Thresholding Images of Historical Documents Using a Tsallis-Entropy Based Algorithm

Carlos A.B.Mello
Department of Computing and Systems, University of Pernambuco, Recife, Brazil
Email: carlos@dsc.upe.br

Luciana A.Schuler
Email: laschuler@yahoo.com.br

***Abstract*— In this paper we present an algorithm for thresholding images of historical documents. The main objective is to generate high quality monochromatic images in order to make them easily accessible through the Internet. Our new algorithm is based on two definitions of entropy: Shannon's classical concept and a variation called Tsallis entropy. For historical documents, our method proved to be more efficient than several known thresholding algorithms by several measures.**

***Index Terms*—Image processing, Document Processing, Historical documents, thresholding, entropy**

I. INTRODUCTION

Thresholding or binarization is the first step in several image processing applications [1]. It refers to the conversion of an image from true color or grayscale into a bi-level version (just two colors, in general, black and white). A threshold or cut-off value is defined and the colors above this value are converted to white while the colors below it are converted to black. Thresholding algorithms can be classified as global or local algorithms. Global thresholding algorithms define a unique cut-off value which is used to binarize the complete image. Local thresholding algorithms can be applied to similar regions of an image. These local algorithms are appropriated, for example, to binarize images of bank checks where the different fields are easily defined. In our approach, global algorithms are used as one can have more information about the probability distribution of the colors along the document.

This research is part of a major project called PROHIST [2-6] which aims to develop algorithms and techniques for preserving and broadcasting thousands of historical documents. The archive contains letters and documents from the end of the 19th century onwards amounting to more than 30,000 pages.

The documents are digitized in true color with 200 dpi resolution and they are stored in JPEG file format with 1% loss for a better quality/space storage rate. However, even in this format each image of a document reaches, on average, 400 Kb.

Hence, publishing, even in JPEG file format, all the archive, would consume several Giga bytes of space. A possible solution to this problem is the use of thresholding algorithms to convert the images to black-and-white. This should be done in a way that preserves the contents of the documents. Thus, a perfect threshold value for this application is one that generates a final bi-level image with all the colors that belong to the ink turned into black and all the colors that belong to the paper converted to white. This is quite a simple task when one deals with recent documents where the paper is almost completely white. However, in old documents, the images have low contrast. In these cases, the colors of the ink approach the colors of the paper, making the separation between them harder (Fig. 1-top).

The paper is organized as follows. Section II describes the images of the archive. Section III presents some classical thresholding algorithms while Section IV describes our new proposal. In Section V, we present the results of the application of the new thresholding algorithm and we detail the analysis of these experiments. Section V concludes the paper.

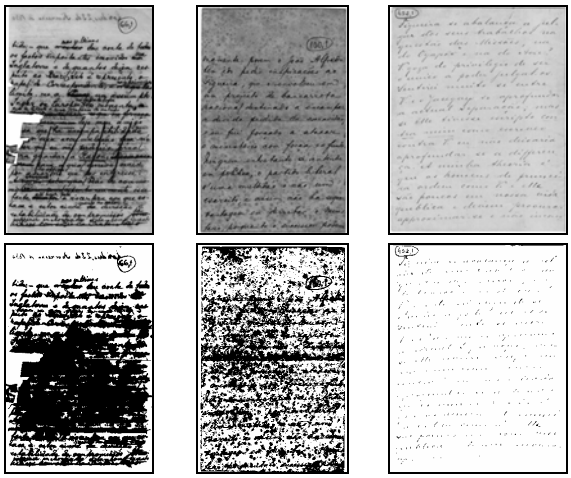


Figure 1. (top) Sample documents and (bottom) their bi-level versions: (left) a document with back-to-front interference; (center) a document with darkened paper and (right) a document with faded ink.

II. Description of the Archive

The object of our study is a collection of images of historical documents from the end of the 19th century and beginning of the 20th century. These documents are digitized in true color (16 million colors) with 200 dpi resolution. For processing, the images are first converted to grayscale by the equation:

$$gray = 0.3*R + 0.59*G + 0.11*B, \quad (1)$$

where R, G and B are the values of the red, green and blue components of a color.

There are several unique features in images of historical documents: 1) some documents are written on both sides of the paper and the ink from one side passes to the other side, creating an effect known as "ink-bleeding" or "bleeding through"; 2) in other cases the paper and the ink have very similar colors which can happen when the paper has darkened over the time or the ink has faded. Fig. 1 presents sample documents of these classes. This figure also presents the results of their binarization using Photoshop™ with its default threshold value. In addition to that features, the archive has:

1) typed and handwritten documents (and documents with a mixture of these) (Figs. 2.a, b, c);
2) documents with crossed out lines or words (Fig 2.d);
3) documents with marks of adhesive tape in the paper (Fig 2.e);
4) documents with difference of illumination (Fig 2.f);
5) documents with dark borders (Fig 2.g);
6) post cards (Fig 2.h);
7) forms (Fig 2.i).

These features are very difficult to deal using automatic image processing algorithms. The first problem encountered is to define a way to classify these different images. In our studies, we have been using entropy as a measure to classify the documents and further threshold them. At first, we used the classical Shannon definition of entropy [7]. Example of such thresholding algorithms can be found in [8] for general images and in [4][6] specifically for images of historical documents.

III. Entropy-based Thresholding Algorithms

Entropy is a measure of information. In information theory, if a source *S* has a set of *n* possible symbols s_i ($1 \le i \le n$), each with an associated probability $p(s_i)$, then, according to Claude Shannon [7], the entropy, *H*, can be evaluated as:

$$H(S) = -\sum_{i=1}^{n} p[s_i]\log(p[s_i]) \quad (2)$$

where it can be measured by bits *per* symbol.

In his work, Shannon did not define what logarithmic basis should be used (although he used basis 2 for a binary source), but Kapur [9] suggested that changes in this basis do not modify the concept of entropy. Hence, this can be very useful for thresholding purposes as can be seen in [10].

Figure 2. Sample documents from the archive: a) typewritten document; b) handwritten document; c) mixture of styles; document with d) crossed words, e) adhesive marks, f) difference of illumination and g) dark borders; h) a post card, and i) a form.

IV. New Proposed Algorithm

In our algorithm, the entropy is divided into two parts: *Hb* and *Hw*, *i.e.*, the entropy of the dark tones and the entropy of the clear tones. The break point is the most frequent color of the image, *t*. As we are working with images of documents, it is natural that this color belongs to the background (the paper). However, as some documents have dark borders, we must prevent such a color from being found to be the most frequent. So we hypothesized that the most frequent color must be above the gray level of 50. *Hb* and *Hw* can be seen as projections of the entropy itself. They are evaluated as:

$$Hb = -\sum_{i=0}^{t} p[i]\log(p[i]) \quad (3)$$

and

$$Hw = -\sum_{i=t+1}^{255} p[i]\log(p[i]) \quad (4)$$

where *p[i]* is the probability distribution of the gray tones in the image. The logarithm is taken using the area of the image as its basis, *i.e.*, height *versus* width.

The value of *H* can be found as the sum of *Hb* and *Hw*.

Shannon's entropy says that if a system can be decomposed into two statistically independent subsystems, say A and B, then *H* has the extensive or additivity property. This means that *H(A+B) = H(A) +*

H(B) (just as we did in Eqs. (3) and (4)). A newer definition of entropy was proposed by C.Tsallis in [11] for nonextensive systems.

According to Tsallis:

$$H_\alpha(S) = \frac{1 - \sum_i p(i)^\alpha}{\alpha - 1} \quad (5)$$

where p(i) is a probability as in the classical definition of entropy and α is a real parameter which value is not defined by Tsallis. When α tends to 1, Tsallis entropy reduces to Boltzmann-Gibbs entropy:

$$H(S) = -\sum_i p(i)\ln(p(i)). \quad (6)$$

Li *et al.* [12] proposed a thresholding algorithm based on Tsallis entropy. They also presented a study on the variation of the α parameter and its consequence to binarization. The method proposed, however, is more suitable for images with a bi-modal histogram (a histogram with two clearly different regions: one referring to the background objects and another to the foreground objects). This is not the case here as some documents have very low contrast, making harder to separate the objects. Fig. 3 presents an image with its bimodal histogram and a sample document from our database with its histogram.

Different values of α are defined for each class of documents. These classes group documents with some similar features and they are determined by the value of the Shannon entropy of the image, *H*, as follows:

- Class 1: $0.28 \le H \le 0.32$;
- Class 2: $H > 0.32$;
- Class 3: $H \le 0.23$;
- Class 4: $0.23 < H < 0.28$.

Details about each class are explained next.

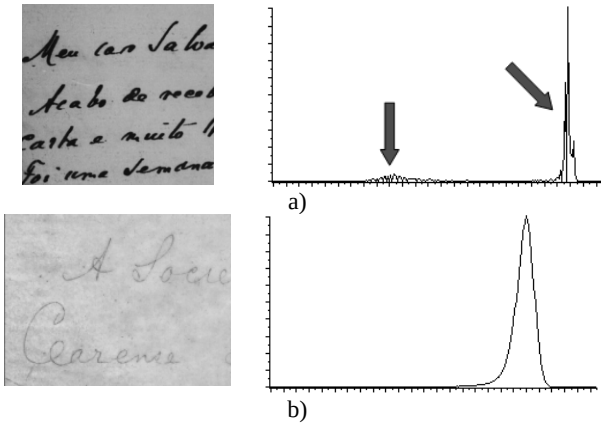


Figure 3. a) Example of an image from our archive with a clear separation between background and foreground (easily seen in its bimodal histogram) and b) a zooming into another sample document where the ink and the paper have very close colors, which produces an unimodal histogram as shown.

The *a priori* Tsallis entropy for each distribution is:

$$H_\alpha^b = \frac{1 - \sum_{i=0}^{t} (p(i)/P_b)^\alpha}{\alpha - 1} \quad (9)$$

and

$$H_\alpha^w = \frac{1 - \sum_{i=t+1}^{255} (p(i)/P_w)^\alpha}{\alpha - 1}. \quad (10)$$

With the value of α, one can evaluate the values of H_α^b and H_α^w as presented in Eqs. (9) and (10). The threshold value is then given by:

$$th = H_\alpha^b + H_\alpha^w. \quad (11)$$

The colors above *th* are converted into white and the colors below *th* are converted into black, creating a bi-level image.

Let us now explain each class and the choice of the α value for them.

Class 1 Documents ($0.28 \le H \le 0.32$)

This class is composed mostly by handwritten documents. This means that an amount of black pixels bigger than normal for documents is expected. This class also contains documents with dark borders.

This is the simplest class to deal with. For this class, α is defined as 0.3.

Class 2 Documents ($H > 0.32$)

In this class we have two types of documents. The first are the documents with ink bleeding effect. They are very different from the other documents as they have a third object besides ink and paper: the intermediary ink that comes from the verse of the paper. Some of these papers have also a dark border which must be considered. This first part of documents from class 2 is characterized by $Hw > 0.13$. For these, if the most frequent color is below 100 (which mean documents with dark border and ink bleeding) then α is equal to 0.5; otherwise, α is 0.1.

The second type of documents from this class has $Hw \le 0.13$. These are the documents with ink bleeding. The amount of ink that trespasses to the other side depends on the thickness of the paper. This can be analyzed by the most frequent color of the document, which can change the value of the α parameter. As defined before, let *t* be the most frequent color. So:

- if $220 < t < 227$, then $\alpha = 0.05$;
- if $t \le 200$, then $\alpha = 0.3$;
- and if $t \ge 227$, then $\alpha = 0.5$.

The first condition refers to the most common document papers; the second deals with documents with thinner pages; and the last condition is related to documents written in thicker papers.

Class 3 Documents (*H > 0.32*)

Class 3 contains documents with few ink parts. However, this may come from documents with just few words or documents were the ink has faded, and just a little of it remains.

For this class, if the most frequent color is greater than 200, we have to check whether this corresponds to a document with faded ink or not. For this purpose, it is necessary to check the standard deviation of the histogram. Faded ink means that the color of the ink is very close to the color of the paper; in other words, the standard deviation is low (in our case, below 15). For these cases, α is equal to 0.04; otherwise, it is 0.07. However, if the most frequent color is lower than or equal to 200, then the parameter α is 0.1.

Class 4 Documents (*0.23 < H < 0.28*)

The last class is the most difficult to deal with. It contains typewritten documents or documents with similar features (handwritten but with small letters). In general, typewritten letters have a great amount of characters but the ink is not as strong as in handwritten letters. In fact, part of the ink is always faded making the thresholding process harder. In order to binarize these images, we must preprocess the images. Initially, α is defined as 0.05 but this can change.

Most of the images are binarized after this classification. With the definition of the value of the parameter α, the algorithm applies Eqs. (9), (10) and (11) to find the threshold value.

The images with α equals to 0.05 (documents from class 4 and some of the class 2) are preprocessed before binarization and they may even have their α value changed. As some of these documents are very clear, first we apply a histogram specification [1] to the images. One of the images of Class 1 which achieved a high quality bi-level image through our algorithm is used as the base image. Its histogram is evaluated and it is considered as the expected histogram for the documents. With this, the images of this class are adjusted so they have this same histogram. Fig. 4 presents this reference histogram and the result after the application of histogram specification to an image of this class. The final binary version of this image is shown in Fig. 9 at the end of this paper.

After this, Class 4 images are filtered using a square root filter as follows. To apply the filter, the colors of the image are normalized so that they go from 0 to 1, instead of 0 to 255. The square root of each normalized pixel is evaluated and denormalized back to the normal color range from 0 to 255. Fig. 5 presents a graph that represents the changes in the color distribution with the evaluation of the square root. It shows that the colors increase to clear tones more rapidly, making the image brighter.

With the use of the square root filter, the main characteristics of the image changes and it must be analyzed again in the search of the correct α value.

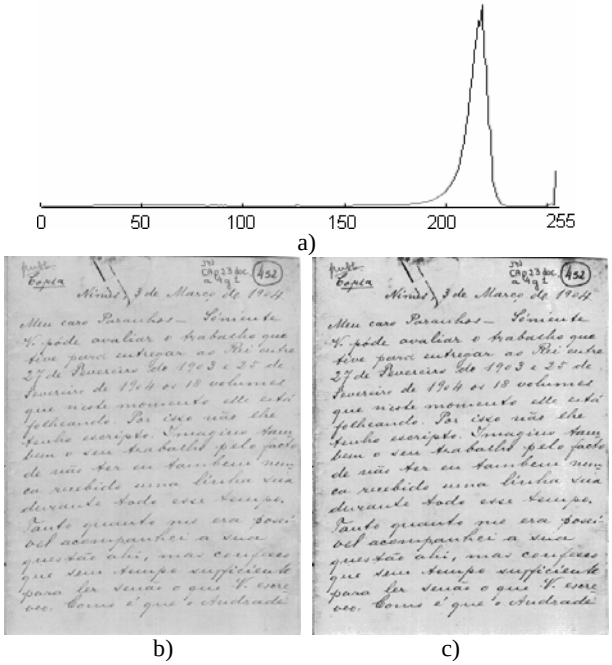


Figure 4. Histogram specification: a) Reference histogram; b) sample original document and c) the same document after the use of histogram specification based on the reference histogram.

The process is the same as before, with the search for the most frequent color, the evaluation of Shannon's entropy *H* and its decomposition into *Hw* and *Hb*, followed by the definition of α parameter. However, as we have just few classes of images, the rules are more simple. Let *H1*, *Hb1* and *Hw1* be the entropy and its decompositions evaluated over the image processed by the square root filter. α is:

- If *0.24 < H < 0.27* and *0.145 < Hb < 0.182* then:
  - Image = Image – 50 (each pixel is subtracted by 50)
  - α remains 0.05.
- Otherwise
  - If *H1 ≥ 0.28* then
    - α = 0.3
  - Otherwise
    - If *H1 ≥ 0.28* then
      - α = 0.02
    - Otherwise
      - α remains 0.05.

As said before, with α, Eqs. (9), (10) and (11) are used to define the threshold value and binarize the image.

V. RESULTS

A. Assessment Metrics

As to our knowledge there is no formal method to compare binarization algorithms, we analyzed our algorithm using four different approaches.

The first analysis was done by visual inspection. A set of 140 images were used for test. Figs. 7 and 8 at the end of the paper presents sample documents and the final bi-level image produced by the algorithm.

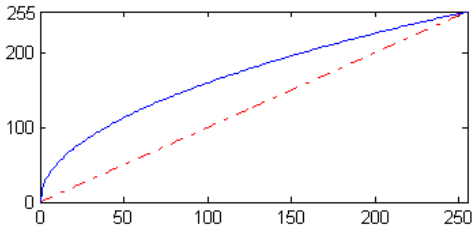


Figure 5. The effect of the square root filter in the color distribution of an image: both axis represent the colors of a grayscale palette; the dashed line represents a direct mapping of one color into itself; the continuous curve is the function y = square_root(x) evaluated over the normalized value of the colors and denormalized back.

For a quantitative evaluation of the algorithm, an ideal bi-level image was created by visually thresholding each image in the set. These images (which we used as gold standards) allow a quantitative comparison with other thresholding algorithms. For this analysis, 21 algorithms were implemented and ran in the images of the test set. As global algorithms we have: Brink, C-Means, Fisher, Huang, Johannsen, Kapur, Kittler, Li-Lee, Otsu, percentage of black, Pun, Renyi, iterative selection, two peaks, Wu-Lu, Yager, Yen, (all of these detailed in [8]) and daSilva-Lins-Rocha [13]. Examples of local thresholding algorithms are NiBlack, Savoula, White and Bernsen (also detailed in [8]).

The comparison is made using the concepts of: precision, recall, accuracy and specificity, defined based on the values of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN):

- Precision = TP/(TP + FP);
- Recall = TP/(TP + FN);
- Accuracy = (TP + TN)/(TP + TN + FP + FN);
- Specificity = TN/(FP + TN).

As a third metric we evaluated Peak Signal-to-Noise Ratio (PSNR) which is related to the Mean Square Error (MSE). PSNR (in dB) is expressed as follows:

$$PSNR = 10\log_{10}(C^2 / MSE) \qquad (11)$$

where C represents the maximum color value.

The fourth analysis is based on distances between vectors or matrixes. The following distances were evaluated: Euclidian, City Block, Minkowski, Canberra and Angular [14]. For bi-level images, Canberra and City Block distances present the same value.

B. Analysis of the Results

The results of the three quantitative metrics are disposed in Tables I which presents the average values of each metric when evaluated in a set of 140 images created by our proposal and other classical thresholding algorithms. These comparisons are made with the gold standard images as reference. Some conclusions can be taken based on this Table and the metrics themselves.

For precision, recall, accuracy and specificity, a good algorithm must have:

- Precision→1: which means FP→0, or there were few mistakes in the classification of the paper elements;
- Recall→1: meaning that FN→0 or there were few mistakes in the classification of the ink elements;
- Accuracy→1: (FP + FN)→0, *i.e.*, there was a little misclassification as possible;
- Specificity→1: indicating that FP→0 and every pixel that belongs to the paper were classified as that.

A good algorithm must have all of these four measures tending to 1 at the same time. The proposed algorithm achieved very high results for the four measures. As the best algorithm must have the four measures close to 1, their sum must tend to 4. This sum is equal to 3.59 for our method which is greater than the others.

For PSNR and MSE, high quality images must be the lower MSE which causes higher PSNR values as they inversely proportional. As both average values are shown in Table I, one can see that the images produced by our algorithm achieve the high PSNR values.

For the distances, as it is intuitive, the smaller the better and our binarized images have the smaller values for the five distances.

We also applied the proposed algorithm to another base of 145 different images from our database and the results were similar. Different image databases were also tested as documents from LOC (Library of Congress – www.loc.gov) and some documents from the site: www.site.uottawa.ca/~edubois/documents. Again the proposed algorithm proved to be very efficient as can be seen in Fig. 6, where it is shown a document from that second site which presents the ink bleeding effect and the results achieved by our algorithm.

VI. CONCLUSIONS

There is a lot of interest in having a digitally available database of historical documents. Researchers all over the world need information that is part of these databases. However, in general, to have access to the documents is not as simple as going in person to the Institution that holds the document in paper. Digitization comes as a suitable solution as it covers aspects of preservation and broadcasting.

In order to make the publishing of the contents of the documents viable, they can be digitized and converted into back-and-white images in a way that the text part (the ink) is preserved.

In this paper, we proposed a new thresholding algorithm based on the concepts of entropy as defined by the works of C.Shannon and C.Tsallis. Our algorithm can be applied to images of historical documents, and we considered several possible cases, such as faded ink, darkened paper, bleeding through effect, amongst others. In some cases, preprocessing is needed through histogram specification and the use of a square root filter.

Our algorithm proved to be more efficient than several other thresholding algorithms. This could be concluded analyzing its results using dissimilarities measures, and elements from signal detection theory as precision, recall, accuracy and specificity.

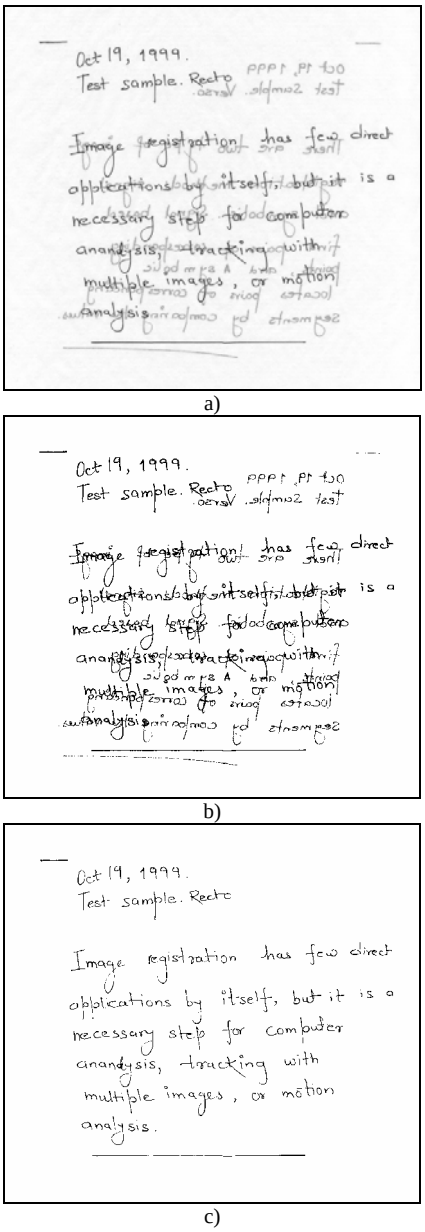
The next phase of the PROHIST project is going to use these bi-level images to segment the documents. Future

works on binarization are going to use different color systems besides RGB.

ACKNOWLEDGMENTS

The authors wish to thank CNPq, FACEPE, University of Pernambuco and the Joaquim Nabuco Foundation.

We would also like to thank Dr. Claudia Mello-Thoms for her valuable suggestions.



a) b) c)

Figure 6. a) Sample document from another database which presents bleeding through interference, b) the result after applying Otsu's algorithm and c) the bi-level image generated by our new algorithm.

REFERENCES

[1] J.R.Parker, *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons, 1997.

[2] http://recpad.dsc.upe.br/prohist

[3] C.A.B.Mello, A.L.I.Oliveira, A.Sanchez and A.Lopes Filho, "An Efficient Gray-Level Thresholding Algorithm for Historic Document Images", *Journal of Cultural Heritage*, Ed. Elsevier, 2008, vol. 9, no. 2, pp. 109-116.

[4] C.A.B.Mello, A.L.I.Oliveira and A.Sanchez, "Historical Document Image Binarization," *International Conference on Computer Vision Theory and Applications*, 2008, Funchal, pp. 108-113.

[5] E.R.Silva Jr et al., "Feature Selection and Model Design through GA Applied to Handwritten Digit Recognition from Historical Document Images," *International Conference on Frontiers in Handwritten Recognition*, 2008, *in press*.

[6] C.A.B.Mello and L.A.Schuler, "Tsallis Entropy-Based Thresholding Algorithm for Images of Historical Documents," *International Conference on Systems, Man and Cybernetics*, 2007, Montreal, pp. 1112-1117.

[7] C.Shannon, "A Mathematical Theory of Communication," in *Bell System Technology Journal*, 1948, vol. 27, pp. 370-423, 623-656.

[8] M.Sezgin and B.Sankur: "Survey over image thresholding techniques and quantitative performance evaluation", *Journal of Electronic Imaging*, 2004, vol. 13, no. 1, pp. 146-168.

[9] J.N.Kapur, *Measures of Information and their Applications*, J.Wiley & Sons, 1994.

[10] C.A.B.Mello, "A New Entropy and Logarithmic Based Binarization Algorithm for Grayscale Images," *IASTED International Conference on Signal and Image Processing*, 2004, pp. 418-423.

[11] C.Tsallis, "Possible Generalization of Boltzmann-Gibbs statistics", *Journal of Statistical Physics*, 1988, vol. 52, nos. 1-2, pp. 479-487.

[12] Y.Li, X.Fan and G.Li, "Image Segmentation based on Tsallis-entropy and Renyi entropy and Their Comparison", *International Conference on Industrial Informatics*, 2006, Singapore, pp. 943-948.

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

[14] A.Webb, *Statistical Pattern Recognition*, Ed. John Wiley Professional, 2002.

Carlos A.B.Mello was born in Recife, Brazil, in 1971. He received the B.Sc. in Electronic Engineering from the Federal University of Pernambuco (UFPE), Brazil, in 1994. He received a M.Sc. and Ph.D. in Computer Science from UFPE in 1996 and 2002, respectively.

He is currently at the Department of Computing and Systems at University of Pernambuco, Recife, Brazil, where he was Head of Department (2002-2003), undergraduate Coordinator of Computer Engineering (2002-2005), and graduate Coordinator in Computer Engineering (2006-2008). His research interests are document and image processing, texture analysis, segmentation, and pattern recognition in medical images.

Luciana A.Schuler was born in Recife, Brazil, in 1972. She received the B.Sc. in Electronic Engineering from the Federal University of Pernambuco (UFPE), Brazil, in 1996.

She is currently at Claro, an América Móvil subsidiary, in Recife, Brazil. Her research interests are in telecommunications.

TABLE I.
AVERAGE VALUES OF PRECISION (P), RECALL (R), ACCURACY (A), SPECIFICITY (S), PSNR, MSE AND SOME DISTANCES IN A SET OF 140 DOCUMENTS BINARIZED BY THE NEW ALGORITHM AND CLASSICAL ALGORITHMS COMPARED WITH THEIR GOLD STANDARD VERSIONS GENERATED MANUALLY.

Algorithm	P	R	A	S	PSNR	MSE	Distances				
							Euclidian	CityBlock	Minkowski	Canberra	Angular
New Algorithm	**0.7785**	**0.871**	**0.9699**	**0.9784**	**21.6521**	**0.0301**	**223.63**	**52852**	**36.629**	**52852**	**0.0007**
Brink	0.8972	0.7176	0.9378	0.9898	20.5486	0.0622	2794	110300	41.4436	110300	0.0008
CMeans	0.9656	0.4752	0.7276	0.9919	15.5124	0.2724	5763	507900	65.4192	507900	0.0013
daSilva-Lins-Rocha	0.8882	0.7277	0.9459	0.9908	20.3931	0.0541	2922	99000	43.3403	99000	0.0007
Fisher	0.9085	0.7274	0.9329	0.9903	20.5406	0.0671	2822	115500	41.6218	115500	0.0008
Huang	0.9174	0.6946	0.9146	0.9904	19.8345	0.0854	3286	181300	45.2472	181300	0.0008
Johannsen	0.9398	0.6548	0.9317	0.9942	19.3038	0.0683	5333	420300	62.8037	420300	0.0015
Kapur	0.0148	0.4073	0.9021	0.9013	16.8976	0.0979	3053	111700	44.6049	111700	0.0007
Kittler	0.8587	0.7982	0.9429	0.9847	21.1809	0.0571	3889	161600	52.9015	161600	0.0007
Li-Lee	0.872	0.7313	0.9447	0.9899	20.3123	0.0553	2580	105400	39.2511	105400	0.0008
Otsu	0.8431	0.7128	0.9422	0.9683	19.1462	0.0578	2407	62400	38.3805	62400	0.0007
Percentage of Black	0.9902	0.2156	0.6165	0.9985	10.3754	0.3835	3114	111300	45.2421	111300	0.0007
Pun	0.941	0.6513	0.9395	0.9957	19.3791	0.0605	8344	757800	87.7744	757800	0.0009
Renyi	0.9894	0.1716	0.3683	0.9011	8.6996	0.6317	3006	103400	44.2557	103400	0.0007
Iterative Selection	0.3368	0.3356	0.9418	0.9514	18.9542	0.0582	2730	90000	41.295	90000	0.0007
TwoPeaks	0.9402	0.6288	0.9236	0.9947	18.7636	0.0764	10523	1229700	101.9167	1229700	0.0012
Wu-Lu	0.7792	0.7711	0.9411	0.9595	19.7098	0.0589	3107	104900	45.4374	104900	0.0007
Yager	0.7985	0.2502	0.7475	0.9609	12.0418	0.2525	2448	68200	38.594	68200	0.0007
Yen	0.5276	0.8058	0.9358	0.944	19.0102	0.0642	3258	124300	46.5856	124300	0.0008
Niblack	0.862	0.3911	0.8531	0.985	14.6877	0.1469	6782	490100	76.6828	490100	0.0008
White	0.8925	0.1888	0.6045	0.9755	10.0915	0.3955	3093	105600	45.2605	105600	0.0007
Bernsen	0.3855	0.757	0.7348	0.9524	15.6284	0.2652	5189	311600	63.5995	311600	0.0007
Savoula	0.839	0.8135	0.9636	0.9847	21.4086	0.0364	8542	788500	89.3042	788500	0.0009

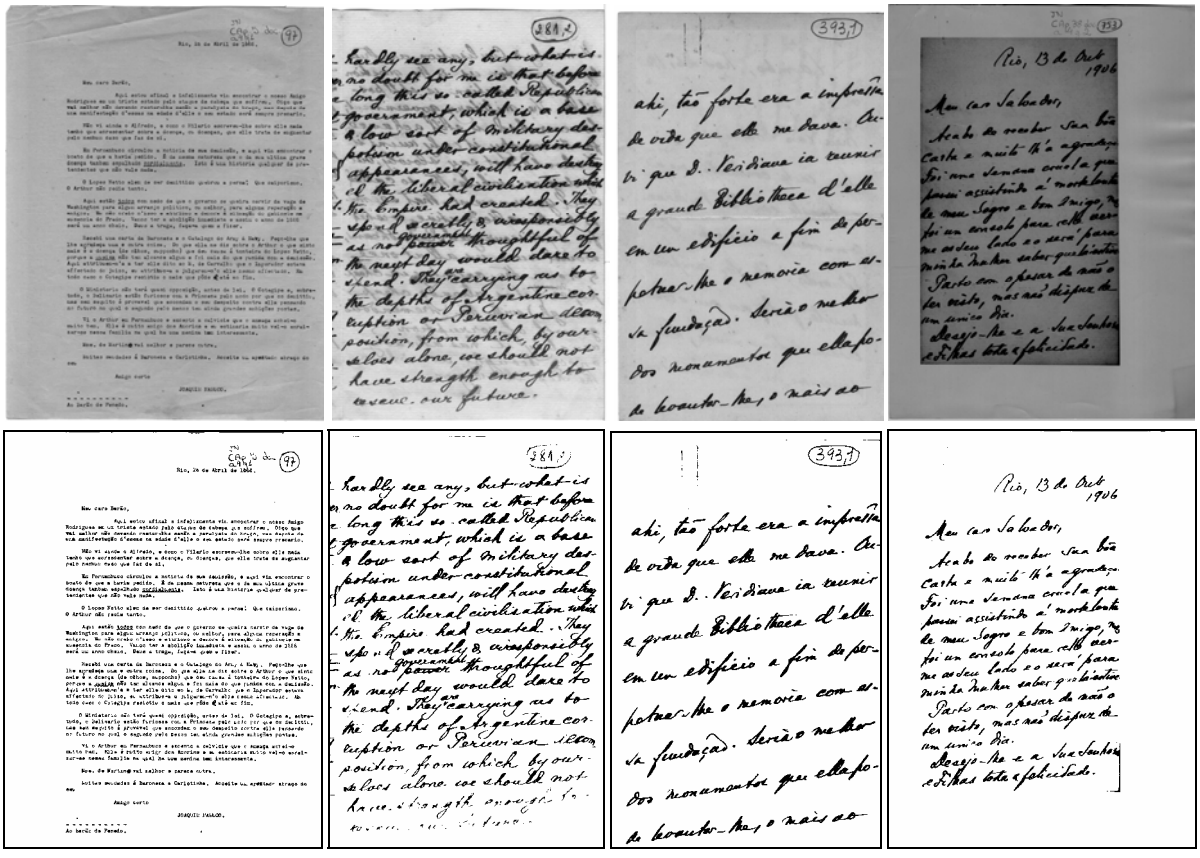


Figure 7. (top) Original sample documents and (bottom) their bi-level images generated by our algorithm.

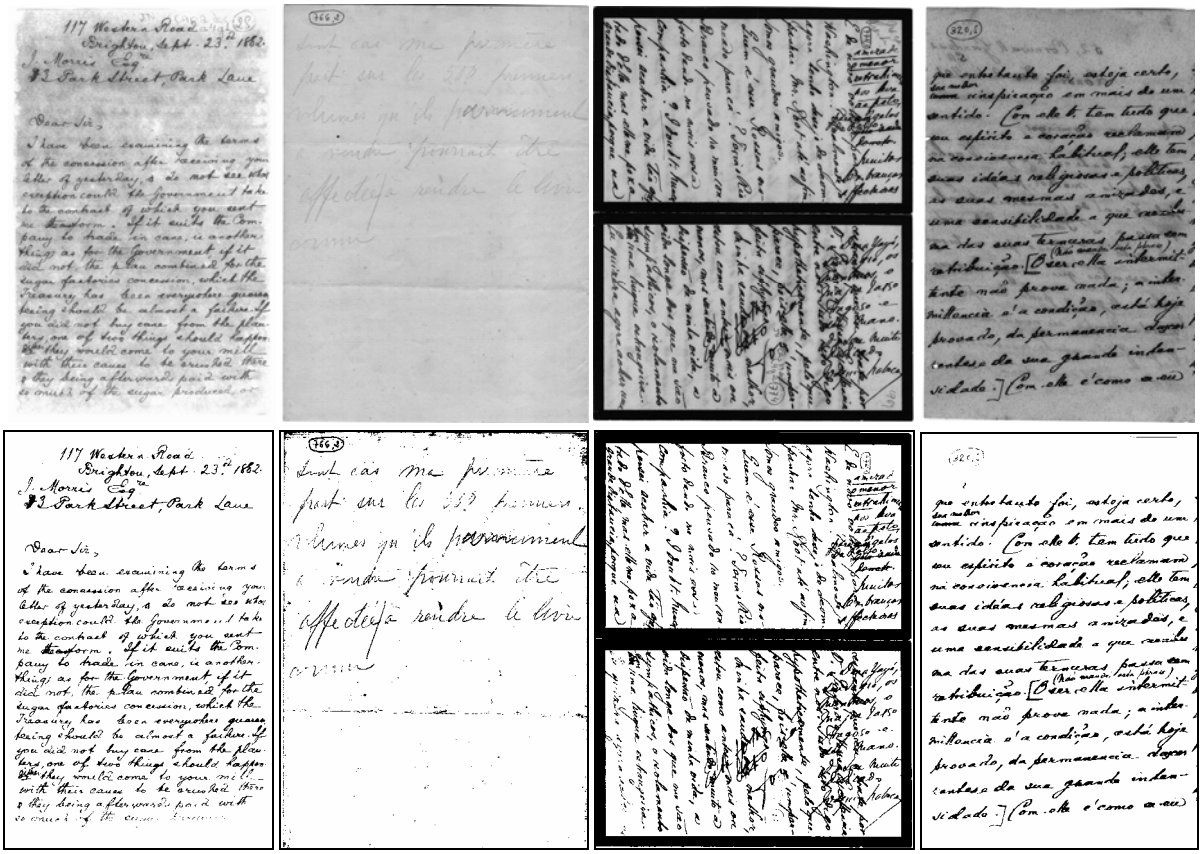


Figure 8. (top) More sample documents and (bottom) the results after binarization using the algorithm proposed in this paper.

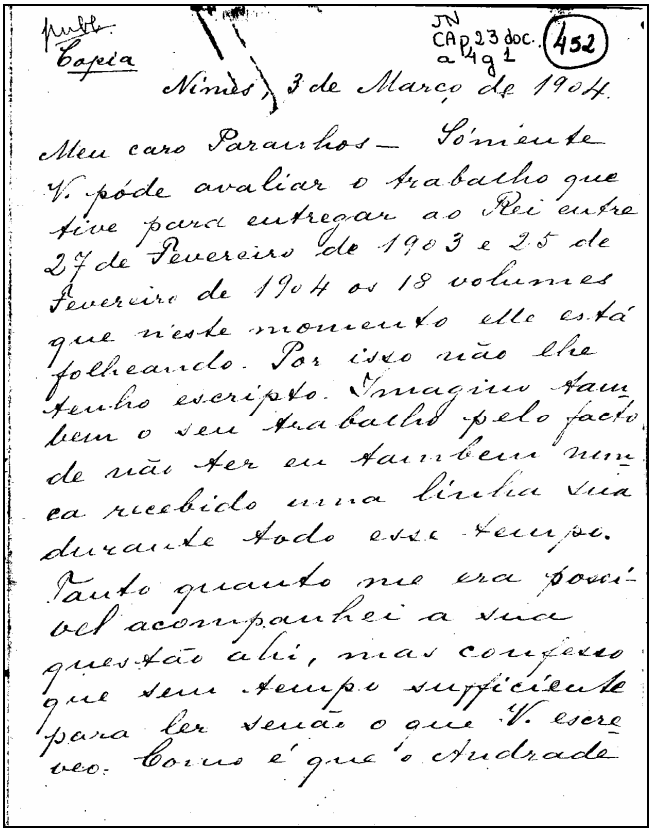


Figure 9. Binarization of the document presented in Fig. 4 which needed histogram specification.