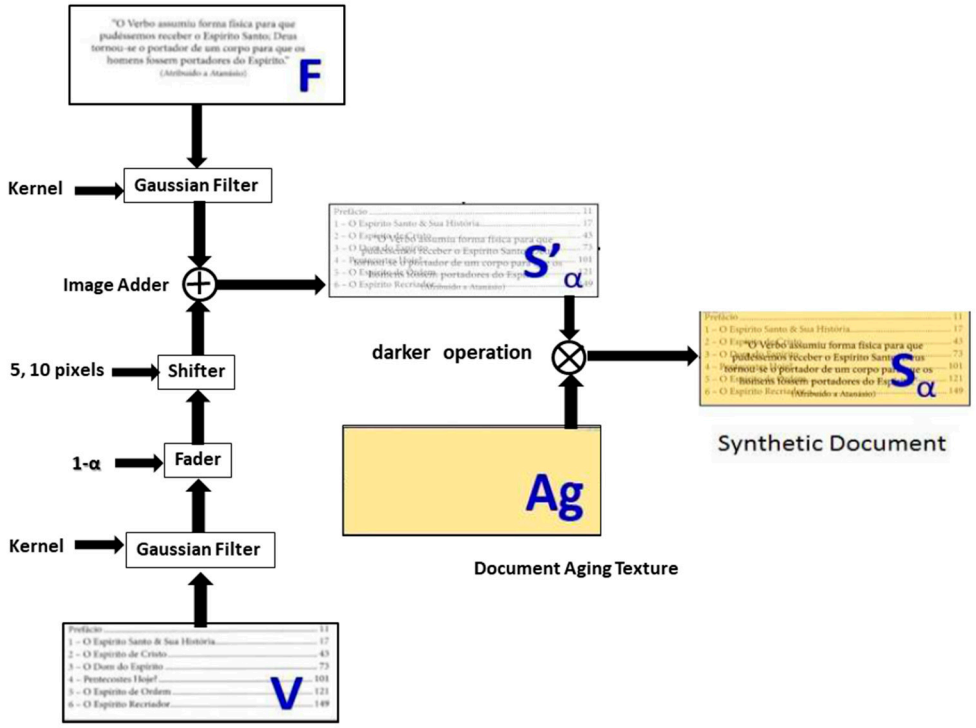

Figure 4. Block diagram of the scheme for the generation of synthetic images for the Decision-Making Block.

The verso image is "blurred" by passing through two different Gaussian filters that simulate the low-pass effect of the translucidity of the verso as seen in the front part of the paper. Two different parameters were used to simulate two different classes of paper translucidity. The "blurred" verso image is now faded with a coefficient α varying between 0 and 1 in steps of 0.01. Then, a circular shift of the lines of the document is made of either 5 or 10 pixels, to minimize the chances of the front and verso lines coincide entirely. Finally, 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 texture pattern was extracted from document from late 19th century to the year 2000. The analysis of 3450 documents representative of a wide variety of documents of such a period was analyzed yielding 100 different clusters of textures. The synthetic texture to be applied to the image to simulate paper aging is generated using those 100 clusters by image quilting [11] and randomly, as explained in reference [9]. The training performed in the current version of the presented algorithm was made with 16 of those 200 synthetic textures. The total number of images used for training here was thus 16 (textures), times 10 ($0 < \alpha < 1$ in steps of 0.10), times 2 blur parameters for the Gaussian filters, times 100 different binary images, totaling 32,000 images. Details of the full generation process of the synthetic image database are out of the scope of this paper and may be found in reference [9].

2.3. The Bilateral Filter

The bilateral filter was first introduced by Aurich and Weule [12] under the name "nonlinear Gaussian filter". It was later rediscovered by Tomasi and Manduchi [13] who called it the "bilateral filter" which is now the most commonly used name according to reference [14].

The bilateral filter is a 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, while 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 the pixels in a window of size N × N. Let I(x,y) be a 2D discrete image of size N × N, such that {x,y} ∈ {0, 1, ..., N − 1} X {0, 1, ..., N − 1}. Assume that I(x,y) is corrupted by an additive white Gaussian noise of variance σ_n^2. For a pixel (x,y), the output of a bilateral filter can be as described by Equation (1):

$$I_{BF}(\mathrm{x,y}) = \frac{1}{K}\sum_{i=\mathrm{x}-\mathrm{d}}^{x+d}\sum_{j=\mathrm{y}-\mathrm{d}}^{\mathrm{x}+d} G_s(i;x,j;y)G_r[I(i,j),I(x,y)]I(i,j), \tag{3}$$

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 and d is a non-negative integer such that (2d + 1) × (2d + 1) stands for the size of the neighborhood window. Let G_s and G_r be the domain and the range components, respectively, which are defined as:

$$G_s(i;x,\ j;y) = e^{-\frac{|(i-x)^2+(j-y)^2|}{2\sigma_s^2}} \tag{4}$$

and

$$G_r(I(i,j);I(x,y)) = e^{-\frac{|I(i,j)-I(x,y)|^2}{2\sigma_r^2}} \tag{5}$$

The normalization constant K is given as:

$$\mathrm{K} = \frac{1}{\sum_{i=\mathrm{x}-\mathrm{d}}^{x+d} \sum_{j=\mathrm{y}-\mathrm{d}}^{\mathrm{x}+d} G_s(i;x,j;y) G_r[I(i,j), I(x,y)]} \tag{6}$$

Equations (4) and (5) show that the bilateral filter has three parameters: σ_s^2 (the filter domain), σ_r^2 (the filter range), and the third parameter is the window size N × N [15].

The geometric spread of the bilateral filter is controlled by σ_s^2. If the value of σ_s^2 is increased, more neighbours are combined in the diffusion process yielding a "smoother" image, while σ_r^2 represents the photometric spreading. Only pixels with a percentage difference of less than σ_r^2 are processed [13].

2.4. Otsu Filtering

After passing through the bilateral filter, the image is split into its original (non-gamma corrected) Red, Green and Blue components, as shown in the block diagram in Figure 2. The kernel of the bilateral filter alters the balance of the colors in the original image in such a way to widen the differences between the color of the front and back-to-front interference. A modified version of Otsu [16] algorithm is applied to each RGB channel using the thresholds determined by the Decision Making Block, which may be considered as the "optimal" threshold for each RGB channel, and then three binary images are generated.

2.5. Image Classification

The image classification block was also trained with the synthetic images in such a way to analyze the three binary images generated in each of the channels and outputs the one that is considered the best one. This decision was also made by a naïve Bayes automatic classifier which was trained using the calculated co-occurrence matrix for each of the 32,000 synthetic images by comparing each of them with the original ground truth image, the Front image.

3. Experiments and Results

As already explained, the enormous variety of kinds of text documents makes extremely improbable that one single algorithm is able to satisfactorily binarize all kinds of documents. Depending on the nature (or degree of complexity) of the image several or no algorithm will be able to provide good results. This paper follows the assessment methodology proposed in reference [9], in which one compares the numbers of background and foreground pixels correctly matched with a ground-truth image. Twenty-three binarization algorithms were tested using the methodology described:

1. Mello-Lins [5]
2. DaSilva-Lins-Rocha [6]
3. Otsu [16]
4. Johannsen-Bille [17]
5. Kapur-Sahoo-Wong [18]
6. RenyEntropy (variation of [18])
7. Li-Tam [19]
8. Mean [20]
9. MinError [21]
10. Mixture-Modeling [22]
11. Moments [23]
12. IsoData [24]
13. Percentile [25]

14. Pun [26]
15. Shanbhag [27]
16. Triangle [28]
17. Wu-Lu [29]
18. Yean-Chang-Chang [30]
19. Intermodes [31]
20. Minimum (variation of [31])
21. Ergina-Local [32]
22. Sauvola [33]
23. Niblack [34]

A ground-truth image for each "real" world one is needed to allow a quantitative assessment of the quality of the final binary image. Only the DIBCO dataset [10] had ground-truth images available. This makes the assessment task of real-world images extremely difficult [35]. All care must be taken to guarantee the fairness of the process. The ground-truth images for the other datasets were generated by applying the 23 algorithms above and the bilateral algorithm to all the test images in the Nabuco [7] and LiveMemory [36] datasets. Visual inspection was made to choose the best binary image in a blind process, a process in which the people who selected the best image did not know which algorithm generated it. To increase the degree of fairness and the number of filtering possibilities, the three component images produced by the Decision Making block were all analyzed. The binary images chosen using the methodology above went through salt-and-pepper filtering and were used as ground-truth image for the assessment below. All the processing time figures presented in this paper are from Intel i7-4510U@ 2.00 GHzx2, 8 GB RAM, running Linux Mint 18.2 64-bit. All algorithms were coded in Java, possibly by their authors.

3.1. The Nabuco Dataset

The Nabuco bequest encompasses about 6500 letters and postcards written and typed by Joaquim Nabuco [7], totaling about 30,000 pages. Such documents are of great interest to whoever studies the history of the Americas, as Nabuco was one of the key figures in the freedom of black slaves, and was the first Brazilian Ambassador to the U.S.A. The documents of Nabuco were digitalized by the second author of this paper and the historians of the Joaquim Nabuco Foundation using a table scanner in 200 dpi resolution in true color (24 bits per pixel), back in 1992 to 1994. Due to serious storage limitations then, images were saved in the jpeg format with 1% loss. The historians in the project concluded that 150 dpi resolution would suffice to represent all the graphical elements in the documents, but choice of the 200-dpi resolution was made to be compatible with the FAX devices widely used then. About 200 of the documents in the Nabuco bequest exhibited back-to-front interference. The 15 document images used in this dataset were chosen for being representative of the diversity of documents in such a universe.

Table 1 presents the quantitative results obtained for all the documents in this dataset. P(f/f) stands for the ratio between the number of foreground pixels in the original image mapped onto black pixels and the number of black pixels in the ground-truth image. Similarly, P(b/b) is proportion between the number of background pixels in the original image mapped onto white pixels of the binary image and the number of white pixels in the ground-truth image. The figures for P(b/b) and P(f/f) are followed by "$\pm$" and the value of the standard deviation. The time corresponds to the mean processing time elapsed by the algorithm to process the images in this dataset. The results were ranked in P(b/b) decreasing order.

The results presented in Table 1 shows the bilateral filter in third place for this dataset in terms of image quality, however the standard deviation is much lower than the two first. That implies that its quality is more stable for the various document images in this dataset. Figure 5 presents the document

for which the bilateral filter presented the best and the worst results in terms of image quality with two zoomed areas from the original and the binarized document.

Table 1. Binarization results for images from Nabuco bequest.

AlgName	P(f/f)	P(b/b)	Time (s)
IsoData	98.08 ± 3.39	99.38 ± 0.60	0.0171
Otsu	98.08 ± 3.39	99.36 ± 0.63	0.0159
Bilateral	99.57 ± 1.23	99.29 ± 0.93	1.0790
Huang	99.40 ± 2.14	98.69 ± 0.88	0.0200
Moments	99.39 ± 1.34	98.40 ± 1.70	0.0160
Ergina-Local	99.99 ± 0.03	98.13 ± 0.64	0.3412
RenyEntropy	100.00	97.56 ± 1.17	0.0188
Kapoo-Sahoo-Wong	100.00	97.51 ± 1.07	0.0172
Yean-Chang-Chang	100.00	97.38 ± 1.26	0.0161
Triangle	100.00	95.94 ± 1.46	0.0160
Mello-Lins	98.61 ± 5.14	89.63 ± 24.43	0.0160
Mean	100.00	81.77 ± 5.99	0.0168
Johannsen-Bille	98.87 ± 2.97	59.77 ± 48.80	0.0164
Pun	100.00	55.44 ± 2.57	0.0185
Percentile	100.00	53.21 ± 1.33	0.0185
Sauvola	85.51 ± 12.93	99.95 ± 0.11	1.2977
Niblack	99.75 ± 0.34	77.06 ± 5.63	0.2135

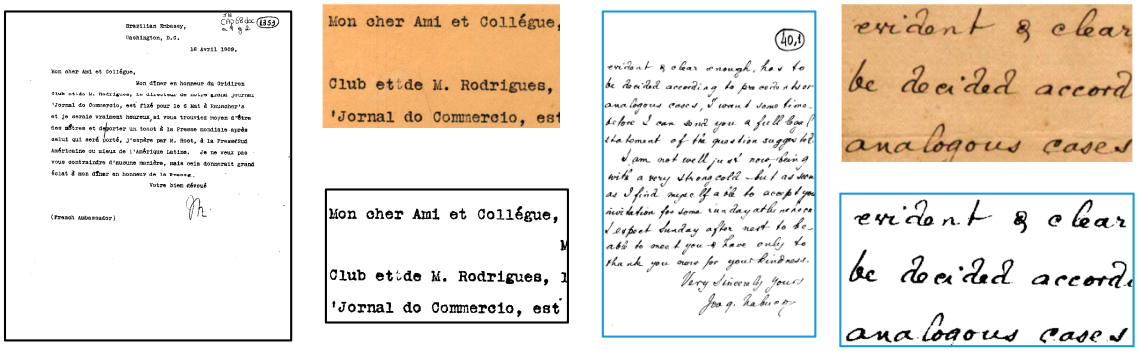

Figure 5. Historical documents from Nabuco bequest with the best ((**left**)—P(f/f) = 100, P(b/b) = 99.99) and the worst ((**right**)—P(f/f) = 89.76, P(b/b) = 99.98) binarization results for the bilateral filter with zooms from the original (**top**) and binary (**bottom**) parts.

3.2. The LiveMemory Dataset

This dataset encompasses 15 documents with 200 dpi resolution selected from the over 8000 documents from the LiveMemory project that created a digital library with all the proceedings of technical events from the Brazilian Telecommunications Society. The original proceedings were offset printed from documents either typed or electronically produced. Table 2 presents the performance results for the 12 best ranked algorithms. The bilateral filter obtained the best results in terms of image filtering. It is worth observing that in the case of the worst quality image (Figure 6, right) the performance degraded for all the algorithms. This behavior is due to the shaded area in the hard-bound spine of the volumes of the proceedings.

Table 2. Binarization results for images from the LiveMemory project.

AlgName	P(f/f)	P(b/b)	Time (s)
Bilateral	100.00	98.90 ± 1.07	3.3325
IsoData-ORIG	99.56 ± 0.69	98.61 ± 1.99	0.0734
Otsu	99.60 ± 0.68	98.57 ± 2.08	0.0735
Moments	99.99 ± 0.03	97.91 ± 1.87	0.0716
Ergina-Local	98.98 ± 2.82	97.62 ± 1.04	0.9917
Huang	99.93 ± 0.27	96.42 ± 4.20	0.0865
Triangle	100.00	94.24 ± 2.15	0.0728
Mean	100.00	83.58 ± 5.59	0.0747
Niblack	99.76 ± 0.76	78.31 ± 2.97	0.6710
Pun	100.00	55.28 ± 3.60	0.0800
Percentile	100.00	53.91 ± 1.96	0.0795
Kapur-Sahoo-Wong	98.62 ± 4.92	97.15 ± 1.44	0.0729

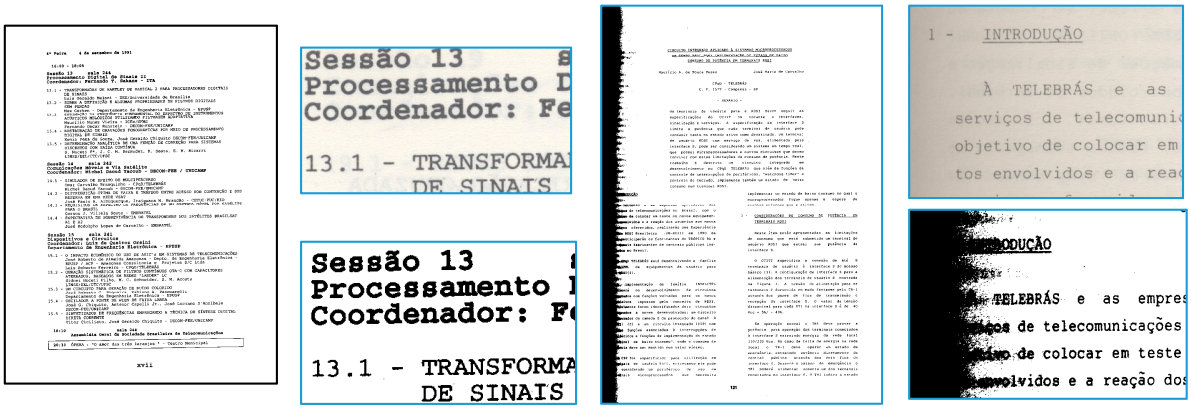

Figure 6. Images from LiveMemory with the best ((**left**)—P(f/f) = 100.00, P(b/b) = 99.99) and the worst ((**right**)—P(f/f) = 100.00, P(b/b) = 95.97) binarization results for the bilateral filter with zooms from the original (**top**) and binary (**bottom**) parts.

3.3. The DIBCO Dataset

This dataset has all the 86 images from the Digital Image Binarization Contest from 2009 to 2016. Table 3 presents the results obtained. The performance of the bilateral filter in this set may be considered good, in general. The overall performance of the bilateral filter was strongly degraded by the single image shown in Figure 7 (right) in which the P(f/f) of 25.93 drastically dropped the average result of the algorithm in this test set. It is important to remark that such an image is almost unreadable even for humans and that it degraded the performance of all the best algorithms.

Table 3. Binarization results for images from Document Image Binarization Competition (DIBCO).

AlgName	P(f/f)	P(b/b)	Time (s)
Ergina-local	91.37 ± 6.25	99.88 ± 1.89	0.1844
RenyEntropy	90.13 ± 14.19	96.77 ± 3.50	0.0125
Yean-Chang-Chang	90.61 ± 14.44	96.16 ± 4.35	0.0112
Moments	90.75 ± 9.91	95.80 ± 5.19	0.0112
Bilateral	92.99 ± 9.06	90.78 ± 16.01	0.6099
Huang	95.62 ± 6.37	84.22 ± 18.36	0.0147
Triangle	96.40 ± 5.72	80.80 ± 23.32	0.0113
Mean	99.35 ± 1.14	78.99 ± 9.35	0.0115
MinError	92.79 ± 23.46	74.29 ± 19.36	0.0115
Pun	99.68 ± 0.82	56.20 ± 6.18	0.0122
Percentile	99.71 ± 0.72	55.06 ± 3.58	0.0121
Sauvola	59.75 ± 30.06	99.58 ± 079	0.6933
Niblack	95.91 ± 2.31	78.61 ± 5.69	0.1241

4. Conclusions

Historical documents are far more difficult to binarize as several factors such as paper texture, aging, thickness, translucidity, permeability, the kind of ink, its fluidity, color, aging, etc. all may influence the performance of the algorithms. Besides all that, many historical documents were written or printed on both sides of translucent paper, giving rise to the back-to-front interference.

This paper presents a new binarization scheme based on the bilateral filter. Experiments performed in three datasets of "real world" historical documents with twenty-three other binarization algorithms. Image quality and processing time figures were provided, at least for the top 10 algorithms assessed. The results obtained showed that the proposed algorithm yields good quality monochromatic images that may compensate its high computational cost. This paper provides evidence that no binarization algorithm is an "all-kind-of-document" winner, as the performance of the algorithms varied depending of the specific features of each document. A much larger test set of synthetic about 250,000 images is currently under development, such a test set will allow much better training of the Decision Making and Image Classifier blocks of the bilateral algorithm presented. The authors are currently attempting to integrate the Decision Making and Image Classifier blocks in such a way to anticipate the choice of the best component image. This would highly improve the time performance of the proposed algorithm.

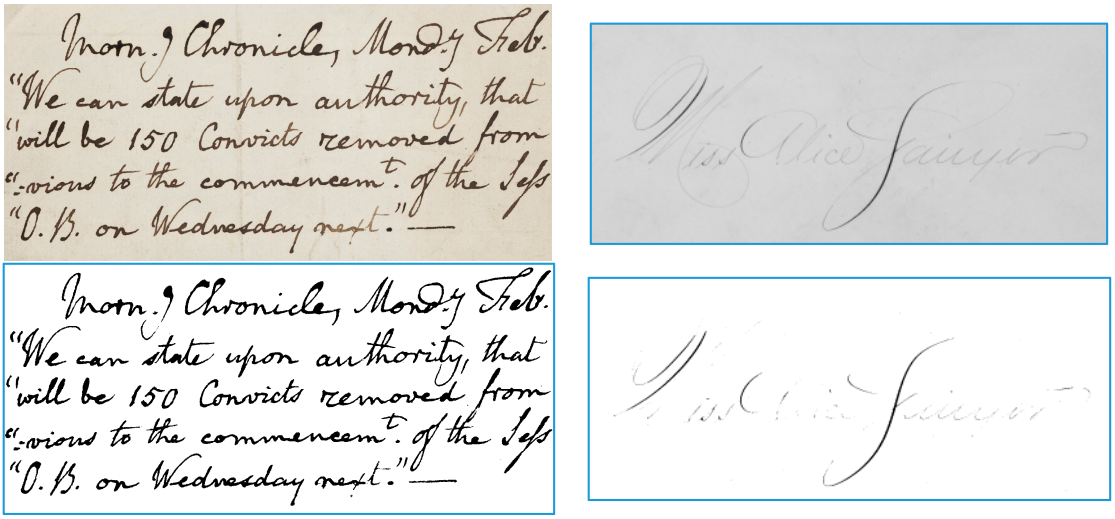

Figure 7. Two documents from DIBCO dataset: (**left-top**) original image (**left-bottom**) binary image obtained using the bilateral filter best result (P(f/f) = 97.05, P(b/b) = 99.88); (**right-top**) original image. (**right-bottom**) the worst binarization results for the bilateral filter (P(f/f) = 25.93, P(b/b) = 99.99).

The authors of this paper are promoting a paramount research effort to assess the largest possible number of binarization algorithms for scanned documents using over 5.4 million synthetic images in the DIB-Document Image Binarization platform. An image matcher, a more general and complex version of the Decision Making block, is also being developed and trained with that large set of images, in order to whenever fed with a real world image, to be able to match with the most similar synthetic one. Once that match is made, the most suitable binarization algorithms are immediately known. If this paper were accepted, all the test images and algorithms will be included in the DIB platform. The preliminary version of the DIB-Document Image Binarization platform and website is publicly available at https://dib.cin.ufpe.br/.

Acknowledgments: The authors of this paper are grateful for the referees whose comments much helped in improving the current version of this paper and to those researchers who made the code of their algorithms publicly available for testing and performance analysis and to the DIBCO team from making their images publicly available. The authors also acknowledge the partial financial support of to CNPq and CAPES—Brazilian Government.

Author Contributions: Marcos Almeida and Rafael Dueire Lins contributed in equal proportion to the development of the algorithm presented in this paper, which was written by the latter author. Bruno Lima was responsible for the first implementation of the algorithm proposed. Rodrigo Bernardino and Darlisson Jesus re-implemented the algorithm and were also responsible for all the quality and time assessment figures presented here.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chaki, N.; Shaikh, S.H.; Saeed, K. *Exploring Image Binarization Techniques*; Springer: New Delhi, India, 2014.
2. Lins, R.D. A Taxonomy for Noise in Images of Paper Documents-The Physical Noises. In Proceedings of the International Conference Image Analysis and Recognition, Halifax, NS, Canada, 6–8 July 2009; Volume 5627, pp. 844–854.
3. Lins, R.D. An Environment for Processing Images of Historical Documents. *Microprocess. Microprogr.* **1995**, *40*, 939–942. [CrossRef]
4. Sharma, G. Show-through cancellation in scans of duplex printed documents. *IEEE Trans. Image Process.* **2001**, *10*, 736–754. [CrossRef] [PubMed]
5. Mello, C.A.B.; Lins, R.D. Generation of Images of Historical Documents by Composition. In Proceedings of the 2002 ACM Symposium on Document Engineering, New York, NY, USA, 8–9 November 2002; pp. 127–133.
6. Silva, M.M.; Lins, R.D.; Rocha, V.C. Binarizing and Filtering Historical Documents with Back-to-Front Interference. In Proceedings of the 2006 ACM Symposium on Applied Computing, New York, NY, USA, 23–27 April 2006; pp. 853–858.
7. Lins, R.D. Nabuco—Two Decades of Processing Historical Documents in Latin America. *J. Univers. Comput. Sci.* **2011**, *17*, 151–161.
8. Roe, E.; Mello, C.A.B. Binarization of Color Historical Document Images Using Local Image Equalization and XDoG. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 205–209.
9. Lins, R.D.; Almeida, M.A.M.; Bernardino, R.B.; Jesus, D.; Oliveira, J.M. Assessing Binarization Techniques for Document Images. In Proceedings of the ACM Symposium on Document Engineering, Valletta, Malta, 4–7 September 2017.
10. Pratikakis, I.; Zagoris, K.; Barlas, G.; Gatos, B. ICDAR 2017 Competition on Document Image Binarization (DIBCO 2017). In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan, 13–15 November 2017; pp. 2140–2379.
11. Efros, A.A.; Freeman, W.T. Image quilting for texture synthesis and transfer. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01), New York, NY, USA, 12–17 August 2001; pp. 341–346.
12. Aurich, V.; Weule, J.B. Non-Linear Gaussian Filters Performing Edge Preserving Diffusion. In Proceedings of the DAGM Symposium, London, UK, 13–15 September 1995; pp. 538–545.
13. Tomasi, C.; Manduchi, R. Bilateral Filtering for Gray and Color Images. In Proceedings of the 6th International Conference on Computer Vision, Washington, DC, USA, 4–7 January 1998; pp. 836–846.
14. Paris, P.; Kornprobst, P.; Tumblim, J.; Durand, F. Bilateral Filtering: Theory and Applications. *Found. Trends Comput. Graph. Vis.* **2008**, *4*, 1–73. [CrossRef]
15. Shyam Anand, C.; Sahambi, J.S. Pixel Dependent Automatic Parameter Selection for Image Denoising with Bilateral Filter. *Int. J. Comput. Appl.* **2012**, *45*, 41–46.
16. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [CrossRef]
17. Johannsen, G.; Bille, J.A. A Threshold Selection Method Using Information Measure. In Proceedings of the 6th International Conference on Pattern Recognition (ICPR'82), Munich, Germany, 19–22 October 1982; pp. 140–143.
18. Kapur, N.; Sahoo, P.K.; Wong, A.K.C. A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram. *Comput. Vis. Graph. Image Process.* **1985**, *29*, 273–285. [CrossRef]
19. Li, C.H.; Tam, P.K.S. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognit. Lett.* **1998**, *19*, 771–776. [CrossRef]

20. Glasbey, C.A. An analysis of histogram-based thresholding algorithms. *Graph. Models Image Process.* **1993**, *55*, 532–537. [CrossRef]
21. Kittler, J.; Illingworth, J. Minimum error thresholding. *Pattern Recognit.* **1986**, *19*, 41–47. [CrossRef]
22. Mixture Modeling. ImageJ. Available online: http://imagej.nih.gov/ij/plugins/mixture-modeling.html (accessed on 20 January 2018).
23. Tsai, W.H. Moment-preserving thresholding: A new approach. *Comput. Vis. Graph. Image Process.* **1985**, *29*, 377–393. [CrossRef]
24. Doyle, W. Operation useful for similarity-invariant pattern recognition. *J. Assoc. Comput. Mach.* **1962**, *9*, 259–267. [CrossRef]
25. Pun, T. Entropic Thresholding, A New Approach. *Comput. Vis. Graph. Image Process.* **1981**, *16*, 210–239. [CrossRef]
26. Shanbhag, A.G.G. Utilization of Information Measure as a Means of Image Thresholding. *Comput. Vis. Graph. Image Process.* **1994**, *56*, 414–419. [CrossRef]
27. Zack, G.W.; Rogers, W.E.; Latt, S.A. Automatic measurement of sister chromatid exchange frequency. *J. Histochem. Cytochem.* **1977**, *25*, 741–753. [CrossRef] [PubMed]
28. Wu, U.L.; Songde, A.; Haqing, L.U.A. An Effective Entropic Thresholding for Ultrasonic Imaging. In Proceedings of the International Conference Pattern Recognition, Brisbane, Australia, 16–20 August 1998; pp. 1522–1524.
29. Yen, J.C.; Chang, F.J.; Chang, S. A New Criterion for Automatic Multilevel Thresholding. *IEEE Trans. Image Process.* **1995**, *4*, 370–378. [PubMed]
30. Ridler, T.W.; Calvard, S. Picture Thresholding Using an Iterative Selection Method. *IEEE Trans. Syst. Man Cybern.* **1978**, *8*, 630–632.
31. Prewitt, M.S.; Mendelsohn, M.L. The Analysis of Cell Images. *Ann. N. Y. Acad. Sci.* **1996**, *128*, 836–846. [CrossRef]
32. Kavallieratou, E.; Stamatatos, S. Adaptive binarization of historical document images. In Proceedings of the 18th International Conference on Pattern ICPR 2006, Hong Kong, China, 20–24 August 2006; Volume 3.
33. Sauvola, J.; Pietikainen, M. Adaptive document image binarization. *Pattern Recognit.* **2000**, *33*, 225–236. [CrossRef]
34. Niblack, W. *An introduction to Digital Image Processing*; Prentice-Hall: Upper Saddle River, NJ, USA, 1986.
35. Ntirogiannis, K.; Gatos, B.; Pratikakis, I. Performance Evaluation Methodology for Historical Document Image Binarization. *IEEE Trans. Image Process.* **2013**, *22*, 595–609. [CrossRef] [PubMed]
36. Lins, R.D.; Silva, G.F.P.; Torreão, G.; Alves, N.F. Efficiently Generating Digital Libraries of Proceedings with the LiveMemory Platform. In *IEEE International Telecommunications Symposium*; IEEE Press: Rio de Janeiro, Brazil, 2010; pp. 119–125.