Article

A New Binarization Algorithm for Historical Documents

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

[1] Federal University of Pernambuco, Recife-PE, Brazil
[2] Federal Rural University of Pernambuco; Recife-PE, Brazil
* Correspondence: mmar@ufpe.br; Tel.: +55-81-2126-7129

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 The binarization of historical documents is far more complex than recent ones as paper aging, color, texture, translucidity, stains, back-to-front interference, kind and color of ink used in handwritting, 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 This article presents a new binarization algorithm for historical documents.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 21 binarization algorithms in three sets of "real world" documents showed very good results.

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

1. Introduction

Binary documents claim for far less storage space and computer bandwidth for network transmission than color or grayscale documents. Document image binarization plays an important role in the document image analysis, compression, transcription, and recognition pipeline. Historical documents drastically increase the degree of difficulty for binarization algorithms. Physical noises [1] such as stains and paper aging affect the performance of binarization algorithms. Besides that, historical documents were often typed or written on both sides of sheets of paper and the opacity of the paper is often such as to allow the back printing or writing to be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* [2], was later known as *bleeding* or *show-though* [3]. Figure 1 presents an example of a document with such a noise. 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. Depending on the strength of the interference present, that depends on the opacity of the paper, its permeability, the kind and degree of fluidity of the ink used, the degree of difficulty for obtaining good segmentation capable of filtering-out such a noise increases enormously, as 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 [4]. OCRs are also unable to work properly for such documents. Several algorithms were developed specifically to binarize documents with back-to-front interference [2] [3][5-8].

There is no binarization technique to be an all case winner as many parameters may interfere in the quality of the resulting image [8]. The development of new binarization algorithms is still an important research topic. International competitions on binarization algorithms, such as DIBCO - Document Image Binarization Competition [9], are an evidence of the relevance of this area. Having

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 November 2017 doi:10.20944/preprints201711.0143.v1

Peer-reviewed version available at J. Imaging 2018, 4, 27; doi:10.3390/jimaging4020027

quantitative criteria to choose which is the best binarization algorithm, in terms of image quality and performance, for a specific image is of paramount importance.

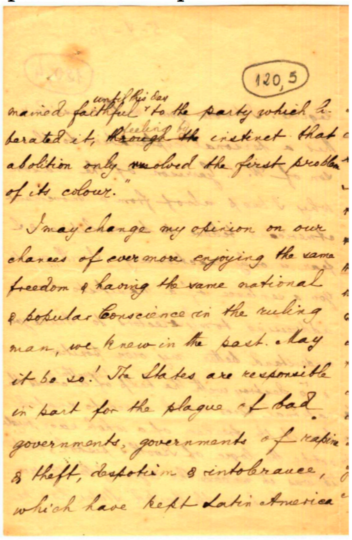
120,5
married faithful to the party which liberated it, through the cristnet that abolition only involved the first problem of its colour.
I may change my opinion on our chances of ever more enjoying the same freedom & having the same national & popular Conscience in the ruling men, we knew in the past. 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 1. Historical document from Nabuco bequest with back-to-front interference.

This paper presents a new global filter 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 (late 19th century to present, both printed and handwritten, using a different kind 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: 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. Figure 2 presents the block diagram of the proposed algorithm [10]. The functionality of each block is detailed as follows.

2.1. The Bilateral Filter

The bilateral filter was first introduced by Aurich and Weule [11] under the name "nonlinear Gaussian filter". It was later rediscovered by Tomasi and Manduchi [12] who called it the "bilateral filter" which is now the most commonly used name according to reference [13].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 November 2017 doi:10.20944/preprints201711.0143.v1

Peer-reviewed version available at *J. Imaging* **2018**, *4*, 27; doi:10.3390/jimaging4020027

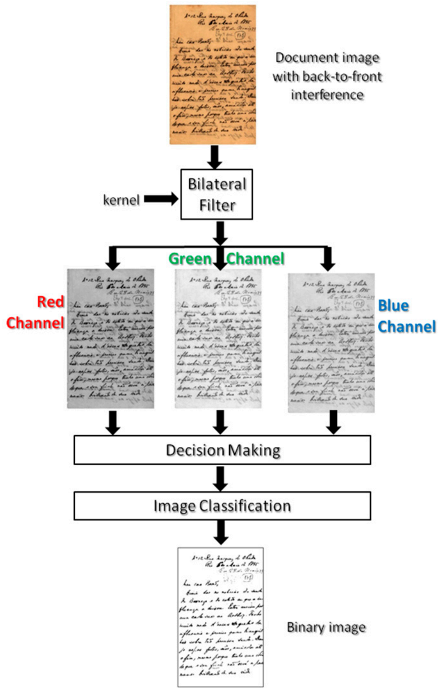

Figure 2. Block diagram of the proposed algorithm.

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 NxN window size. For a pixel (x,y), the output of a bilateral filter can be as described by equation:

$$I_{BF}(x,y) = \frac{1}{K}\sum_{x,y=(\hat{x},\hat{y})-N}^{(\hat{x},\hat{y})+N} e^{-\frac{||x-\hat{x}||^2+||y-\hat{y}||^2}{2\delta_d^2}}\, e^{-\frac{(I(x,y)-(I(\hat{x},\hat{y}))^2}{2\delta_d^2}}, \quad (1)$$

where I(x,y) is the pixel intensity in the image before applying the bilateral filter, IBF(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:

$$K = \sum_{x,y=(\hat{x},\hat{y})-N}^{(\hat{x},\hat{y})+N} e^{-\frac{||x-\hat{x}||^2+||y-\hat{y}||^2}{2\delta_d^2}}\, e^{-\frac{(I(x,y)-(I(\hat{x},\hat{y}))^2}{2\delta_d^2}}. \quad (2)$$

Equations (1) and (2) how that the bilateral filter has three parameters. The parameters δ_d (filter domain) and δ_r (filter range) are $e^{-\frac{||x-\hat{x}||^2+||y-\hat{y}||^2}{2\delta_d^2}}$ and $e^{-\frac{||x-\hat{x}||^2+||y-\hat{y}||^2}{2\delta_d^2}}$, respectively. The third parameter is the window size NxN.

The geometric spread of the bilateral filter is controlled by δ_d. As δ_d is increased, more neighbours are combined in the diffusion process resulting in a more "smooth" image, while δ_r represents the photometric spreading. Only pixels with a percentage difference of less than δ_r are processed [13].

2.2. 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 2. 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

 doi:10.20944/preprints201711.0143.v1

Peer-reviewed version available at J. Imaging 2018, 4, 27; doi:10.3390/jimaging4020027

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:

$$P(b/b) = f(\alpha,R,G,B). \tag{3}$$

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

$$t^* = \text{Max}P(b/b), \tag{4}$$

subject to a given criterion P(f/f) ≥ M. The criterion used here was M=97%, that is at most 3% of the foreground pixels may be incorrectly mapped. 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. The generation of the synthetic images is explained below.

2.3. Generating synthetic images

The Decision-Making Block needs training to "learn" about the optimal threshold parameters. 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 3 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.

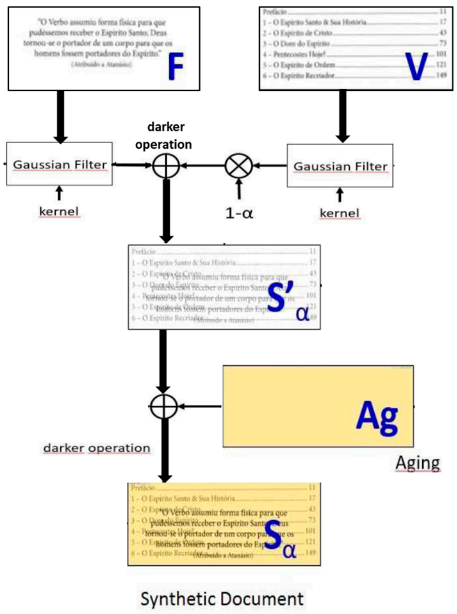

Figure 3. Block diagram of the scheme for the generation of synthetic images for the Decision-Making Block.

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, this parameter is currently being changed for ten. The "blurred" verso image is now faded with a coefficient α varying between 0 and 1 in steps of 0.01. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 November 2017 doi:10.20944/preprints201711.0143.v1

Peer-reviewed version available at J. Imaging 2018, 4, 27; doi:10.3390/jimaging4020027

analysis of 3,450 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 [14] and randomly, as explained in reference [8]. The training performed in the current version of the algorithm presented was made with 16 of those 200 synthetic textures. The total number of images used for training here was thus 16 (textures), times 100 (0< α <1 in steps of 0.01), times 2 blur parameters for the Gaussian filters, times 10 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 [8].

2.4. 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" naïve Bayes automatic classifier which was trained using 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 [8]. Twenty-one binarization algorithms were tested using the methodology described:

1. DaSilva-Lins-Rocha [5]
2. Intermodes [15]
3. Ergina-Local [33]
4. IsoData [16]
5. Johannsen-Bille [17]
6. Kapur-Sahoo-Wong [18]
7. Li-Tam [19]
8. Mean [20]
9. Mello-Lins [4]
10. MinError [21]
11. Minimum (variation of [15])
12. Mixture-Modeling [22]
13. Moments [23]
14. Otsu [24]
15. Percentile [25]
16. Pun [26]
17. RenyEntropy (variation of [18])
18. Shanbhag [27]
19. Triangle [28]
20. Wu-Lu [29]
21. Yean-Chang-Chang [30]

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 [9] had ground-truth images available. This makes the assessment task of real-world images extremely difficult [32]. 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 21 algorithms above and the bilateral algorithm to all the test images in the Nabuco and LiveMemory 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 November 2017 doi:10.20944/preprints201711.0143.v1

Peer-reviewed version available at J. Imaging 2018, 4, 27; doi:10.3390/jimaging4020027

as ground-truth image for the assessment below. All the processing time figures presented in this paper are from Intel i7-4510U@ 2.00GHzx2, 8GB 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 6,500 letters and postcards written and typed by Joaquim Nabuco [6], totaling about 30,000 pages. The images 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 universe.

Table 1 presents the quantitative results obtained for all the documents in this dataset. P(f/f) stands for the number of foreground pixels in the ground truth image mapped onto black pixels in the binarized image. P(b/b) is the number of background pixels in the ground-truth image mapped onto white pixels of the binary image. The SDP(f/f) and SDP(b/b) the standard deviation of P(f/f) and P(b/b). 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(f/f) decreasing order.

Table 1. Binarization results for images from Nabuco bequest.

AlgName	P(f/f)	P(b/b)	SD P(f/f)	SD P(b/b)	Time (s)
IsoData	98.08	99.38	3.39	0.60	0.0171
Otsu	98.08	99.36	3.39	0.63	0.0159
Bilateral	99.57	99.29	1.23	0.93	1.0790
Huang	99.40	98.69	2.14	0.88	0.0200
Moments	99.39	98.40	1.34	1.70	0.0160
Ergina-Local	99.99	98.13	0.03	0.64	0.3412
RenyEntropy	100.00	97.56	0.00	1.17	0.0188
Kapoo-Sahoo-Wong	100.00	97.51	0.00	1.07	0.0172
Yean-Chang-Chang	100.00	97.38	0.00	1.26	0.0161
Triangle	100.00	95.94	0.00	1.46	0.0160
Mello-Lins	98.61	89.63	5.14	24.43	0.0160
Mean	100.00	81.77	0.00	5.99	0.0168
Johannsen-Bille	98.87	59.77	2.97	48.80	0.0164
Pun	100.00	55.44	0.00	2.57	0.0185
Percentile	100.00	53.21	0.00	1.33	0.0185

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 it is a more stable documents among the various images in this dataset. Figure 4 presents the document for which the bilateral filter presented the worst results in terms of image quality with two zoomed areas from the original and the binarized document.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 November 2017 doi:10.20944/preprints201711.0143.v1

Peer-reviewed version available at J. Imaging 2018, 4, 27; doi:10.3390/jimaging4020027

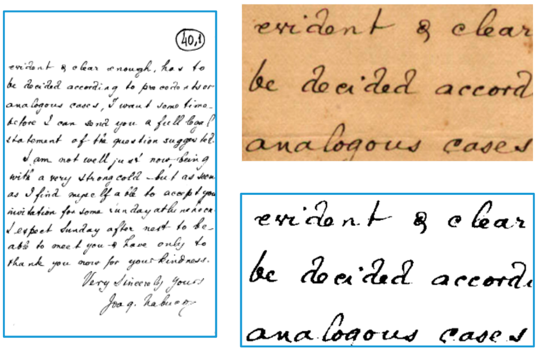

Figure 4 – Historical document from Nabuco bequest with the worst binarization results for the bilateral filter with zoom from original and binary parts

3.2. The LiveMemory dataset

This dataset encompasses 15 documents with 200 dpi resolution selected from the over 8,000 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 from typed or electronically produced. Table 2 presents the performance results for the 10 best ranked algorithms. The bilateral filter obtained the best results in terms of image filtering. It is in worth observing that the image quality degraded for all the algorithms. The shaded area due to the hard bound spine of the volumes of the proceedings, as one can see in Figure 5, were possibly the responsible for such lower quality results.

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

AlgName	P(f/f)	P(b/b)	SD P(f/f)	SD P(b/b)	Time (s)
Bilateral	100.00	98.97	0.00	1.07	3.1220
IsoData	93.98	98.22	20.78	2.84	0.0600
Otsu	94.02	98.18	20.79	2.90	0.0594
Moments	94.46	97.52	20.69	2.76	0.0579
Ergina-local	93.46	97.23	20.56	2.09	0.9619
Huang	94.78	96.03	19.25	4.95	0.0728
Triangle	94.85	93.85	19.26	3.13	0.0597
Mean	95.66	83.26	16.25	5.85	0.0612
Oun	97.91	55.15	7.80	3.67	0.0662
Percentile	97.91	53.78	7.80	1.99	0.0640

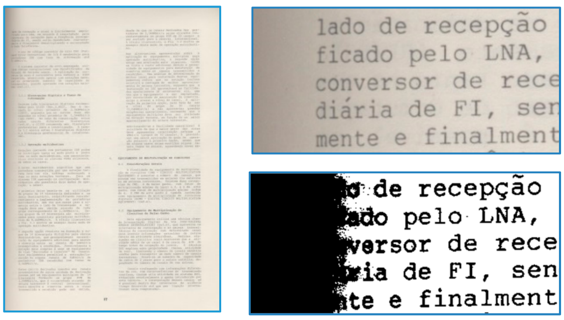

Figure 5 –LiveMemory with the worst binarization results for the bilateral filter with original and binary zoom.

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 poor. The reason for that is possibly that all the training images for the bilateral filter were 300 dpi synthetic images, while the DIBCO images are very small sized high-resolution images. Figure 6 presents the DIBCO image for which the bilateral filter presented the worst binarization results.

Table 3. Binarization results for images from DIBCO.

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

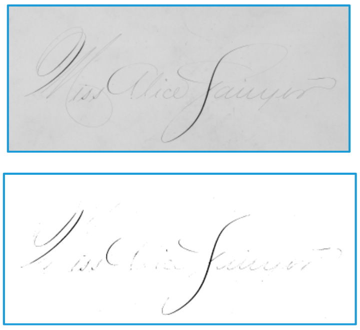

Figure 6 – Document from DIBCO with the worst binarization results for the bilateral filter

4. Conclusions

Historical documents are far more difficult to binarize as 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 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 one other binarization algorithms 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 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 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 made that match, the most suitable binarization algorithms are immediately known. If this paper were accepted, all the test images and algorithms will be included in the DIBplatform. The preliminary version of the DIB-Document Image Binarization platform and website is publically available at www.cin.ufpe.br/~dib.

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 the DIBCO team from making their images publically available. The authors also acknowledge the partial financial support of to CNPq and CAPES - Brazilian Government.

References

1. Lins, R.D. A Taxonomy for Noise in Images of Paper Documents - The Physical Noises. *ICIAR* **2009**, *Volume 5627.pp. 844-854.*
2. Lins, R.D. at al. An Environment for Processing Images of Historical Documents. *Microproc. and Microprogramming* **1995**. *pp. 111-121.*
3. Sharma, G. Show-trough cancellation in scans of duplex printed documents. *IEEE Transaction Image Processing* **2001**. *Volume 10. N. 5, pp. 736-754.*
4. Mello, C. A. B. and Lins, R. D. Generation of Images of Historical Documents by Composition. Symp. On Document Engineering **2002**. *pp. 127-133.*
5. Silva, M. M., Lins, R. D., Rocha, V. C. Binarizing and Filtering Historical Documents with Back-to-Front Interference. *ACM Symposium on Applied Computing* **2006**, *pp. 853-858.*
6. Lins, R. D. Nabuco – Two Decades of Processing Historical Documents in Latin America. *Journal of Universal Computer Science* **2011** *Volume 17. N. 1, pp. 151-161.*
7. Roe, E. and Mello, C. A. B. Binarization of Color Historical Document Images Using Local Image Equalization and XDoG. *12th International Conference on Document Analysis and Recognition* **2013**. *pp. 205-209.*
8. Lins, R.D., Almeida, M. A. M., Bernardino, R. B., Jesus, D., Oliveira, J. M. Assessing Binarization Techniques for Document Images. *In Proceedings of ACM Symposium on Document Engineering, Valetta, Malta* **2017**.
9. DIBCO.
10. Almeida, M. A. M. Statistical Analysis Applied to Data Classification and Image Filtering. Doctorate, Federal University of Pernambuco, Recife-PE, Brazil, 21 December 2016.
11. Aurich, V. and Weule, J. B. Non-linear gaussian filters performing edge preserving diffusion. *In Proceedings of the DAGM Symposium* **1995**. *pp. 538 - 545.*
12. Tomasi, C. and Manduchi, R. Bilateral filtering for gray and color images. *In IEEE Proc. 6th International Conference on Computer Vision* **1998**. *pp. 836-846.*
13. Paris, P., Kornprobst, P., Tumblim, J., Durand, F. Bilateral Filtering: Theory and Applic. Found. *Trends in Comp. Graphics and Vision* **2008**. *Volume 4. N. 1, pp. 1-73.*
14. Efros, A. A. and Freeman, W. T. Image quilting for texture synthesis and transfer. *SIGGRAPH '01 28th Annual Conference on Computer Graphics and Interactive Techniques* **2001**. *pp. 341-346.*
15. Prewitt, M. S. and Mendelsohn, M. L. The Analysis of Cell Images. *Ann. N. Y. Acad. Sci.* **1996**. *Volume 128, N. 3, pp. 836-846.*
16. Ridler, T. W. and Calvard, S. Picture Thresholding Using an Iterative Selection Method. *IEEE Trans. Systems, Man., and Cybernetics* **1978**. *Volume 8, N. 8, pp. 630-632.*
17. Johannsen, G. and Bille, J. A. A Threshold Selection Method Using Information Measure. *ICPR'82 – 6th International Conference on Pattern Recognition* **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. *C. Vision Graphics and Image Processing* **1985**. *Volume 29, pp. 273-285.*
19. Li, C. H. and Tam, P. K. S. An iterative algorithm for minimum cross entropy thresholding. *Pattern Recognition Letters* **1998**. *Volume 19, N. 8, pp. 771-776.*
20. Glasbey, C. A. An analysis of histogram-based thresholding algorithms. *CVGIP: Graphical Models and Image Processing* **1993**. *Volume 55, pp. 532-537.*
21. Kittler, J. and Illingworth, J. Minimum error thresholding. *Pattern Recognition* **1986**. *Volume 19, N. 1, pp. 41-47.*

22. Title of Site. Available online: https://imagej.nih.gov/ij/plugins/mixture-modeling.html (accessed on 24 October 2017).
23. Tsai, W. H. Moment-preserving thresholding: A new approach. *Computer Vision, Graphics, and Image Processing* **1985**. *Volume 29, N. 3, pp. 377-393*.
24. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transaction on Systems, Man and Cybernetics* **1979**. *Volume SMC-9, N. 1, pp. 62-66*.
25. Doyle, W. Operation useful for similarity-invariant pattern recognition. *Journal of the Association for Computing Machinery* **1962**. *Volume 9, pp. 259-267*.
26. Pun, T. Entropic Thresholding, A New Approach. *Computer Vision Graphics, and Image Processing* **1981**. *pp. 210-239*.
27. Shanbhag, A. G. G. Utilization of Information Measure as a Means of Image Thresholding. *Computer Vision Graphics, and Image Processing* **1994**. *Volume 56, N. 5, pp. 414-419*.
28. Zack, G. W., Rogers, W. E., Latt, S. A. Automatic measurement of sister chromatid exchange frequency. *Journal Histochem Cytochem* **1977**. *Volume 25, N. 7, pp. 741-753*.
29. Wu, U. L., Songde, A., Haqing, L. U. A. An Effective Entropic Thresholding for Ultrasonic Imaging. *International Conference Pattern Recognition* **1998**. *pp. 1522-1524*.
30. Yen, J. C., Chang, F. J., Chang, S. A New Criterion for Automatic Multilevel Thresholding. *IEEE transaction Image Process IP-4* **1995**. *pp. 370-378*.
31. 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* **2010**. *pp. 119-125*.
32. Ntirogiannis, K., Gatos, B., Pratikakis, I. Performance Evaluation Methodology for Historical Document Image Binarization. *IEEE Transaction Image Process* **2013**. *Volume 22, N. 2, pp. 595-609*.
33. Kavallieratou, Ergina, and Stamatatos Stathis, Adaptive binarization of historical document images. *Pattern* **2006**, *ICPR 2006. 18th International Conference on. Volume 3, IEEE*, **2006**.