Image Thresholding of Historical Documents based on Tsallis Entropy

Carlos A.B.Mello
University of Pernambuco, Department of Computing Systems
Recife, Brazil, 51021-370
carlos@dsc.upe.br

Abstract
This paper presents an algorithm for thresholding images of historical documents.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 and achieve high recognition rates by Optical Character Recognition algorithms. Our new algorithm is based on the classical entropy concept and a variation defined as the Tsallis Entropy and it proved to be more efficient than classical thresholding algorithms. The images generated are analyzed by quantitative measures comparing the results of the new algorithm with the ones produced by other thresholding algorithms.

Keywords: Image Thresholding, Document Processing, Image Processing, Tsallis Entropy.

1. INTRODUCTION
This research takes place in the scope of the Image Processing of Historical Documents Project (DocHist) [16][17][18] for preserving and broadcasting of a file of thousands of historical documents. This file is composed of more than 6,500 letters and documents which amounts more than 30,000 pages from the end of the nineteenth century onwards.

To preserve the file, the documents are digitized in 200 dpi resolution in true color and stored in JPEG file format with 1% loss for better quality/space storage rate. Even in this format each image of a document reaches, in average, 400 KB. In spite of the common use of broadband Internet access nowadays, the visualization of an archive of thousand of files is not a simple task. Even in JPEG file format all the archive must consume Giga Bytes of space. A possible solution to this problem is to convert the images to bi-level using thresholding techniques.

Image thresholding or binarization [20] is subject of several research efforts. It is the first step in some image processing applications as optical character recognition (OCR). Threshold algorithms search for a point that separates object and background in an image: the threshold or cut-off value. It defines which colors belong to one or another class. In the case of document images these two classes are the paper (the background) and the ink (the foreground). A good threshold value is one that generates a final bi-level image with all the colors that belong to the ink turned to black and all the colors that belong to the paper converted to white. This is quite a simple task when we deal with recent documents where the paper is almost completely clear. This process however is not so easily done in images with low contrast or other kind of noise. For these cases, image enhancement techniques could be used first to improve the visual appearance of the image for further thresholding.

Images of historical documents present some unique features that make the binarization process very hard: 1) some documents are written on both sides of the paper and the ink from one side passes to the other side, creating a back-to-front interference (also known as ink-bleeding); 2) some paper sheets are very consumed and the paper has darkened over the time (making its color similar to the ink); 3) the last case presents the documents where the ink has faded so much that it has almost the same color as the paper. Examples of these classes of documents can be seen in Figure 1. This Figure also presents the results of their binarization using a commercial tool.

Figure 2 presents a zooming into a document with back-to-front interference (the document of Figure 1-right). One can see how it is difficult to separate the colors from the transposed ink from the ink at the foreground of the document. This makes very hard a binarization process.

Next Section discusses some of the works being developed for image processing of historical documents. Classical thresholding algorithms are detailed in Section 3 of this paper. The new proposed method is explained in Section 4 and its results are analyzed in Section 5 which is followed by the Conclusions of the paper.
Figure 1. (top) Sample documents and (bottom) their bi-level versions: (left) a very faded document; (center) a document with darkened paper and (right) a document with back-to-front interference.

Figure 2. Zooming into a sample document which is written on both sides of the paper: the ink has transposed from one side to the other.

2. IMAGE PROCESSING OF HISTORICAL DOCUMENTS

Previous works related to image processing of historical documents can be found in literature. The problem of back-to-front interference is dealt in [28] where a Canny edge detector is used to detect and to suppress undesired background patterns. The method is based on the direction of the writing considering that the angle of the writing in the foreground opposes the angle of the writing in the background. This approach, however, does not deal with horizontal and vertical lines as can be found in a letter "H" for example (even if it is handwritten). The same authors also propose a new method to work with ink bleeding using the matching of the images from both sides of the paper. A wavelet reconstruction process then iteratively enhances the foreground strokes and smears the interfering strokes [29].

In [14], the authors propose the use of a multi-stage thresholding. This means that different algorithms are used in different stages of the complete process in order to create the best image possible. The authors propose this use of different algorithms and they also claim that global thresholding algorithms must not be used in this kind of images. An algorithm for background normalization is proposed in [27] to decrease the background influence for further binarization. Unfortunately, the method is adjusted for documents written on just one side of the paper.

Chen and Leedham in [3] propose the use of a quadtree decomposition to break down the image into sub-regions and to apply different thresholding algorithms in each of these regions. The choice for the best algorithm is done after a training process based on a set of documents from the Library of Congress, USA. Background removal is also considered in [12] and [1] as a first step for text segmentation of historical documents.

A combination of global and local thresholding algorithms is presented in [10]. The first step is the binarization of the document using Iterative Global Thresholding (IGT). This method decreases the contrast of the image and it darkens the image for a further histogram stretch. After this, sub-areas n by n of the image are analyzed to verify if they have more black pixels than they should have. These areas are processed again using IGT. The authors do not explain how the size of the sub-areas must be defined as this clearly changes the results of the algorithm.

A new method is detailed in [4] which consists of five distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original
image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. This algorithm is not set to documents with back-to-front interference.

Several other projects are being developed world around with historical documents as object of study:

- **HUMI Project**: The Humanities Media Interface (HUMI) Project is an initiative to launch a digital library, presenting a digital version of Gutenberg Bible. More information at its web site: http://www.humi.keio.ac.jp.
- **Glasgow University Emblem**: The Glasgow University owns the Stirling Maxwell Collection of Emblems Books which is the largest emblem collection of the world. Additional information can be found at: http://www.emblems.arts.gla.ac.uk.
- **ARTFL Project**: The Project for American and French Research on the Treasury of the French Language (ARTFL) has several collections of French texts and encyclopedias from the 16th century: http://humanities.uchicago.edu/orgs/ARTFL
- **Koninklijke Bibliotheek**: Books and journals from the Netherlands (http://www.konbib.nl/index-en.html).
- **The Digital Scriptorium**: As defined in its web site “The Digital Scriptorium is an image database of medieval and renaissance manuscripts, intended to unite scattered resources from many institutions into an international tool for teaching and scholarly research”: http://sunsite.berkeley.edu/Scriptorium
- **Cervantes Digital Library**: A digital library dedicated to the work of Miguel de Cervantes: http://www.csdl.tamu.edu/cervantes/english/index.html
- **Madonne Project**: A French initiative to use document image analysis techniques for the purpose of preserving and exploiting heritage documents: http://l3iexp.univ-lr.fr/madonne/
- **George Washington’s Manuscripts**: Digital database at the US Library of Congress containing historical manuscripts that were scanned from microfilm: http://memory.loc.gov/ammem/gwhtml/gwhome.html

Classical thresholding algorithms were tested as it is presented in the next Section. None of them achieved satisfactory results when applied to our images. Because of this, a new algorithm is presented using the classical definition of Entropy and a variation proposed by C. Tsallis [29] adjusted for the images in analysis. After digitized in true colors, the images are converted to 256 gray levels and processed in this format.

3. **THRESHOLDING ALGORITHMS**

Thresholding [20] is an important part of several image processing or pattern recognition applications. In our case, we search for means to classify the foreground (the ink) and the background (the paper). After this, the pixels classified as ink are turned to black and the pixels classified as paper are turned to white. The main problem comes when the images have low contrast. This means that there exists a gray level interval of pixel intensities where it is hard to determine whether any pixel belongs to the foreground or the background region. To obtain a perfect separation, an appropriate threshold value must be defined. The gray levels below this value are classified as ink and the gray levels above are part of the paper. The threshold value is considered correct if all the essential information of the image is preserved.

The oldest thresholding algorithms were based on simple features of the images or their histograms. The average value of the grayscale histogram is used as cut-off value in the thresholding by mean gray level algorithm [20]. Another algorithm is based on the percentage of black pixels desired [20]. For documents, in general, it is expected that 10% of the image belongs to the ink, the rest being part of the paper. So a percentage of black set to 10% achieves good results for the major part of the common documents.

In two peaks algorithm, the threshold is found at the lower point between two peaks of the histogram [20]. The authors in [25] classify thresholding algorithms based on histogram, entropy, maximization or minimization functions or fuzzy theory.

Entropy [26] is a measure of information content. In Information Theory, it is assumed that there are n possible symbols, s, which occur with probability p(s). The entropy associated with the source S of symbols is:

\[ H(S) = -\sum_{i=0}^{n} p(s_i) \log(p(s_i)) \]  

(1)

where the entropy can be measured in bits/symbols.

Five entropy-based segmentation algorithms are briefly described: Pun [21], Kapur et al [9], Li-Lee [15], Wu-Lu [31] and Renyi [24].

Pun’s algorithm [21] analyses the entropy of black pixels, \( H_b \), and the entropy of the white pixels, \( H_w \), bounded by the threshold \( t \):

\[ H_b = -\sum_{i=0}^{t} p(s_i) \log(p(s_i)) \]  

(2)
The algorithm suggests that \( t \) is such that maximizes the function \( H = H_b + H_w \).

In [9], Kapur \textit{et al.} defines a probability distribution \( A \) for the object and a distribution \( B \) for the background of the document image, such that:

\[ A: \frac{p_0}{P_t}, \frac{p_1}{P_t}, ..., \frac{p_t}{P_t} \]
\[ B: \frac{p_{t+1}}{1 - P_t}, \frac{p_{t+2}}{1 - P_t}, ..., \frac{p_{255}}{1 - P_t} \]

The entropy values \( H_w \) and \( H_b \) are as before with \( p[i] \) defined by these new distributions. The function \( H_w + H_b \) is maximized to find the threshold value \( t \).

Li-Lee algorithm [15] uses the minimum cross entropy thresholding, where the threshold selection is solved by minimizing the cross entropy between the image and its segmented version.

The main idea of the Wu-Lu algorithm [31] is the use of the lower difference between the minimum entropy of the objects and the entropy of the background. This method is very useful in ultra-sound images which have few different contrast values.

Renyi method [24] uses two probability distribution functions (one for the object and the other for the background), the derivatives of the distributions and the methods of Maximum Sum Entropy and Entropic Correlation. Other algorithms are based on the maximization or minimization of functions. Brink method [11] identifies two threshold values (\( T_1 \) and \( T_2 \)), using Brink’s maximization algorithm. The colors below \( T_1 \) are turned to black and the colors above \( T_2 \) are turned to white. The values between \( T_1 \) and \( T_2 \) are colorized analyzing the neighbors of the pixel. A 25 by 25 area is analyzed and, if there is a pixel in this area which color is greater than \( T_2 \), the pixel is converted to white.

In Kittler and Illingworth algorithm [13], based on Yan’s unified algorithm [34], the foreground and background class conditional probability density functions are assumed to be Gaussian, but, in contrast to the previous method, the equal variance assumption is removed. The error expression can be interpreted also as a fitting expression to be minimized.

Fisher method [2] consists in the localization of the threshold values between the gray level classes. These threshold values are found using a minimization of the sum of the inertia associated to the two classes.

Otsu [19] suggests minimizing the weighted sum of within-class variances of the foreground and background pixels to establish an optimum threshold. The algorithm has its basis in the linear discriminant analysis.

In a fuzzy set, an element \( x \) belongs to a set \( S \) with a grade membership \( u_e(x) \). This definition of fuzzy sets can be easily applied to the segmentation problem. Most of the algorithms use a measure of fuzziness which is a distance between the original graylevel image and the segmented one. The minimization of the fuzziness produces a most accurate binarized version of the image. We can cite three binarization algorithms that use fuzzy theory: C Means [7], Huang [6] and Yager [32].

As an adaptive algorithm, iterative selection [22] makes an initial guess of a threshold value which is refined by improving this value. The initial guess is the mean graylevel which separates two areas and the mean values of these areas are evaluated (\( T_b \) and \( T_o \)). A new estimative of the threshold is evaluated as \( (T_b + T_o)/2 \). The process is repeated using this new value of the threshold until no change is found in it in two consecutives steps. Another adaptive algorithm is the Ye-Danielsson [5] which is also implemented as an iterative thresholding.

Figure 3 shows the results of the application of some of these algorithms on the sample documents presented in Figures 1-right and center. Some cases generated very low quality images.

4. A NEW THRESHOLDING ALGORITHM

Tsallis entropy [29] has been considered a new information measure. It has been used in several image processing applications as Content Based Image Retrieval (CBIR) [23] and even thresholding [33][34]. According to Tsallis, a universal definition of entropy is given by:

\[
H_\alpha(S) = \frac{1 - \sum p(i)^\alpha}{\alpha - 1} \tag{3}
\]

where \( p(i) \) is a probability as in the classical definition of entropy and \( \alpha \) is a real parameter. When \( \alpha \) tends to 1, Tsallis entropy reduces to Boltzmann-Gibbs entropy:

\[
H(S) = -\sum p(i) \ln(p(i))
\]

Tsallis entropy is also related to Renyi’s definition of entropy \( H_\beta \) by:

\[
H_\beta(S) = \frac{1}{1 - \alpha} \ln[1+(1-\alpha)H_\alpha(S)]
\]
Shannon’s definition of entropy ($H$) [26] established in Eq. 1 settles that if a system can be decomposed into two statistical independent subsystems, say $A$ and $B$, then $H$ has the extensive or additivity property. This means that $H(A+B) = H(A) + H(B)$. This fact is used in Pun’s thresholding algorithm, for example. Tsallis entropy has a nonextensive property for statistical independent subsystems, defined by the following pseudo additivity entropic rule:

$$H_\alpha(A + B) = H_\alpha(A) + H_\alpha(B) + (1 - \alpha) H_\alpha(A) H_\alpha(B)$$

However, mathematically, Tsallis entropy (Eq. 3) can be broken into two parts:

$$H_\alpha(S) = \frac{1}{\alpha - 1} \sum_{i=0}^{255} p(i)^\alpha$$

$$H_\alpha(S) = \frac{1}{\alpha - 1} \sum_{i=0}^{255} p(i)^\alpha$$

$$H_\alpha(S) = \frac{X_b}{\alpha - 1} + \frac{X_w}{\alpha - 1} - \frac{1}{\alpha - 1} \sum_{i=0}^{255} p(i)^\alpha - \frac{1}{\alpha - 1} \sum_{i=0}^{255} p(i)^\alpha$$

$$H_\alpha(S) = \left( \frac{X_b}{\alpha - 1} - \frac{1}{\alpha - 1} \sum_{i=0}^{255} p(i)^\alpha \right) + \left( \frac{X_w}{\alpha - 1} - \frac{1}{\alpha - 1} \sum_{i=0}^{255} p(i)^\alpha \right)$$

where $X_b + X_w = 1$. It can be defined then that:

$$H_\alpha(S) = H_{b\alpha}(A) + H_{w\alpha}(B)$$

with

$$H_{b\alpha}(A) = \frac{X_b}{\alpha - 1} - \frac{1}{\alpha - 1} \sum_{i=0}^{t} p(i)^\alpha$$

and

$$H_{w\alpha}(B) = \frac{X_w}{\alpha - 1} - \frac{1}{\alpha - 1} \sum_{i=t+1}^{255} p(i)^\alpha$$

In the equations above, $t$ is the threshold value. In our case, $t$ is the most frequent color in the image. As said before, as most part of the document belongs to the paper, it is reasonable to consider that the most frequent color is part of the background. $H_{b\alpha}$ is the measure of the pixels below the color $t$ and $H_{w\alpha}$ is the measure of the colors above the threshold $t$. The variable $t$ is also used to define the values of $X_b$ and $X_w$ as $X_b$ is the percentage of colors below $t$ and $X_w$ is the percentage of colors above $t$.

The $\alpha$ parameter is a real number and it characterizes the degree of nonextensivity. Its value is not fixed in the Tsallis entropy. For thresholding purposes, variations in its value can modify the quality of the bi-level image. For our
The value of $\alpha$ is equal to 0.3 for the most part of the images. This value changes just in one case in the algorithm for one sub-class as it is shown next.

At first, the document images are separated into classes. There are three main classes of documents:

- Class 1: documents with few parts of text or documents where the ink has faded;
- Class 2: common documents with around 10% of text elements;
- Class 3: documents with more black elements than it should have; this includes documents with a black border or documents with back-to-front interference.

In order to classify an image as one of these classes, Shannon entropy ($H$) is evaluated using Equation 1 but with the logarithmic basis taken as the product of the dimensions of the image. As defined in [8], changes in the logarithmic basis do not alter the definition of the entropy. $H$ value classifies the document as:

- $H \leq 0.26$: Class 1 documents;
- $0.26 < H < 0.30$: Class 2 documents;
- $H \geq 0.30$: Class 3 documents.

These boundaries were defined analyzing a set of 500 images representatives of the complete archive. For example, the sample documents of Figure 1 belong, from left to right, to classes 1 ($H = 0.23$), 2 ($H = 0.29$) and 3 ($H = 0.32$). These classes are adjustments from previous studies presented in [16][17].

Also, the values of $H_b$ and $H_w$ (Eq. 2) are evaluated, using the most frequent color ($t$) as the separation point between the two entropies.

For each of these classes, an analysis must be made to process the images that belong to them as can be seen next. The final threshold value, $th$, is defined by:

$$th = mb * H_{ba} + mw * H_{wa}$$

where $mb$ and $mw$ are multiplicative constants that are going to be defined for each class. $H_{ba}$ and $H_{wa}$ can be seen as projections of the $H\alpha$ value; changes in those values (generated by the product by $mw$ or $mb$) make changes in $H\alpha$ itself.

**Class 1 Documents:**

As said before, this class involves documents with few ink elements. Some situations can create these documents: the letter can have just a few words or the ink can have faded. In this class, it can also be found most part of the typewritten documents as, in general, the typewriter ink is not as strong as handwritten characters and they are more susceptible to degradation decreasing of their colors.

Although the images of this class have similar features in some way, they differ in basic aspects as, for example, typewritten documents must occupy a complete sheet of paper. Because of this, another aspect must be considered within this class. We must consider the distribution of the pixels of the original image based on $H_w$ or $H_b$. We choose $H_w$ with no loss of generality.

For these images, we have:

- If ($H_w \geq 0.1$), then $mb = 2.5$ and $mw = 4.5$:
  - For example, typewritten documents with dark ink and bright paper;
- If ($0.08 < H_w < 0.1$), then $mb = mw = 4$ and $H = 0.23$:
  - As documents with the ink faded;
- If ($H_w \leq 0.8$), then $mb = mw = 4$:
  - For example, documents with dark ink and paper.

Figure 4 presents some documents from this class and its final bi-level images. Figure 5 shows in details one of the most faded documents (also presented in Figure 1-left) and the bi-level resultant image generated by the new algorithm. Although the presence of noise, this is the best image ever generated for this document, allowing an user to read its contents.

**Class 2 Documents:**

The most common documents just need a boost in $H_b$ and $H_w$ to achieve the best threshold value. So, in general, the algorithm defines $mb = 2.2$ and $mw = 3$. Some darkened documents need another treatment. If a document belongs to class 2 and $H_w > 0.1$, then the value of $mw$ decreases by half (i.e., $mw = 1.5$); $mb$ remains the same. One can see in Figure 6 sample documents from class 2 darkened or not and its thresholded images.
Class 3 Documents:
These are the documents with more black pixels than normal in a document. In this class, we have documents with a black border or documents with back-to-front interference. As the ink from one side transposes to the other side, it creates an intermediary element in the image: there is no more just paper or background; the transposed ink is an element between them. In these cases, there is no need to increase the dark measures. The system must deal just with the paper and the transposed ink turning them to white. Because of this, the \( mb \) parameter is fixed as 1. The only situation that must be dealt in this class is the documents with black border; this strongly increases the amount of black pixels in the image and the brighter components must be equally boosted. So, for common documents of this class, \( mw = 2 \), unless \( Hw < 0.099 \) which identifies documents with black borders and it makes \( mw = 9 \).

Figure 7 presents examples of the application of this algorithm for documents from class 3. Figure 8 shows the conversion to black-and-white of the detailed document of Figure 2 with a high back-to-front interference. Figure 9 presents the result of the application of the algorithm in another document with back-to-front interference.
5. **RESULTS**

The proposed algorithm was tested in a set of 500 images that are considered representative of the complete file. The results were considered very satisfactory by visual inspection. However, a most objective measure is also necessary.

In order to make a quantitative evaluation of the performance of the new algorithm, a “clean” image (with the background removed manually) was created for the documents and compared to the images generated by the algorithm. This comparison is made using the concepts of: precision, recall, accuracy and specificity, defined based on the values of true positives ($TP$ – number of ink pixels correctly classified as ink), true negatives ($TN$ – number of pixels correctly classified as paper), false positive ($FP$ – number of pixels that are part of the paper, but are misclassified as ink) and false negative ($FN$ – number of ink elements classified as paper). Together they define:

- **Precision** = $TP / (TP + FP)$
- **Recall** = $TP / (TP + FN)$

"Figure 7. (top) Some documents from class 3 and (bottom) their monochromatic versions created by the new algorithm."
• Accuracy = (TP + TN)/(TP + TN + FP + FN)
• Specificity = TN/(FP + TN)

Based on these measures, a good algorithm must have:
• Precision → 1: which means FP → 0, i.e., 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; there was no misclassification at all;
• Specificity → 1: indicating that FP → 0 and every pixel that belongs to the paper were classified as that.

A good algorithm must have all these measures tending to 1. Table 1 presents the average result for these four measures applied to a set of 50 documents binarized by the new proposed algorithm and classical algorithms in comparison with their “clean” images generated manually. As one can see, the algorithm achieved very good values for the four measures.

Table 1. Average values of precision, recall, accuracy and specificity in a set of 50 bi-level documents generated by the new proposal and classical methods compared with their “clean” version generated manually

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Algorithm</td>
<td>0.8786</td>
<td>0.8176</td>
<td>0.9713</td>
<td>0.9892</td>
</tr>
<tr>
<td>Pun</td>
<td>1</td>
<td>0.175</td>
<td>0.577</td>
<td>1</td>
</tr>
<tr>
<td>Kapur</td>
<td>1</td>
<td>0.541</td>
<td>0.915</td>
<td>1</td>
</tr>
<tr>
<td>Renyi</td>
<td>1</td>
<td>0.319</td>
<td>0.759</td>
<td>1</td>
</tr>
<tr>
<td>Brink</td>
<td>1</td>
<td>0.580</td>
<td>0.921</td>
<td>1</td>
</tr>
<tr>
<td>Otsu</td>
<td>1</td>
<td>0.657</td>
<td>0.951</td>
<td>1</td>
</tr>
<tr>
<td>Kittler</td>
<td>0</td>
<td>0</td>
<td>0.910</td>
<td>0.91</td>
</tr>
<tr>
<td>Ye</td>
<td>1</td>
<td>0.514</td>
<td>0.905</td>
<td>1</td>
</tr>
<tr>
<td>Iterative Selection</td>
<td>1</td>
<td>0.491</td>
<td>0.885</td>
<td>1</td>
</tr>
<tr>
<td>Mean Grey Level</td>
<td>1</td>
<td>0.484</td>
<td>0.882</td>
<td>1</td>
</tr>
<tr>
<td>Percentage of Black</td>
<td>1</td>
<td>0.627</td>
<td>0.947</td>
<td>1</td>
</tr>
<tr>
<td>C-Means</td>
<td>1</td>
<td>0.421</td>
<td>0.747</td>
<td>1</td>
</tr>
</tbody>
</table>

In the complete set of 500 documents, the only case that the algorithm did not generated good quality images is documents written for all over the sheet of paper but with the ink very faded (Figure 10). For these cases, ThRoc algorithm [16] achieved better quality images.

Figure 10. Type of document which did not achieved high quality images by the new algorithm: (left) original document and (right) its bi-level version.

Figure 11 shows another type of document (forms) also binarized by the new algorithm with high quality results. Figure 12 shows more examples of documents and the application of the algorithm.

6. CONCLUSIONS

This paper presents a new entropy-based thresholding algorithm for images of historical documents. The algorithm uses both Shannon and Tsallis definition of entropy to find the best cut-off value. The algorithm was applied in a set of 500 representative images of a file from the 19th century and beginning of the 20th century. The use of the algorithm was
analyzed by visual inspection and by comparison with perfect bi-level images. The values of precision, recall, accuracy and specificity were evaluated for a set of 50 documents and the algorithm achieved satisfactory results.

Three classes of documents are identified using the classical Shannon entropy definition. After this, a set of rules is used to define the best threshold value. For this, Tsallis entropy is separated into two components which are boosted in order to define the cut-off value.

The new algorithm fails just in documents with large parts of text but faded ink. This represents just 5% of the complete archive but they will be treated in further versions of the algorithm identifying the documents and applying a contrast adjustment first.

Figure 11. (top) Forms (dated from 1908 and 1907) and their bi-level images generated by the new proposed algorithm.

References


[12] Kennard, D.J. and Barrett, W.A.: Separating Lines of Text in Free-Form Handwritten Historical Documents,


Figure 12. (top) Sample documents and (bottom) their thresholded images produced by the new algorithm.