A Quality and Time Assessment of Binarization Algorithms

Rodrigo Bernardino
UFPE
Recife, Brazil
rbb.ufpe@gmail.com

Rafael Lins
UFRPE – UFPE
Recife, Brazil
rdl.ufpe@gmail.com

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

***Abstract*—Binarization algorithms are an important step in most document analysis and recognition applications. Many aspects of the document affect the performance of binarization algorithms, such as paper texture and color, noises such as the back-to-front interference, stains, and even the type and color of the ink. This work focuses on determining how each document characteristic impacts the time to process and the quality of the binarized image. This paper assesses thirty of the most widely used document binarization algorithms.**

Keywords - Binarization; documents; back-to-front-interference; show-through; algorithms;

I. INTRODUCTION

Binarization is the name of the process by which a color image is converted into its monochromatic version. Such a process is often applied on document images, as its black and white version is much easier for computers to process, require less storage space and bandwidth when transmitting through computer networks. Due to its importance, there is an ever-growing variety of binarization methods, which produce images with good quality not only for visual inspection but also for numberless applications within the context of document analysis. Thus, it is essential to have a proper quality and processing time evaluation methodology.

Several document binarization algorithms, assessments, and competitions have been published in the last two decades. Sezgin and B. Sankur [1] conducted a comprehensive study with a vast number of approaches and techniques, presenting also an overall image quality assessment of the algorithms developed up to 2004. The international competitions also witness the importance of this area. Maybe the most traditional and best-regarded competition on documents binarization is DIBCO – Document Image Binarization Competition, which was first organized at the ICDAR – International Conference on Document Analysis and Recognition, in 2009, and have been occurring every year ever since [2].

The DIBCO methodology consists of applying the algorithms to about 10 excerpts of high dpi real images and comparing them with a ground-truth (GT) image. The GT is the binary equivalent manually generated or retouched "by hand", which produces a good-quality monochromatic image under visual inspection, taken as reference. DIBCO compares only the images produced using the competitors' binarization algorithms with the GT images using several measures: F-Measure, pseudo F-Measure, PSNR and DRD [5]. However, the specific characteristics of the images are rarely considered in nearly all assessments seen in the literature. Different algorithms have different strengths and weaknesses and no binarization algorithm is capable of performing well for all kinds of images.

Another aspect that has been missing in binarization studies is assessing the time taken for each algorithm to process the images. If, for instance, an algorithm is capable of outperforming in image quality any other for a given dataset, but takes far longer to process a small portion of document (as the DIBCO test images), it might not be suitable to process a whole document, or even a large batch of documents in large scale document processing plant. Thus, the context in which the algorithm will be used should also be considered in the analysis.

However, in order to find precisely in which context each algorithm performs best, one needs to specify, for example, which kind of noise and its strength that is present in the image. This work attempts to analyze which factors affect most the quality and time performance of binarization algorithms taking into account the effects of the back-to-front interference, texture, type of print and ink on the thirty most widely used binarization techniques. A test set of 2,257 scanned images, encompassing controlled parameter synthetic images from the publicly available IAPR TC10-TC11 DIB dataset (https://dib.cin.ufpe.br), was used here. The documents focus of this work are text only scanned documents. The binarization of camera-acquired images is far more complex due to the uneven resolution and illumination, among other problems, being thus out of the scope of this paper. Similarly, illustrated documents with graphical elements such as color diagrams or photos are not addressed here.

II. ELEMENTS OF A TEXT DOCUMENT IMAGE

It is fundamental to be explicit about the kind of document one wants to binarize. Figure 1 shows a real-world historical document used to compose some images from the test set in which one may see: (i) The paper texture. (ii) The ink from the foreground. (iii) The ink from the background showing-trough the foreground (the back-to-front interference). (iv) Other physical noises, such as stains, folding marks, etc.

In the case of most historical documents, the ink in the foreground presents minimum variation in the intensity, while the ink in the backside is seen blurred in the front-side. The document image in Figure 1 also shows that the paper background presents variations in the texture, besides some stains possibly due to fungi. The analysis of several documents shows that the back-to-front interference tends to show a uniform offset in relation to the writing in the foreground, such an offset is called a *shift*, and it corresponds to the mean number of pixels of the offset.

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00232

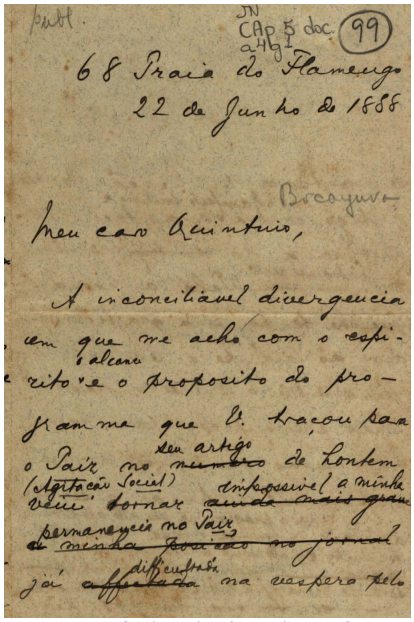

Figure 1 – Part of a handwritten letter from Nabuco bequest in the test set, showing a non-uniform texture, back-to-front interference, folding marks, etc.

III. A New Quality Assessment Method

The first challenge met in the task proposed here is to find a valid quality assessment method. DIBCO takes the F-Measure, pseudo F-Measure, PSNR, and DRD between the GT image and the result of each competitor's algorithm to generate the final scores of the algorithms. DIBCO makes no processing time assessment. The final rank calculation is done by summing up the rank positions in each of the listed measures.

This paper proposes a new evaluation method for binarization algorithms using (1):

$$Quality\_Score = 100 \times (k + (P_{if} \times 1.5)) / (2.5), \quad (1)$$

where k is the Cohen's Kappa coefficient [8] and P_{if} is the proportion of pixels that have been wrongly mapped from the back-to-front interference as being part of the text, i.e. as foreground pixels. The Kappa can be interpreted as a weighted summarization of the error (or confusion) matrix of the number of foreground and background correctly mapped pixels, taking the GT image as reference. Kappa compares the observed accuracy with an expected accuracy, indicating how well a given classifier performs. As indicated in [37], the Kappa may be preferred in some cases when evaluating binary classifiers. In (2), the equation for Kappa is presented:

$$k = (P_o - P_c) / (1 - P_c), \quad (2)$$

where P_o is the number of correctly mapped pixels (accuracy) and P_c is calculated by using (3):

$$k = (n_{bf} \times n_{gf} + n_{bb} \times n_{gb}) / N^2, \quad (3)$$

where n_{bf} and n_{bb} are the number of pixels mapped as foreground and background on the binary image, respectively, while n_{gf} and n_{gb} are the number of foreground and background pixels on the GT image and N is the total number of pixels. Furthermore, the value of P_{if} is multiplied by 1.5, a weight value empirically found to generate the appropriate comparisons between different binary images. Finally, to normalize the *Quality_Score* value to (0, 100), the equation is multiplied by 100 and divided by 2.5.

The use of the *Quality_Score* proposed here, in (1), showed results consistent with what is expected by visual inspection of the resulting binarized images and a global quality rank not much different from the one obtained using the DIBCO ranking system, although far simpler and more straightforward. However, it is important to note that one must know beforehand the precise positions of the back-to-front interference pixels that count for P_{if}.

Furthermore, in addition to quality, the time required to binarize the image is also considered in the analysis here. First, the algorithms are sorted according to the quality of the images produced. Then, their time performance is assessed such that, for example, if the 2 best-performing algorithms in its category of images have similar quality results, but one has an overall better time performance, the faster one will be recommended.

IV. Performance Variation with Image Elements

The assessment of the thirty binarization algorithms studied here, with the chosen images from the DIB data set, showed that the quality and time performance presented high sensitivity to the intensity and the degree of blur of the back-to-front noise and its mean offset in relation to the lines in the foreground. One of the advantages of using the DIB platform (https://dib.cin.ufpe.br) is the possibility of any user to generate test images controlling such parameters in different categories. The intensity of the back-to-front noise is modeled by the parameter α, whose values used here are 1.0 (no interference), 0.8 (weak interference), 0.7 (medium interference), and 0.6 (strong interference). The degree of blur is modeled by using two Gaussian blur filters with kernel size 3 and 5. The mean pixel offset (*shift*) used three values: 10, 20, and 30 pixels. Three equal size subsets of test images were used here: Modern (Inkjet printed and ballpen handwritten documents), Nabuco (quill pen handwritten and typewritten historical documents from Nabuco bequest [7]), and SBT (inkjet and offset printed documents from the Live Memory project [6]). Eight very different texture seeds from the one hundred existing in the DIB dataset were chosen to generate random and quilt type textures for this sensitivity analysis. In total, 2,736 parameter controlled synthetic images have been initially considered for the analysis, however, after visually inspecting each combination of text and texture, some of them were considered invalid (e.g. modern printer ink with historical paper texture), and were discarded, leading to a final number of 2,257 images. Each image was then processed by each of the thirty algorithms and several performance measures were taken, in addition to collecting the processing time.

Eleven of the thirty binarization algorithms were discarded because of the low-quality images produced (average *Quality_Score* ≤ 50): Bernsen [9], Huang [10], Johannsen-Bille [11], Mean [12], MinError [13], Niblack [14], Percentile [15], Pun [16], Rosin [17], Shanbhag [18] and Triangle [19]. Howe algorithm [4] was also discarded from this analysis for being several orders of magnitude slower than the others with no quality gains if compared with the other best-ranked algorithms.

A. Parameter-specific Assessment of the Algorithms

The first part of this study focused on understanding how each of the listed parameters impacted the quality and time of the resulting binary image. Linear regression was performed over the parameter variation. Tables I-II present the overall results, in which "**L**ow/**M**edium/**H**igh" indicates the level of significance of each parameter for each algorithm. This significance indication was inferred from the t-value output of the regression, then, in other words, a high significance means that it is very likely or, on average, if the reference parameter varies, the outcome (*Quality_Score* or time) will also vary. The values after each parameter significance indication represent the magnitude of the average increase or decrease of the outcome when the parameter varies.

TABLE I. SCORE SENSITIVITY ON PARAMETER VARIATION

Algorithm	Quality_Score		
	Shift	***Blur***	***α***
Bradley [20]	H/α0.6-7/-2.5	H/α0.6-7/+5	H/α0.6-7/-13.5
dSLR [21]	-	L/α0.6/+1.4	H/all/-4.8
Ergina-G. [22]	-	-	H/all/-16.7
Ergina-L. [23]	-	L/α0.7/+3	H/all/-14.9
Intermodes [24]	-	H/α0.6/+3	H/α0.6/-4.7
IsoData [25]	-	H/α0.6/+16	H/α0.6/-17.8
Kapur-SW [26]	-	-	H/0.6-7/-8.0
Li-Tam [27]	-	H/α0.6/+4	H/α0.6/-4.1
Mello-Lins [28]	-	-	H/all/-7.1
Minimum [24]	-	L/α0.6/+1.4	H/0.6-7/-3.4
Moments [29]	-	H/α0.6-7/+6.9	H/0.6-7/-10.1
Nick [30]	H/α0.6/-2.0	H/α0.6-7/+7.1	H/α0.6-7/-7.3
Otsu [31]	M/α0.6/-2.0	H/α0.6/+22	H/α0.6/-25.3
RenyE. [32]	H/α0.6/+11.5	-	H/α0.6/-39.2
Sauvola [33]	-	-	L/0.6/-1.1
Wolf [34]	-	-	H/all/-0.2
Wu-Lu [35]	-	H/α0.6/-0.76	H/all/-0.7
Yen-CC [36]	-	-	L/0.6-7/-42,3

Either "all" or "α0.6-7" indicates if the parameter variation impacts the "*Quality_Score*" or "Time" for any value of other parameters or only when, for example, α=0.6 or 0.7. For example, the parameter α is highly significant to Bradley algorithm, what it means that, on average, when α goes from the reference value (in this case, 1.0 – no interference) to α=0.6 or a=0.7, the *Quality_Score* decreases about –13. On the other hand, when it goes from 1.0 to 0.8, the *Quality_Score* does not vary. Now, for algorithm Yen-CC, the significance is low, as the *Quality_Score* is constant for most images, but when it varies it is as high as –42.3.

The graphs presented in Fig. 2-3 are samples of a visual representation of the results that were used in the analysis to have an overall picture of the impact on each algorithm. The space between the colored lines indicates the magnitude of variance. The black horizontal line highlights the *Quality_Score* = 96, which is a value that indicates a good quality of binarized image, as observed through visual inspection. Thus, the number of points close to that value may give a rough idea of how good an algorithm could binarize the test images.

TABLE II. TIME SENSITIVITY ON PARAMETER VARIATION

Algorithm	Time (ms)		
	Shift	***Blur***	***α***
Bradley [20]	H/+14.0	-	M/α0.6/+34.8
dSLR [21]	M/+2.4	-	H/ α0.6/+14.5
Ergina-G. [22]	-	-	H/all/-199.4
Ergina-L. [23]	-	-	-
Intermodes [24]	-	-	H/α0.6L7/+6.4
IsoData [25]	-	L/+4.5	L/all/-6.5
Kapur-SW [26]	L/+2.2	-	M/α0.6/+5.3
Li-Tam [27]	H/-3.2	L/+3.3	L/α0.7/+7.7
Mello-Lins [28]	L/-2.2	-	M/α0.6/+8.8
Minimum [24]	-	-	H/α0.6/+4.4
Moments [29]	-	-	M/0.7-8/-7.0
Nick [30]	-	-	M/α0.6/+35.1
Otsu [31]	L/-1.3	-	H/α0.7/-8.3
RenyE. [32]	H/-6.0	-	M/α0.6/+10.5
Sauvola [33]	-	L/+87.4	M/α0.7/+128.4
Wolf [34]	-	-	-
Wu-Lu [35]	M/+2.9	-	-
Yen-CC [36]	M/+3.6	-	M/α0.7-8/-75.5

1) Shift Variation

For some algorithms, the variation of the shift parameter implied in significant changes in the *Quality_Score*. Here, the shift equal to ten pixels was taken as reference. Bradley is the one which varies the most and its variation is illustrated on Fig. 2. As depicted on Table I, on average, when *s* varies from s=10 to 20 or s=20 to 30 pixels, the *Quality_Score* varies –2.60, either for α=6 or 7. Shift also has significance for Nick, Otsu, Yen and RenyEntropy, but only for α=0.6. As for the rest of the algorithms, shift has little to no impact on the *Quality_Score*. An example of an algorithm that doesn't suffer from shift variation is Intermodes and on Fig. 3 this situation is presented.

Regarding the impact on processing time, it has been noticed that for three algorithms: Bradley, Li-Tam and Reny, a variation in Shift had a significant impact on the time required time to process the image. For dSLR, Wu-Lu and Yen-CC algorithms, the variation had medium impact in the *Quality_Score* and for Kapur-SW, Mello-Lins and Otsu, it had low impact. This highlights the importance of better understanding how specific characteristics of the document image may impact the binarization process.

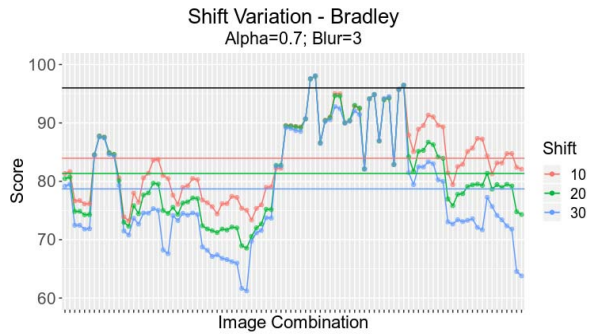

Figure 2 - Impact of Shift variation on Score for Bradley algorithm.

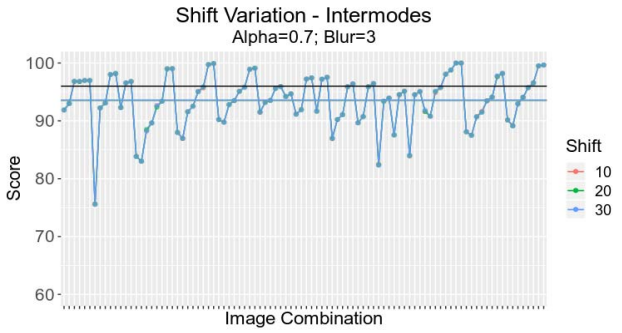

Figure 3 - Shift variation on Score for Intermodes algorithm.

2) Blur Variation

For several algorithms, the blur variation had a significant impact on *Quality Score*. The blur value equal to three was taken as reference and the Shift parameter has been arbitrarily set to 20. Only algorithms Ergina-Global, Kapur-Sahoo-Wong, Mello-Lins, RenyEntropy, Sauvola, Wolf and Yen-Chang-Chang suffered zero impact on the *Quality_Score*. For DaSilvaLinsRocha, Ergina-Local and Minimum, there was little impact in the *Quality_Score*. Now, for Bradley, Intermodes, IsoData, Li-Tam, Moments, Nick, Otsu, and Wu-Lu, the impact in the *Quality_Score* was high, but only when α=0.6. From those results, it is possible to infer that blur has little impact on the quality of the binary images of the studied algorithms. The variation of the *blur* brought no processing time variation for the algorithms assessed, except for IsoData, Li-Tam and Sauvola, which do not vary for most images (low significance *Quality_Score*), and when it does, it is around 4.3 and 87 milliseconds, respectively.

3) Alpha Variation

The parameter α, the strength of the back-to-front interference, is the one which most affects the performance of the algorithms, whenever compared to an image with no interference (α=1). The only algorithms that presented little or no variation in the *Quality_Score* due to α were those by Sauvola, Wolf, Wu-Lu and Yen, being only affected for α<=0.6 (strong interference). Bradley, Ergina Global and Local, IsoData, Moments, and specially Otsu, Reny and Yen algorithms were the most affected by the increase in strength of the back-to-front interference, with impact of over 10 and, for the last three mentioned, over 20 in the *Quality_Score*, what leads to totally unreadable document images with strong interference.

Figures 4 to 11 show the mentioned variation for the algorithms ranked as top-five, as discussed in the next session. It is important to note that the results of Table I are averaged, thus, for example, as seen on Fig. 4, Bradley algorithm has significant variation for α=0.6-7 for DIB and Nabuco dataset, however the variation is not just less intense (about half), but also only significant for α=0.6 if the document is from SBT dataset. Another point is the remarkable stability of Intermodes, IsoData and Otsu, which variation for α other than 0.6 is zero.

Regarding the impact on the processing time, most of the algorithms studied have their processing time affected by variation in α, as may be observed in Table II. Only Ergina-Local, Wolf and Wu-Lu algorithms suffered no impact in their *Quality_Score*. dSLR, Ergina Global, Intermodes, Minimum and Otsu varied in processing time for nearly all images. It is also noticeable how the impact on time only occurred for strong interference values (0.6 and 0.7), except for Moments and Yen, which even for α=0.8 (light interference), suffered variation in time. Another important point to note is that for Ergina-Local, IsoData, Moments, Otsu and Yen algorithms the increase in the strength of the back-to-front interference reduced the time required to process, while for all other algorithms, it increased.

B. Overall Quality Assessment of Binarization Algorithms

The evaluation of the several image characteristics impact on quality and time highlighted several aspects of each binarization algorithm. However, the estimation of the specific characteristics of a given real world document image is not trivial and, thus, when it is necessary to evaluate binarization algorithms, the only straightforward characteristics that can be determined are writing type, ink, date of writing, etc. The second assessment conduct in this work is the overall time-quality performance in the context of each image major characteristics, i. e. the age, type of writing and ink type.

Table III-V presents the overall assessment of each group of images. Note that the ranking is done first for every single image and then the ranking number of all images within each group is summed up. The Kappa, P_{if} and Time columns are the *average* values for all images in each group. Hence, the ranking by average *Quality_Score* may differ from the final ranking, the latter being more precise and correct.

1) Discussion

It is remarkable that the classical algorithm by Otsu is still in the top-5 for all categories of images in the test set, performing better than the several more recent algorithms. IsoData, Minimum and Kapur stand out as the three best for most categories.

Another point to remark is that Minimum and Intermodes have the lowest P_{if} values for all categories, being, in some cases, almost zero. That means those algorithms can effectively remove the back-to-front interference for most images. However, that comes with a cost: the foreground is

also strongly affected and those algorithms do not show up in the top-5 for Nabuco handwritten and modern SBT documents.

For modern handwritten documents, which have been written with ballpen, the Minimum algorithm is ranked as the top quality performing, however, its average Kappa is slightly smaller than the second and third algorithms. As for Nabuco typewritten documents, the situation is inverted: even IsoData having a much larger average P_{if}, it still outperformed Minimum in relation to Kappa, suggesting that Minimum has taken too many foreground pixels out. Comparing two binary images with similar and good values for all other measures, but P_{if} over 3%, it can be shown that values higher than 3% imply in images visually bad, with many interference pixels converted into foreground (black).

As a final remark on the overall quality assessment, one may find at Fig. 4-11 the plots for α variation impact on the score, where each line indicates the score of each image on all possible α values. As the shift and blur parameter impacts much less than α (as seen on previous sections), they have been set to shift=10 and blur=3 in order to have a cleaner picture of how each algorithm performed. The only images with score greater than 50 are shown and also, all algorithms that appear on top-5 tables are present, except for Yen CC algorithm, which only appeared once at the 5th position of SBT typewritten dataset.

C. Time Assessment

As discussed in the previous sections, the time required for an algorithm to fully process the document image is of utter importance when deciding which binarization algorithm to choose for a specific application. Thus, this paper also provides a global analysis of the performance in terms of time. For each document image, the algorithms have been sorted according to its processing time (the faster, the better) and the ranking was summed up across all images.

1) Discussion

Table VI presents the final ranking for each image group. The first group (DIB) had the same ranking and very similar average processing time for both handwritten and printed documents. Once DIB dataset images have standardized dimensions for both printing types, that outcome was expected. One may also notice the existence of two main groups: the one on the top, with faster (all global) algorithms. And the other one on the bottom, composed by the local algorithms, on average two orders of magnitude slower than the global and hybrid algorithms. That is also quite understandable, once the local and hybrid algorithms repeatedly apply the same strategy on small portions of the images.

In Figures 4-11, besides the Score variation on α, it is also possible to have a rough idea of overall quality and confirm the results of Table III. The closer the colored lines are to each other and also to the black horizontal line (Score=96), the better the algorithm performed. One expected that the more time costly (local and hybrid) approaches would yield better results in quality; however, for the studied datasets and algorithms, this did not happen.

TABLE III. QUALITY RANKING FOR THE DIB DATASET

Print	#	Algorithm	Quality Score	Kappa (10^{-2})	P_{if}	Time ($s10^{-2}$)
HW ballpen	1	Minimum	94.98	87.53	0.05	3.91
	2	IsoData	92.55	88.23	4.58	4.71
	3	Intermodes	93.43	88.02	2.96	3.83
	4	Otsu	91.95	87.99	5.41	4.96
	5	Li-Tam	94.17	87.32	1.27	4.73
PR inkjet	1	Minimum	94.54	86.37	0.01	4.52
	2	IsoData	93.58	87.28	2.23	4.64
	3	Intermodes	93.77	86.76	1.56	4.68
	4	Otsu	91.36	86.35	5.29	4.61
	5	Li-Tam	93.46	84.34	0.46	4.50

TABLE IV. QUALITY RANKING FOR THE NABUCO DATASET

Print	#	Algorithm	Quality Score	Kappa (10^{-2})	P_{if}	Time ($s10^{-2}$)
HW quillpen	1	IsoData	90.30	84.90	6.11	0.66
	2	Kapur-SW	92.88	87.56	3.58	0.74
	3	Reny E	84.33	83.98	15.44	0.80
	4	Otsu	89.40	84.57	7.38	0.66
	5	Intermodes	91.63	82.70	2.41	0.66
TW historic	1	IsoData	91.53	87.22	5.59	1.26
	2	Minimum	93.95	84.90	0.02	1.26
	3	Intermodes	94.37	86.34	0.28	1.26
	4	Otsu	90.19	86.69	7.47	1.24
	5	Li-Tam	94.07	85.54	0.25	1.25

TABLE V. QUALITY RANKING FOR THE SBT DATASET

Print	#	Algorithm	Quality Score	Kappa (10^{-2})	P_{if}	Time ($s10^{-2}$)
PR inkjet	1	IsoData	92.65	85.58	2.63	1.79
	2	Otsu	89.60	84.34	6.89	1.74
	3	Li-Tam	92.85	84.13	1.34	1.77
	4	Minimum	92.67	81.70	0.02	1.79
	5	Bradley	91.35	84.51	4.10	25.73
TW electr.	1	Reny E	85.83	82.97	12.27	4.87
	2	Kapur SW	92.04	84.98	3.26	4.79
	3	IsoData	88.96	81.45	6.03	4.76
	4	Otsu	88.52	81.50	6.79	4.74
	5	Yen CC	85.38	82.69	12.83	4.77

TABLE VI. TIME RANKING FOR ALL ALGORITHMS

#	Algorithm	DIB	Nabuco HW	Nabuco TW	SBT PR	SBT TW
Average Time in Seconds×10^{-2}						
1	Li-Tam	4.61	0.65	1.25	1.77	4.76
2	Mello-Lins	4.68	0.65	1.24	1.77	4.75
3	Moments	4.63	0.65	1.24	1.79	4.76
4	Yen-CC	4.72	0.66	1.24	1.78	4.77
5	Otsu	4.78	0.66	1.24	1.74	4.74
6	IsoData	4.67	0.66	1.26	1.79	4.76
7	dSLR	4.69	0.66	1.26	1.81	4.76
8	Intermodes	4.26	0.66	1.26	1.81	4.81
9	Minimum	4.21	0.66	1.26	1.79	4.69
10	Kapur SW	4.77	0.74	1.33	1.84	4.79
11	Wu-Lu	4.84	0.76	1.32	1.87	4.90
12	Reny-E	5.06	0.80	1.36	1.84	4.87
Average Time in Seconds						
13	Bradley	0.685	0.706	0.098	0.179	0.257
14	Ergina-G	1.002	0.979	0.117	0.214	0.332
15	Sauvola	1.022	1.068	0.093	0.233	0.405
16	Ergina-L	1.336	1.448	0.163	0.298	0.469
17	Nick	1.474	1.500	0.215	0.392	0.557
18	Wolf	1.528	1.604	0.221	0.402	0.575

V. CONCLUSIONS

This paper proposes a new quality assessment methodology for binarization algorithms based on Cohen's kappa, which has shown to be consistent with the visual inspection of the resulting images and the global rank used in DIBCO, although much simpler and direct than the multi-measure DIBCO ranking.

The thirty most widely used binarization algorithms for text document images were assessed here both in terms of quality of the resulting image and processing time. A test set of 2,257 document images was used here. Several controlled parameters were used to generate a wide variety of documents and infer how each parameter impacts the final quality and processing time in the binarization process.

It has been observed that the vertical shift of the back-to-front interference have not much impact on the quality of the binarized image. However, it significantly increases the required processing time for most algorithms. The level of blur has a small, but significant, impact on quality and no impact on processing time. For some algorithms, a decrease in blur level strongly impacted the quality of the binary image. The back-to-front interference strength (alpha) had a very strong impact on the final image quality, as it is the most important parameter to control the amount of noise added.

These results lead to conclude that studying the specific characteristics of the binary image can clarify where each algorithm performs best.

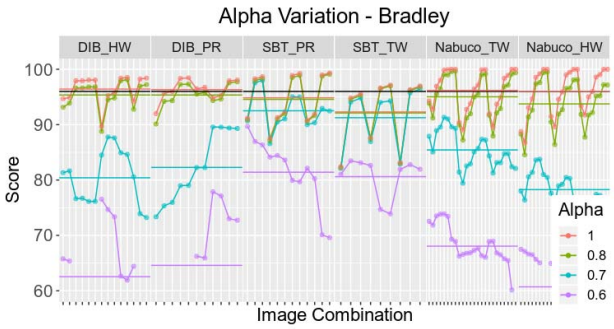

Figure 4 - Alpha variation on Score for Bradley algorithm.

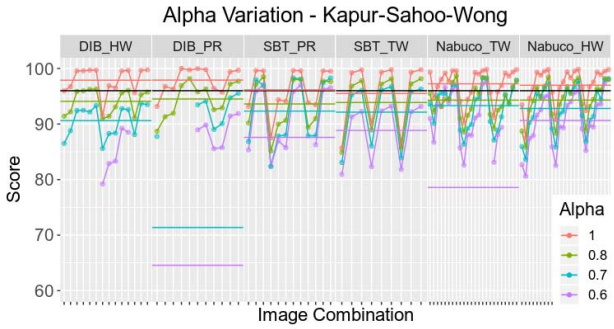

Figure 7 - Alpha variation on Score for Kapur algorithm.

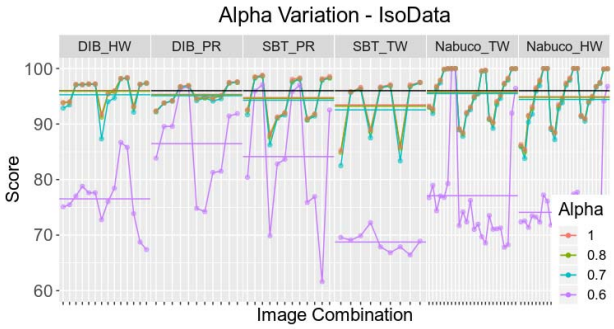

Figure 6 – Alpha variation on Score for IsoData algorithm.

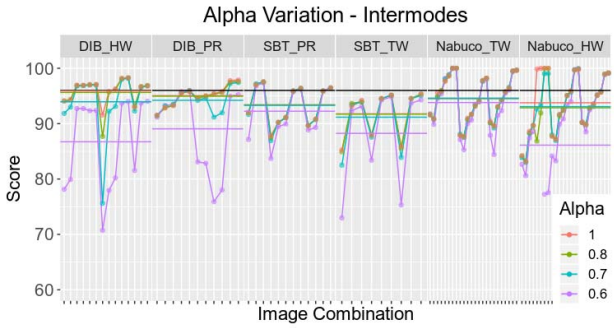

Figure 5 - Alpha variation on Score for Intermodes algorithm.

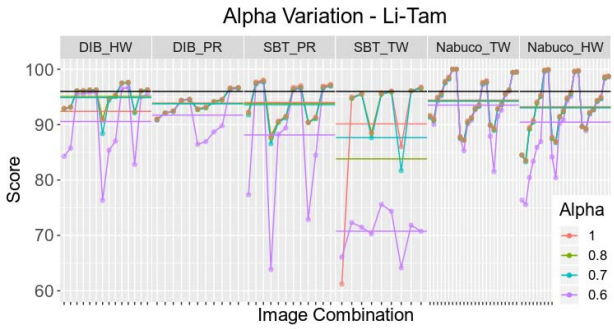

Figure 8 – Alpha variation on Score for Li-Tam algorithm.

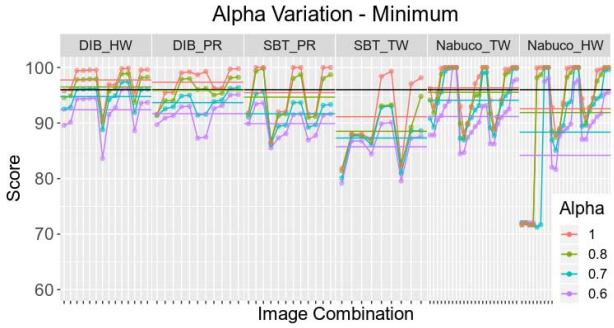

Figure 9 - Alpha variation on Score for Minimum algorithm.

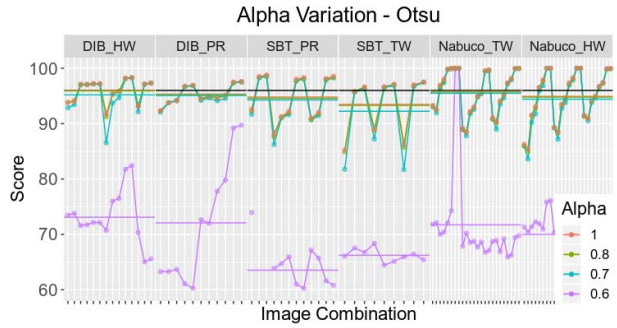

Figure 10 - Alpha variation on Score for Otsu algorithm.

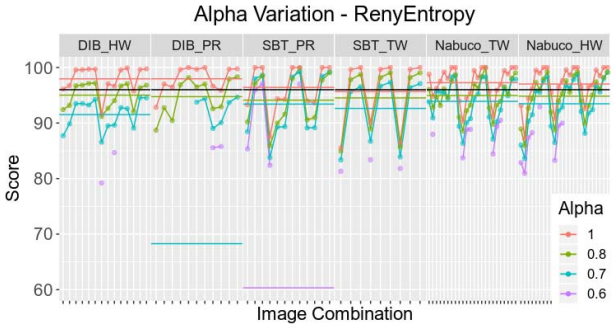

Figure 11 - Alpha variation on Score for Reny algorithm.

ACKNOWLEDGMENTS

This research was sponsored by CNPq – Brazilian Government and FACEPE.

REFERENCES

[1] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imaging*, vol. 13, no. 1, p. 146, Jan. 2004.

[2] I. Pratikakis, K. Zagori, P. Kaddas, and B. Gatos, "ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018)," *16th ICFHR* , 2018, pp. 489–493.

[3] D. Lu, X. Huang, and L. X. Sui, "Binarization of degraded document images based on contrast enhancement," *Int. J. Doc. Anal. Recognit.*, vol. 21, no. 1–2, pp. 123–135, 2018.

[4] N. R. Howe, "Document binarization with automatic parameter tuning," *Int. J. Doc. Anal. Recognit.*, vol. 16(3): 247–258, Sep. 2013.

[5] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance Evaluation Methodology for Historical Document Image Binarization," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 595–609, Feb. 2013.

[6] R. D. Lins, G. De Pereira, G. Torreão, and N. F. Alves, "Efficiently Generating Digital Libraries of Proceedings with The LiveMemory Platform," in *International Telecommunications Symposium*, 2010.

[7] R. D. Lins, "Two Decades of Document Processing in Latin America," *J. Univers. Comput. Sci.*, vol. 17(1), pp. 151–161, 2011.

[8] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sens. Environ.*, vol. 37, no. 1, pp. 35–46, Jul. 1991.

[9] J. Bernsen, "Dynamic thresholding of gray-level images," in *Inter. Conference on Pattern Recognition*, 1986, pp. 1251–1255.

[10] L. K. Huang and M. J. J. Wang, Image thresholding by minimizing the measures of fuzziness, *Pattern Recognit.*, vol. 28(1):41–51, 1995.

[11] J. Johannsen and G. Bille, A threshold selection method using information measures, *Int'l Conf. Patt. Recog.*, 1982, pp. 140–143.

[12] C. Glasbey, "An Analysis of Histogram-Based Thresholding Algorithms," *Graph. Model. Image Process.*, 55(6): 532–537, 1993.

[13] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognit.*, vol. 19, no. 1, pp. 41–47, Jan. 1986.

[14] W. Niblack, *An introduction to digital image processing.* Strandberg Publishing Company, 1985.

[15] W. Doyle, "Operations Useful for Similarity-Invariant Pattern Recognition," *J. ACM*, vol. 9, no. 2, pp. 259–267, Apr. 1962.

[16] T. Pun, "Entropic thresholding, a new approach," *Comput. Graph. Image Process.*, vol. 16, no. 3, pp. 210–239, 1981.

[17] P. L. Rosin, "Unimodal thresholding," *Pattern Recognit.*, vol. 34, no. 11, pp. 2083–2096, 2001.

[18] A. G. G. Shanbhag, "Utilization of Information Measure as a Means of Image Thresholding," *CVGIP Graph. Model. Image Process.*, vol. 56, no. 5, pp. 414–419, 1994.

[19] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency.," *J. Histochem. Cytochem.*, vol. 25, no. 7, pp. 741–753, 1977.

[20] D. Bradley and G. Roth, "Adaptive Thresholding using the Integral Image," J. Graph. Tools, vol. 12, no. 2, pp. 13–21, Jan. 2007.

[21] J. M. M. da Silva, R. D. Lins, and V. C. da Rocha, "Binarizing and filtering historical documents with back-to-front interference, ACM - SAC '06, 2006, pp. 853–858.

[22] E. Kavallieratou, "A binarization algorithm specialized on document images and photos" , ICDAR 2005, no. 1, pp. 463–467, 2005.

[23] E. Kavallieratou and S. Stathis, "Adaptive binarization of historical document images," Int. Conf. P. Recognit., vol. 3, pp. 742–745, 2006.

[24] J. M. S. Prewitt and M. L. Mendelsohn, "The analysis of cell images," Ann. N. Y. Acad. Sci., vol. 128(3): 1035–1053, Dec. 2006.

[25] F. R. Velasco, "Thresholding Using the Isodata Clustering Algorithm," Mar. 1979.

[26] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," Comput. Vision, Graph. Image Process., vol. 29, no. 1, p. 140, 1985.

[27] C. H. Li and P. K. S. Tam, "An iterative algorithm for minimum cross entropy thresholding," Pat. Recognit. Lett., vol. 19(8):771–776, 1998.

[28] C. A. B. Mello and R. D. Lins, "Image segmentation of historical documents," Visual 2000, 2000.

[29] W.-H. Tsai, "Moment-preserving thresolding: A new approach," Comput. Vision, Graph. Image Process., vol. 29(3):377–393, 1985.

[30] K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of Niblack inspired binarization methods for ancient documents," in SPIE Proceedings, 2009, p. 72470U.

[31] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man. Cybern., vol. 9, no. 1, pp. 62–66, 1979.

[32] P. Sahoo, C. Wilkins, and J. Yeager, "Threshold selection using Renyi's entropy," Pattern Recognit., vol. 30, no. 1, pp. 71–84, 1997.

[33] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.

[34] C. Wolf, J.-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in Object recognition supported by user interaction for service robots, 2003, vol. 2, pp. 1037–1040.

[35] W. Lu, M. Songde, and H. Lu, "An effective entropic thresholding for ultrasonic images," Proc. 14th Int. Conf., vol. 2:1552–1554, 1998.

[36] J. C. Yen, F. J. Chang, and S. Chang, "A New Criterion for Automatic Multilevel Thresholding," IEEE Trans. Image Process., vol. 4, no. 3, pp. 370–378, 1995.

[37] D. M. W. Powers, "Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technolagies*, vol. 2, no. 1, pp. 37–63, 2011.