Journal of Universal Computer Science, vol. 14, no. 2 (2008), 299-313
submitted: 15/6/07, accepted: 17/11/07, appeared: 28/1/08 © J.UCS

A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference

João Marcelo Monte da Silva
(Universidade Federal de Pernambuco, Recife, Brazil
joaommsilva@gmail.com)

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

Fernando Mário Junqueira Martins
(Universidade do Minho, Braga, Portugal
fmm@di.uminho.pt)

Rosita Wachenchauzer
(Universidad de Buenos Aires, Buenos Aires, Argentina
rositaw@gmail.com)

Abstract: "Back-to-front interference", "bleeding" and "show-through" is the name given to the phenomenon found whenever documents are written on both sides of translucent paper and the print of one side is visible on the other one. The binarization of documents with back-to-front interference with standard algorithms yields unreadable documents. This paper presents a fast entropy-based segmentation method for generating high-quality binarized images of documents with back-to-front interference.

Keywords: Document engineering, Back-to-front interference, Show through, Bleeding
Categories: H.3.3

1 Introduction

The algorithm presented here is part of a larger project for processing historical documents from before the nineteenth century belonging to the bequest of Joaquim Nabuco, a Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910). This rich file is kept by the Joaquim Nabuco Foundation [FUNDAJ, 07] (a social science research institute in Recife – Brazil) and encompasses over 6,000 letters. Those letters are of paramount importance to understand the formation of political and social structure of countries in the Americas and their relationship with other countries.

Although mankind use paper for over 2,000 years to record information, paper manufacturing in Nabuco's time added too much bleach and those documents are at risk of fast decomposition. Thus, the aim of the Nabuco Project is to preserve such information for future generations and make possible the widest access to those documents making them available on-line. To allow good compaction of such file and

efficient network transmission, documents are made available for consultation in their monochromatic version. More recently, the compression technique developed by Silva and Lins [Silva-a, 07] will make available a color version documents that "resemble" the original ones with little space overhead in relation to the monochromatic one.

The binarization of documents in Nabuco and similar bequests is more difficult than more recent ones because while the paper darkens with age, the printed part, either handwritten or typed, tends to fade. A special difficulty appears in the case whenever a document is typed or written on both sides and the opacity of the paper is such as to allow the back printing to be visualized on the front side. A new set of hues of paper and printing colors appears complicating the binarization process in such a way that the direct application of general algorithms is completely unsuitable and yields unreadable documents. This phenomenon, first addressed in the literature by [Lins, 95] was called "back-to-front interference", and later called "show-through" [Sharma, 01] or "bleeding" [Kasturi, 02]. Whenever the document is either in true-color or gray-scale the human eye is able to filter out that sort of noise keeping document readability (see Figure 1). This is not the case of the binarized image in the Figure 2, that was obtained by the direct application of the palette reduction algorithms, provided by standard commercial tools (such as Photoshop™ [Adobe, 07]).

The literature presents several different schemes for removing back-to-front interference: waterflow models [Hyun-Hwa, 05], wavelet filtering [Cao, 01], but the most successful seems to be entropy-based threshold techniques [Lins-a, 07]. A mirror filtering technique, was originally suggested by [Lins, 95], was adopted with success by [Sharma, 01] (using adaptive filter) and [Su, 07] (using Hidden Markov Model). A serious difficulty appears in the last technique: aligning the images of the two sides. All algorithms report limitations in different kinds of images (too dark paper background, too faded printing, interference restricted to part of the document, etc.). A complex filtering scheme is proposed by Nishida and Suzuki [Nishida, 03] where first the foreground components are separated from the background and interference through locally adaptive binarization for each color component and edge magnitude thresholding [Cumani, 91]. Background colors are estimated locally through color thresholding to generate a restored image, and then corrected adaptively through multi-scale analysis along the comparison of edge distributions between the original and the restored image. Due to the nature of the documents in Nabuco's bequest edge detection seems to be of little help in eliminating show-through noise, thus the Nishida-Suzuki method seems to be unsuitable for this kind of document, although this is still to be borne out by experiments.

This paper presents an efficient new algorithm to binarize images of documents with back-to-front interference that yields better images than the algorithm proposed by Silva, Lins and Rocha [Silva, 06], one of the best known algorithms referenced for this problem [Lins-a, 07]. The time performance of the proposed algorithm is compared with the six best algorithms assessed in [Lins-a, 07].

This paper is organized as follows. Section 1 presents this introduction. Section 2 overviews the six best algorithms assessed in [Lins-a, 07]. The new algorithm is described in Section 3. The results are presented and analyzed in Section 4. Finally, conclusions and lines for further work are presented in Section 5.

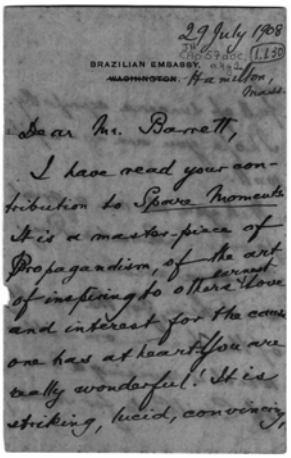

Figure 1: Document in 256 grey levels.

Figure 2: Binarized document from Figure 1.

2 Threshold Techniques

This section describes the six algorithms that exhibited the best performances in removing back-to-front interference, using the methodology outlined in reference [Lins-a, 07]. All those algorithms are based on a gray-scale image histogram threshold. The true-color to gray-scale conversion is performed by:

$$gray = 0.299r + 0.587g + 0.114b \tag{1}$$

where *gray* is the new pixel value and *r*, *g*, and *b* are the red, green and blue values of the original pixel.

All the algorithms presented here take the image histogram and normalize each of its entries by the total number of pixels in the image, yielding a probability distribution provided by relative frequencies. Thus,

$$p_i = \frac{n_i}{N},\ 0 \le i \le 255 \tag{2}$$

$$P_t = \sum_{i=0}^{t} p_i \tag{3}$$

where n_i is the number of pixels with gray level i (0 to 255), N is the total number of pixels in the image, $\{p_0, p_1, \ldots, p_{255}\}$ is the probability distribution of the pixel gray-levels taking into account their relative frequencies, and P_t is the adding of all probabilities up to entry t.

2.1 Kapur, Sahoo and Wong

The algorithm by Kapur, Sahoo and Wong [Kapur, 85] (KSW algorithm) considers the foreground and background images as two distinct sources with the following distributions

$$b:\ p(i) = \frac{p_i}{P_t},\ \ 0 \le i \le t, \tag{4}$$

$$w:\ p(i) = \frac{p_i}{1 - P_t},\ \ t+1 \le i \le 255. \tag{5}$$

Then one calculates the entropy [Abramson, 63] of the two sources

$$H_b(t) = -\sum_{i=0}^{t} p(i)\log(p(i)), \tag{6}$$

$$H_w(t) = -\sum_{i=t+1}^{255} p(i)\log(p(i)), \tag{7}$$

where $p(i)$ is determined by Equations 4 and 5. After, one determines the optimal threshold that maximizes the sum of the two entropies.

$$H(t) = H_b(t) + H_w(t). \tag{8}$$

Figure 3 presents the result of applying that algorithm to the image presented in Figure 1.

2.2 Yen, Chang and Chang

The algorithm by Yen, Chang and Chang [Yen, 95] (YCC algorithm) follows the same idea as the one by Kapur and his colleagues in respect to the foreground and background distributions. An *entropic correlation* is defined as

$$TC(t) = C_b(t) + C_w(t) = -\log\left\{\sum_{i=0}^{t}\left[\frac{p_i}{P_t}\right]^2\right\} - \log\left\{\sum_{i=t+1}^{255}\left[\frac{p_i}{1-P_t}\right]^2\right\} \tag{9}$$

and the threshold is the argument that maximizes that expression. The functions $C_b(t)$ and $C_w(t)$ are known as Ranyi entropy [MathWorld, 07], with ρ=2.

The result of applying this algorithm to the document image of Figure 1 is shown on Figure 4.

2.3 Wu, Songde and Hanqing

This algorithm (WSH algorithm) calculates the same entropies evaluated by the Kapur, Sahoo and Wong's algorithm – Equations 6 and 7. But, instead of maximizing the addition of theses, Wu, Songde and Hanqing [Wu, 98] minimize the difference given by:

$$F(t) = \left|H_b(t) - H_w(t)\right|. \tag{10}$$

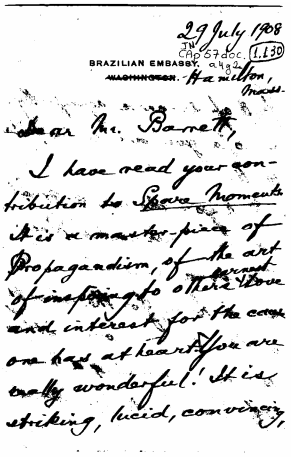

Figure 3: Filtering with algorithm by Kapur-Sahoo-Wong.

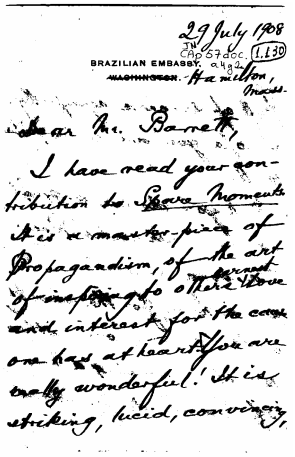

Figure 4: Filtering with algorithm by Yen-Chang-Cheng.

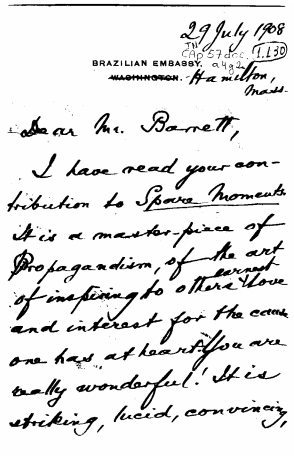

Figure 5: Filtering with algorithm by Wu-Songde-Hanqing.

Figure 5 presents the image obtained by filtering the document image of Figure 1 with the algorithm by Wu, Songde and Hanqing.

2.4 Otsu's Algorithm

The algorithm by Otsu [Otsu, 79] does not belong to the class of algorithms based on entropy. It is included here because it is one of the most often used algorithms in image segmentation. Otsu's algorithm makes discriminator analysis for defining if a gray level t will be mapped into object or background information. This algorithm works to maximize the between-class variance $\sigma_B^2(t)$ given by:

$$\sigma_B^2(t) = P_t\left(\mu_b(t) - \mu_T\right)^2 + \left(1 - P_t\right)\left(\mu_w(t) - \mu_T\right)^2 \tag{11}$$

where

$$\mu_b(t) = \sum_{i=0}^{t} i\frac{p_i}{P_t},\ \mu_w(t) = \sum_{i=t+1}^{255} i\frac{p_i}{1-P_t},\ \mu_T = \sum_{i=0}^{255} i.p_i\,. \tag{12}$$

The threshold is the argument t that maximizes the between-class variance $\sigma_B^2(t)$.

Figure 6 presents the result of the application of Otsu's algorithm to the image presented in Figure 1.

2.5 Mello and Lins

The algorithm by Mello and Lins [Mello, 00][Mello, 02] (ML algorithm) searches the most frequent gray level of the image and takes it like initial threshold to evaluate the values H_b, H_w and H by Equations 6, 7 and 8, respectively, but the entropies must be calculated with the logarithm to the base N and the $p(i) = p_i$ is determined by Equation 2. The entropy H determines two weights m_b and m_w:

- If $H \leq 0.25$, then $m_w = 2$ e $m_b = 3$.
- If $0.25 < H < 0.30$, then $m_w = 1$ e $m_b = 2.6$.
- If $H \geq 0{,}30$, then $m_w = 1$ e $m_b = 1$. (13)

And the threshold is directly calculated by

$$t^* = 256(m_b H_b + m_w H_w). \tag{14}$$

The result of applying this algorithm to the document image of Figure 1 is shown on Figure 7.

2.6 Silva, Lins and Rocha

The rationale of Silva, Lins and Rocha algorithm [Silva, 06] is to perform a statistical adjust, using the normalized entropy, between the distributions of the gray-scale and the black-and-white versions of the document image.

First, one calculates the entropy H of the gray-scale image histogram:

$$H = -\sum_{i=0}^{255} p_i \log(p_i) \tag{15}$$

where $\{p_0, p_1, \ldots, p_{255}\}$ is the *a priori probability distribution* provided by Equation 2. Then, one scans the t levels, calculating for each t the *a posteriori probability distributions* $\{P_t, 1-P_t\}$, where $P_t \leq 0.5$ is the entropy associated with that distribution:

$$H'(t) = h(P_t) \tag{16}$$

and $h(p) = -p.\log_2(p) - (1-p).\log_2(1-p)$ is the entropy function [Abramson, 63] and P_t is provided by Equation 3.

Finally, one determines the optimal limit that minimizes the $|e(t)|$ value given by:

$$|e(t)| = \left| \frac{H'(t)}{H/\log(256)} - \alpha(H/\log(256)) \right|, \tag{17}$$

where α is a loss factor, experimentally determined, given by:

$$\alpha(H/\log(256)) = \begin{cases} -\frac{3}{7} H/\log(256) + 0{,}8 & \text{if } H/\log(256) < 0{,}7 \\ H/\log(256) - 0{,}2 & \text{if } H/\log(256) \geq 0{,}7 \end{cases}. \tag{18}$$

The result of the application of this algorithm on the image presented in Figure 1 is exhibited on Figure 8.

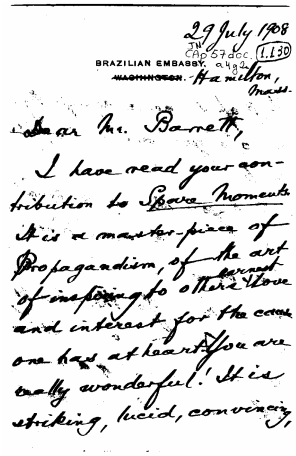

Figure 6: Filtering with algorithm by Otsu.

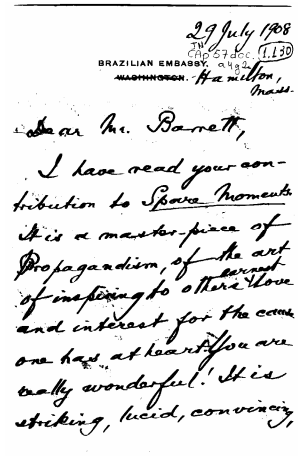

Figure 7: Filtering with algorithm by Mello-Lins.

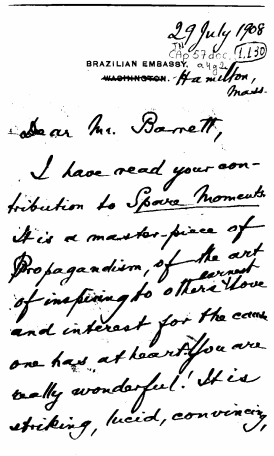

Figure 8: Filtering with algorithm by Silva-Lins-Rocha.

3 Improving the Silva, Lins and Rocha algorithm

First, an improvement for the Silva, Lins and Rocha algorithm (SLR algorithm) is presented taking into account its filtering performance. Then, two new strategies to increase the time efficiency of this algorithm are presented.

3.1 Improving The Filtering Performance

Analyzing the gray-scale images of documents from Joaquim Nabuco's bequest [FUNDAJ, 07] one can evidence that many of them do not present all the 256 gray levels. For instance, some of them have 232, 188, or 167 levels. Thus, the first alteration is in the calculation of the *normalized entropy* that was given by $H/\log(256)$ and now is given by:

$$H_{N_{\text{gray-level}}} = \frac{H}{\log(N_{\text{gray-level}})}, \tag{19}$$

where H is provided by Equation 15 and $N_{\text{gray-level}}$ is the number of gray levels present in the image. Thus, Equation 17 is replaced by

$$|e(t)| = \left| \frac{H'(t)}{H_{N_{\text{gray-level}}}} - \alpha\left(H_{N_{\text{gray-level}}}, s, m, N_{\text{gray-level}}, P_{moda}\right) \right|. \tag{20}$$

Another modification that already appears in Equation 20 is the change in the loss factor α. This factor considers, besides the normalized entropy, the number of gray levels ($N_{\text{gray-level}}$), the s and m measures based on the standard deviation and mean, respectively, of the *a priori distribution* – given by Equation 2, and the

cumulative probability (P_{moda}) from the gray-level zero to the most frequent gray-level in the image. Hence, instead using the Equation 18 the new formula is used:

$$\begin{aligned}\alpha\left(H_{N_{\text{gray-level}}}, s, m, N_{\text{gray-level}}, P_{\text{moda}}\right) = {} & 0.0267 - 0.2965 H_{N_{gray\text{-}level}} + 0.2155 H^2_{N_{gray\text{-}level}} + \\ & + 4.5897 \left(s / N_{\text{gray-level}}\right) - 6.2924 \left(s / N_{\text{gray-level}}\right)^2 + \\ & - 2.0179 \left(m / N_{\text{gray-level}}\right) + 1.3537 \left(m / N_{\text{gray-level}}\right)^2 + \\ & + 1.9632 P_{moda} - 1.2384 P^2_{moda}\end{aligned} \tag{21}$$

where s is given by

$$s = \sqrt{\sum_{j=0}^{N_{\text{gray-level}}-1} (j-m)^2 q_j}, \tag{22}$$

and q_0, q_1, ..., $q_{N\text{gray-level}-1}$ the probabilities of the *a priori distribution* – Equation 2 – that are nonzero, preserving their order, m is calculated by

$$m = \sum_{j=0}^{N_{\text{gray-level}}-1} j \cdot q_j, \tag{23}$$

and P_{moda} is evaluated by

$$P_{moda} = \sum_{j=0}^{moda} q_j, \tag{24}$$

where *moda* is the most frequent gray-level presented in the image.

The s and m measures are used instead the standard deviation and mean wherefore the algorithm may be immune to gray level histogram scaling and shift.

To obtain the loss factor expression presented in Equation 21, we selected 150 images. The selected set included historical and recent documents scanned in various resolutions, from 100 to 300dpi, with and without back-to-front interference. For each image the normalized entropy, $H_{N_{\text{gray-level}}}$; the s and m measures; the number of gray levels presented in the image $N_{gray\text{-}level}$; and the accumulative probability P_{moda} were collected. Besides them, the "best" loss factor α for the "best" binarization were found. Then, the acquired data were fitted with the curve:

$$\begin{aligned}\alpha = {} & a_0 + a_1 H_{N_{gray\text{-}level}} + a_2 H^2_{N_{gray\text{-}level}} + a_3 \left(s / N_{\text{gray-level}}\right) + a_4 \left(s / N_{\text{gray-level}}\right)^2 + \\ & + a_5 \left(m / N_{\text{gray-level}}\right) + a_6 \left(m / N_{\text{gray-level}}\right)^2 + a_7 P_{moda} + a_8 P^2_{moda}\end{aligned} \tag{25}$$

using the fit tool of MATLAB™ v7. Many kinds of curves were tested, and between them, the best results were obtained with α following Equation 25.

The result of the application of this algorithm on the image presented in Figure 1 is exhibited in Figure 9.

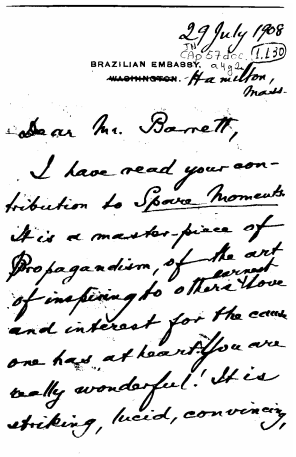

Figure 9: Filtering with improved SLR algorithm.

The effect of the improvement in the SLR algorithm can be seen in next images. In Figure 10 one has a gray-scale document image, in Figures 11 and 12 one may see the results obtained by the SLR and improved SLR algorithms, respectively.

3.2 Increasing the Time Efficiency

Now, one shows how we can improve the time efficiency of the algorithm. Two different strategies are introduced herein and their gains are compared in the next section.

3.2.1 First Strategy

The experience reported by Xerox Corp. [Xerox, 07] with photocopying is that only 5% of the pixels of a copied document are mapped onto black pixels. Statistically analyzing the documents in Nabuco's bequest, after binarization, it was found that approximately 8% of pixels are mapped onto black pixels. The idea is to use this information as a way to speed-up the improved SLR algorithm presented here. Thus, instead of scanning the whole spectrum of the different grey hues, calculating the entropy to minimize $|e(t)|$ (Equation 20), one calculates the entropy at the point which corresponds to the grey level where the accumulated total of dark pixels reaches around 8% of the total number of pixels in the document. This point (t') is called the estimated threshold.

$$E_{t'} = \sum_{i=0}^{t'} p_i \cong 8\% \forall\, p_i \neq 0, \cdot \tag{26}$$

where p_i (i=0, 1, …,t') is obtained by Equation 2, and $E_{t'}$ is the sum of all probabilities from i=0 up to entry i=t' in the histogram. One should notice that the restriction imposed that $p_i \neq 0$ plays an important role computationally as it avoids recalculating the entropy of grey levels that are not present in the image.

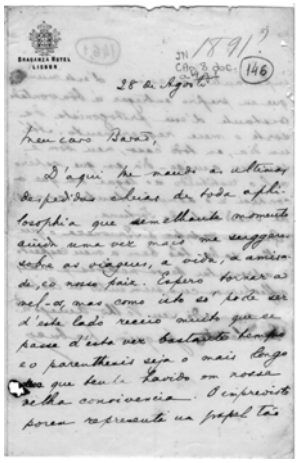

Figure 10: Gray-scale version of a document form Nabuco's bequest.

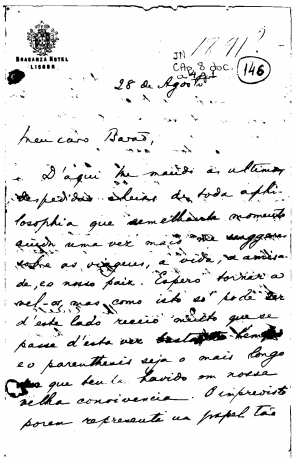

Figure 11: Figure 10 filtered with SLR algorithm.

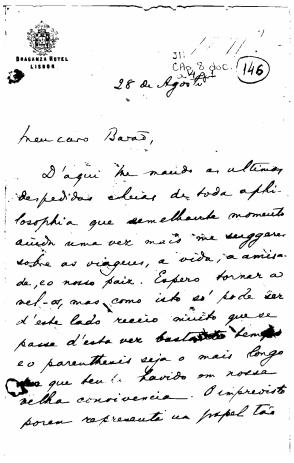

Figure 12: Figure 10 filtered with improved SLR algorithm.

Then the entropy is calculated at the two gray levels surrounding the estimated threshold, also observing that such hues must be present in the image. If the error $|e(t)|$ of the estimated threshold reaches a minimum of the three points that is the cut-off point to be adopted in binarization. Otherwise, the algorithm is repeated taking as the new estimated threshold the point which minimized $|e(t)|$. It is important to say that this first strategy was first proposed in [Silva-b, 07].

3.2.2 Second Strategy

The SLR algorithm minimizes the error $|e(t)|$ in Equation 20. This "minimization" may be expressed as:

$$H'(t) = \alpha H_{N_{\text{gray-level}}} \Rightarrow h(P_t) = \alpha H_{N_{\text{gray-level}}}, \tag{27}$$

where $H_{N_{\text{gray-level}}}$ and α are given by Equations 19 and 21, respectively, and $h(\cdot)$ is the entropy function mentioned before. As one calculates the α and $H_{N_{\text{gray-level}}}$ values *a priori*, the idea is use the injective characteristic of the entropy function $h(\cdot)$, when it is defined in a specific interval, to calculate directly the P_T value, and then one knows the threshold point T. The graphic of $h(P_t)$ is shown in Figure 13. As the SLR algorithm works with values of $P_t \leq 0.5$, the $h(P_t)$ function is injective, thus, for such interval, it exists an inverse function such that

$$h^{-1}\left(\alpha H_{N_{\text{gray-level}}}\right) = P_t. \tag{28}$$

The Figure 14 brings the graph of $h^{-1}(\cdot)$ for the aforementioned interval. Continuing the analysis, Equation 27 can be written as

$$-P_t \log(P_t) - (1-P_t)\log(1-P_t) = \alpha H_{N_{\text{gray-level}}}, \tag{29}$$

To obtain an exactly expression for $h^{-1}(\cdot)$ is not a trivial task.

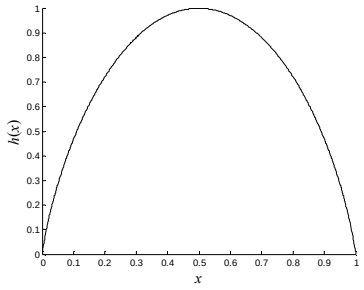

Figure 13: Graph of h(x).

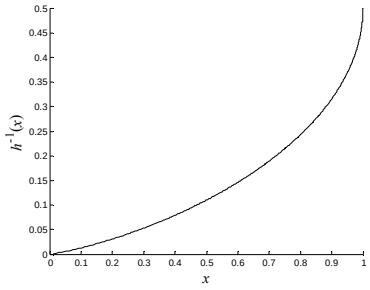

Figure 14: Graph of $h^{-1}(x)$.

From the acquired data set of the 150 images used to adjust the loss factor α, one evidences that the product $\alpha H_{N_{\text{gray-level}}}$ varies from 1.2 to 6.8, therefore, a satisfactory approximation of $h^{-1}\left(\alpha H_{N_{\text{gray-level}}}\right)$ for the interval $0.08 \le \alpha H_{N_{\text{gray-level}}} \le 0.70$ guarantees that one can calculate the P^*_T value as a direct way, and, consequently, the threshold T. The last will be given by the t value that has the closest probability P_t (Equation 3) from P^*_T.

To obtain the approximation of the function $h^{-1}\left(\alpha H_{N_{\text{gray-level}}}\right)$ for $0.08 \le \alpha H_{N_{\text{gray-level}}} \le 0.70$, one may fit it with a quadratic polynomial, using the fit fool of MATLAB™ v7. Figure 15 shows the original curve and its quadratic approximation. The approximated curve has the following expression:

$$P^*_T = 0.2419\left(\alpha H_{N_{\text{gl}}}\right)^2 + 0.09598\left(\alpha H_{N_{\text{gl}}}\right) + 0.002016. \tag{30}$$

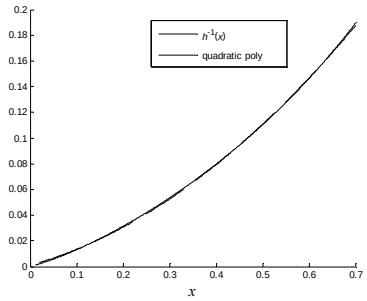

Figure 15: Graphs of $h^{-1}(x)$ and its square approximation for the interval $0.08 \le x \le 0.70$.

3.3 Summary of the Improved SLR Algorithm

1. One calculates: the normalized entropy $H_{N_{\text{gray-level}}}$ (Equation 19); the number of gray levels presented in the image $N_{\text{gray-level}}$; the *s* and *m* measures (Equations 22 and 23); and the probability P_{moda} (Equation 24).
2. One calculates the loss factor α (Equation 21).
3. One calculates the *a posteriori* probability P^*_T (Equation 30).
4. Finally, to find the threshold point, one scans the values of *t* until $P_t = P^*_T$, or as close as possible, where P_t is given by Equation 3.

Bach	KSW (Sec)	YCC (Sec)	Otsu (Sec)	ML (Sec)	WSH (Sec)	SLR (Sec)	Improved SLR	
							1 (Sec)	2 (Sec)
1	1841.38	209.02	228.58	5.42	1841.50	9.06	5.23	4.33
2	1855.71	206.58	228.95	5.34	1855.61	9.06	5.09	4.33
3	1878.61	205.64	230.20	5.25	1878.64	9.15	5.07	4.33
4	1869.97	205.97	229.66	5.27	1870.02	9.03	4.97	4.27
5	1867.57	205.99	229.59	5.38	1867.66	9.00	5.01	4.38
6	1861.81	206.45	229.23	5.60	1861.63	8.94	5.13	4.35
7	1843.97	212.11	228.29	5.52	1844.08	8.86	5.38	4.28
8	1867.78	206.39	229.53	5.60	1867.89	8.89	5.31	4.37
9	1874.89	205.94	229.92	5.57	1874.69	8.94	5.38	4.36
10	1879.25	205.58	230.20	5.30	1879.33	9.06	5.09	4.29
11	1881.39	205.47	230.05	5.31	1881.64	9.05	5.16	4.35
12	1883.05	205.55	230.39	5.17	1883.01	9.11	4.81	4.30
13	1817.17	227.89	228.12	5.00	1817.20	8.91	4.84	4.19
TOTAL	24222.54	2708.57	2982.72	69.72	24222.89	117.05	66.49	56.11
Average	1863.27	208.35	229.44	5.36	1863.30	9.00	5.11	4.32
Speed-Up	431.70	48.27	53.16	1.24	431.70	2.09	1.18	1.00

Table 1: Times in seconds collected for each batch processing repeated 5,000 times.

4 Results and Analysis

In order to analyze the gains in performance of the Improved SLR algorithm in comparison to the other ones 260 images from the Nabuco's bequest were selected, representing a fair part of the universe of documents with back-to-front interference. Images were equally divided into thirteen groups (batches) of documents and their processing times were computed. Processing time elapsed starts after histogram acquisition and stops when the algorithm finds its threshold. This process is repeated 5,000 times for each image and its computed time is added to the batch processing time that appears, in seconds, in the numbered lines on Table 1. The lines in bold TOTAL and Average are the total sum and average of the batch processing times, and the Speed-up line means how much faster the proposed algorithm (Improved SLR) is

related the others (e.g. the Improved SLR algorithm is 53.16 times faster than Otsu's algorithm). This experiments were executed in an Intel Pentium IV, 3.2 GHz, with 512 MB RAM. All the algorithms were written in standard C with Dev-C++ v4.9.8.0.

Here one is taking into account the performance of the algorithms comparing their processing time. However, one must also consider the study by [Lins-a, 07] where there is a quantitative analysis of the filtering capability of the presented algorithms. The Wu, Songde and Hanqing algorithm (WSH) is one of the slowest algorithms (vide Table 1), but the study by [Lins-a, 07] indicates this algorithm as suitable to filter images with very strong interference intensity. That same study shows that the Kapur, Sahoo and Wong's algorithm (KSW) performs similarly to the algorithm proposed by Yen, Chang and Chang (YCC) and as one can see on Table 1 the latter is approximately nine times faster then the former. The algorithm by Otsu had an intermediate performance time (see Table 1) and the study by [Lins, 07] shows that it had a filtering performance intermediate to the others.

Table 1 also shows that the algorithms by Mello and Lins, and by Silva, Lins and Rocha work faster than the other aforementioned, being the former faster than the latter. But the study by [Lins, 07] concluded that the steadiest performance in filtering out the back-to-front interference is provided by Silva, Lins and Rocha's algorithm. Thus, one may point out the Improved SLR 2 algorithm proposed here as the fastest and best quality algorithm amongst the algorithms studied. The Improved SLR 1 algorithm brings images with the same quality of the Improved SLR 2 algorithm, but it is slightly slower than it. This can be justified because the second improved algorithm calculates the threshold practically in a direct way, avoiding the computation presented in the first proposed strategy.

5 Conclusions and Lines for Further Work

This article introduces a new and efficient binarization algorithm for documents written on both sides with back-to-front interference. The new algorithm and the other six methods assessed in [Lins, 07] were applied to a set of 260 samples from Nabuco's document image bequest and it was the fastest algorithm, being twice as fast as its predecessor. The new algorithm is able to work in standalone mode, thus it can be easily incorporated into a processing environment such as BigBatch [Lins, 06].

As mentioned in the introduction, an efficient compression strategy was proposed by Silva and Lins [Silva-a, 07] using the "segmented and synthetic elements" idea to generate color documents [Lins-b, 07]. The first step of this strategy is the document segmentation, thus instead using the Silva, Lins and Rocha's algorithm (as indicated by [Lins-b, 07] and [Silva-a, 07]), the algorithm proposed here could be used for such purpose, since the latter brings better results than the former one.

Acknowledgements

Research reported herein was partly sponsored by CNPq – Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government. The authors also express their gratitude to the Fundação Joaquim Nabuco, for granting the permission to use the images from Nabuco bequest.

References

[Abramson, 63] N. Abramson, "Information Theory and Coding", McGraw-Hill Book Co, 1963.

[Adobe, 07] Adobe Systems Inc.: http://www.adobe.com.

[Cao, 01] R. Cao, C. L. Tan and P. Shen, A wavelet approach to double-sided document image pair processing, Proc. Int. Conf. Image Proc. Oct. 2001.

[Cumani, 91] A. Cumani, "Edge detection in multispectral images", G. Models and Image Processing, 53(1):40-51, 1991.

[FUNDAJ, 07] FUNDAJ – Fundação Joaquim Nabuco: http://www.fundaj.gov.br

[Hyun-Hwa, 05] O. Hyun-Hwa, et al. An improved binarization algorithm based on a waterflow model for document image with inhomogeneous backgrounds. Pattern Recognition 38 (2005) 2612 – 2625, 2005.

[Kapur, 85] 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", C.Vision, Graph. and Im.Proc., 29(3), 1985.

[Kasturi, 02] R. Kasturi, L. O´Gorman and V. Govindaraju, "Document image analysis: A primer", Sadhana, (27):3-22, 2002.

[Lins, 95] R. D. Lins, et al, "An Environment for Processing Images of Historical Documents", Microproc. & Microprogramming, pp. 111-121, North-Holland, 1995.

[Lins, 06] R. D. Lins, B. T. Ávila, and A. A. Formiga, "BigBatch: An Environment for Processing Monochromatic Documents", ICIAR2006, LNCS 4142, pp.886-896, Springer Verlag 2006.

[Lins-a, 07] R. D. Lins and J. M. M. da Silva, "A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents", In: ACM Symposium on Applied Computing, 2007, Seoul, Korea. Proceedings of SAC 2007. New York: ACM Press, 2007. p. 610-616.

[Lins-b, 07] R. D. Lins and J. M. M. da Silva, "Generating Color Documents from Segmented and Synthetic Elements", In: International Conference on Image Analysis and Recognition, 2007, Montreal, Canada. Proceedings of ICIAR 2007. Heidelberg – Germany: Springer Verlag, 2007. v. LNCS. p. 1217-1228.

[MathWorld, 07] MathWorld: http://www.mathworld.com.

[Mello, 00] C. A. B. Mello and R. D. Lins, "Image segmentation of historical documents", Visual 2000, Mexico City, Mexico. 2000.

[Mello, 02] C. A. B. Mello and R. D. Lins, "Generation of images of historical documents by composition", ACM Document Engineering 2002, McLean, VA, USA.

[Nishida, 03] H. Nishida and T. Suzuki, "A Multiscale Approach to Restoring Scanned Color Document Images with Show-trough Effects", Proc. of. ICDAR 2003, 2003.

[Otus, 79] N. Otsu, "A threshold selection method from gray level histograms", IEEE Tran. Syst. Man Cybern., 9, 62–66 (1979).

[Sharma, 01] G. Sharma, "Show-trough cancellation in scans of duplex printed documents", IEEE Trans. Image Processing, v10(5):736-754, 2001.

[Silva, 06] J. M. M. da Silva, R. D. Lins and V. C. da Rocha Jr., "Binarizing and Filtering Historical Documents with Back-to-Front Interference", In: ACM Symposium on Applied Computing, 2006, Dijon, France. Proceedings of SAC 2006. New York : ACM Press, 2006. p. 853-858.

[Silva-a, 07] J. M. M. da Silva, R. D. Lins, "Color Document Synthesis as a Compression Strategy", In: ICDAR - International Conference on Image Analysis and Recognition, 2007, Curitiba, Brazil. ICDAR 2007. New York : IEEE Press, 2007. v. I. p. 466-470..

[Silva-b, 07] J. M. M. da Silva, R. D. Lins, "A Fast Algorithm to Binarize and Filter Documents with Back-to-Front Interference", In: ACM Symposium on Applied Computing, 2007, Seoul, Korea. Proceedings of SAC 2007. New York: ACM Press, 2007. p. 639-640.

[Su, 07] F. Su and A. Mohammad-Djafari, "Bayesian Separation of Document Images with Hidden Markov Model", 2nd International Conference on Computer Vision Theory and Applications, Barcelona, Spain, 2007.

[Wu, 98] L. U. Wu, M. A. Songde, and L. U. Hanqing, "An effective entropic thresholding for ultrasonic imaging", ICPR'98: Intl. Conf. Patt. Recog., pp. 1522–1524 (1998).

[Xerox, 07] Xerox Corporation: http://www.xerox.com.

[Yen, 95] J. C. Yen, F. J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding", IEEE Trans. Image Process. IP-4, 370–378 (1995).