Journal of Universal Computer Science, vol. 14, no. 18 (2008), 3011-3030
submitted: 1/3/08, accepted: 29/9/08, appeared: 1/10/08 © *J.UCS*

An Evaluation Technique for Binarization Algorithms

Pavlos Stathis
(Image Processing and Multimedia Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece
pstathis@ee.duth.gr)

Ergina Kavallieratou
(Dept. of Information and Communication Systems Engineering
University of the Aegean, 83200 Karlovassi, Samos, Greece
kavallieratou@aegean.gr)

Nikos Papamarkos
(Image Processing and Multimedia Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece
papamark@ee.duth.gr)

Abstract: Document binarization is an active research area for many years. The choice of the most appropriate binarization algorithm for each case proved to be a very difficult procedure itself. In this paper, we propose a new technique for the validation of document binarization algorithms. Our method is simple in its implementation and can be performed on any binarization algorithm since it doesn't require anything more than the binarization stage. As a demonstration of the proposed technique, we use the case of degraded historical documents. Then we apply the proposed technique to 30 binarization algorithms. Experimental results and conclusions are presented.

Keywords: Document Image Processing, Binarization, Evaluation
Categories: I.7.5

1 Introduction

Document binarization is a preprocessing task, very useful to document analysis systems. It automatically converts the document images in a bi-level form in such way that the foreground information is represented by black pixels and the background by white ones.

This simple procedure has been proved to be a very difficult task, especially in the case of historical documents that very specialized problems have to be dealt with, such as variation in contrast and illumination, smearing and smudging of text, seeping of ink to the other side of the page and general degradation of the paper and ink due to aging. On the other hand, such a task is necessary for the further stages of document analysis either we are interested in performing OCR, or document segmentation, or just presentation of the document after some restoration stages. The remaining noise,

due to bad binarization, would reduce the performance of the forthcoming processing steps and in many cases could even cause their failure.

Many algorithms have been proposed for the document binarization task. However, the selection of the most appropriate one is not a simple procedure. The evaluation of these algorithms proved to be another difficult task since there is no objective way to compare the results. Weszka and Rosenfeld [Weszka and Rosenfeld 1978] defined several evaluation criteria. Palumbo et al. [Palumbo et al 1986] addressed the issue of document binarization comparing three methods. Sahoo et al. [Sahoo et al 1988] surveyed nine thresholding algorithms and illustrated comparatively their performance. Lee et al. [Lee et al 1990] conducted a comparative analysis of five global thresholding methods. Glasbey [Glasbey 1993] pointed out the relationships and performance differences between histogram-based algorithms based on an extensive statistical study. Leedham et al. [Leedham et al 2002] compared five binarization algorithms by using the precision and recall analysis of the resultant words in the foreground. He et al. [He et al 2005] compared six algorithms by evaluating their effect on end-to-end word recognition performance in a complete archive document recognition system using a commercial OCR engine. Sezgin and Sankur [Sezgin and Sankur 2004] described 40 thresholding algorithms and categorized them according to the information content used. They measured and ranked their performance comparatively in two different contexts of images.

All the above mentioned works presented some very interesting conclusions. However, the problem is that in every case, they try to use results from ensuing tasks of document processing hierarchy, in order to survey the algorithm performance. Although in many cases this is the objective goal, it is not always possible and it is an indirect evaluation approach (through subsequent analysis stages). In case of historical documents where their quality in many cases obstructs the recognition, and sometimes even the word segmentation, this way of evaluation can be proved problematic. On the other hand, we need a different evaluation technique, more direct, able to evaluate just the binarization stage. The ideal way of evaluation should be able to decide, for each pixel, if it has finally succeeded the right color (black or white) after the binarization. This is an easy task for a human observer but very difficult for a computer to perform it automatically for all the pixels of several images.

In this paper, in order to survey the algorithm performance we use for comparison a much wider range of binarization algorithms from the oldest [Doyle 1962] to the newest ones [Vonikakis et al 2008] and some interested conclusions are presented. We perform our experiments on artificial historical documents that imitate the common problems of historical documents, made by using techniques of image mosaicing and combining old blank document pages with noise-free pdf documents. This way, after the application of the binarization algorithms to the synthetic images, it is easy to evaluate the results by comparing the resulted image, pixel by pixel, with the original document. Lins [Lins and da Silva 2007] is using a similar technique to assess algorithms that remove back-to-front interference. Some first experiments of our proposed technique have already been published in [Kavallieratou 2008].

The tested binarization algorithms are very briefly presented in the next section of this paper. Then, the construction of the experimental data is described in detail in the section 3, while the experimental results and the conclusion are given in section 4 and 5, respectively.

2 Tested binarization algorithms

It is common to distinguish the binarization algorithm between global and local methods. The global algorithms calculate one threshold for the entire image, while the local thresholding algorithms calculate different threshold values depending on the local regions of the image. Here, we reference to fourteen global algorithms, fifteen local algorithms, and a hybrid one:

2.1 Global Algorithms

They are probably the faster algorithms. In the existed bibliography, we found global binarization algorithms based on 1) classification procedures, 2) histogram, 3) clustering, 4) entropy and 5) Gaussian distributions. More specifically:

1. Otsu [Otsu 1979] calculates a global threshold by accepting the existence of two classes, foreground and background, and choosing the threshold that minimizes the interclass variance of the thresholded black and white pixels. Reddi et al. [Reddi et al 1984] technique can be considered as an expansion of Otsu technique for the multithresholding case. In this work they have used it as a global thresholding technique. Its goal is the maximization of the interclass variance. Improved Integrated Function Algorithm (IIFA) [Trier and Taxt 1995] applies a gradient like operator, defined as the activity A(x, y), which is the absolute sum of approximated derivatives for both scan and raster directions taken over a small area, on the image. A three-level label-image with pixel levels '+', '-' and '0' is produced. All '+' marked regions are labeled *print*, and '-' marked regions are labeled *background*; a '0' marked region is labeled *print* if a majority of the pixels with 4-connected are '+' marked, otherwise it is labeled *background*.

2. Histogram peaks [Prewitt and Mendelsohn 1966] is the most commonly used global thresholding technique and it is based on histogram analysis. It assumes a bimodal histogram. The histogram is smoothed (using the three-point mean filter) iteratively until it has only two local maxima. Black percentage [Doyle 1962] is a parametric algorithm that assumes that the percentage of black pixels is known (p). The histogram is used and the threshold is set to the highest gray-level which maps at least (100 − p)% of the pixels into the background category. Here, we set p=5. Ramesh et al. [Ramesh et al 1995] use a simple functional approximation to the PMF consisting of a two-step function. Thus, the sum of squares between a bilevel function and the histogram is minimized, and the solution for T_o is obtained by iterative search: Rosenfeld and Kak [Rosenfeld and Kak 1982] select global threshold from the histogram of 2D image. They assume that gray values of each object are possible to cluster around a peak of the histogram of 2D image and try to compute the location of valley or peaks directly from the histogram.

3. K-means [Jain and Dubes 1988] is a clustering-based method, where the gray-level samples are clustered in two parts as background and foreground,

using the corresponding clustering algorithm. Similarly, Fuzzy c-means [Duda and Hart 1973] is a fuzzy clustering approach that the gray-scale values are clustered into two fuzzy classes corresponding to background and foreground pixels.

4. Pun [Pun 1980] considers the gray-level histogram as a G-symbol source, where all the symbols are statistically independent. He considers the ratio of the posteriori entropy as a function of the threshold to that of the source entropy. Yen et al. [Yen 1995] define the entropic correlation and obtain the threshold that maximizes it.

5. Kittler and Illingworth [Kittler and Illingworth 1985] present an algorithm that is based on the fitting of the mixture of Gaussian distributions and it transforms the binarization problem to a minimum-error Gaussian density-fitting problem. Similarly, Lloyd´s [Lloyd 1985] technique considers equal variance Gaussian density functions, and minimizes the total misclassification error via an iterative search. Finally, Riddler and Calvard [Ridler and Calvard 1978] by iterative thresholding advanced one of the first iterative schemes based on two-class Gaussian mixture models. At iteration n, a new threshold T_n is established using the average of the foreground and background class means. In practice, iterations terminate when the changes $|T_n-T_{n+1}|$ become sufficiently small.

2.2 Local Algorithms

In the existed bibliography, we found local binarization algorithms based on 1) clustering procedures, 2) local variation, 3) entropy, 4) neighborhood information, and 5) Otsu's method. More specifically:

1. The Kohonen SOM [Papamarkos and Atsalakis 2000] neural network can be used for general gray-scale reduction. Specifically, gray-level feeds the Kohonen SOM neural network classifier, and after training, the neurons of the output competition layer define the gray-level classes. If we define that the output layer has only two neurons then we perform bilevel clustering. That is, after the training stage, the output neurons specify the two classes obtained. Then, using a mapping procedure, these classes are categorized as classes of the foreground and background pixels.

2. Niblack [Niblack 1986] calculates a local threshold for each pixel that depends on the local mean value and the local standard deviation in the neighborhood of the pixel. A constant determines how much of the total print object boundary is taken as a part of the given object. The neighborhood size should be small enough to preserve local and large enough to suppress noise. It has been proven that a neighborhood *15x15* is a good choice. In a similar way, Sauvola [Sauvola and Pietikainen 2000] calculates local threshold by using the local mean value and the local standard deviation in the neighborhood of the pixel, but using a more complicate formula. Bernsen [Bernsen 1986] uses also local thresholding, calculating by the mean value of

the maximum and minimum values within a window around the pixel. When the difference of the two values is bigger than a threshold the pixel is part of the foreground, otherwise the pixel is considered as background and takes a default value.

3. Abutaleb [Abutaleb 1989] uses a local technique that considers the joint entropy of two related random variables, namely, the image gray value at a pixel, and the average gray value of a neighborhood centered at that pixel. Using the 2-D histogram, for any threshold pair, one can calculate the cumulative distribution, and then define the foreground entropy. Brink and Pendock [Brink and Pendock 1996] suggest a modification of Abutaleb's technique by redefining class entropies and finding the threshold as the value that maximizes the minimum of the foreground and background entropies. A local technique similar to previous ones is also considered from Kapur et al. [Kapur et al 1985]. The maximization of the entropy of the thresholded image is interpreted as indicative of maximum information transfer. The image foreground and background are considered as two different signal sources, so that when the sum of the two class entropies reaches its maximum, the image is said to be optimally thresholded. Johannsen and Bille [Johannsen and Bille 1982] propose an entropy-based algorithm trying to minimize the function $S_b(t) + S_w(t)$, with:

$$S_w(T) = \log\left(\sum_{i=t+1}^{255} p_i\right) + \left(1/\sum_{i=T+1}^{255} p_i\right)\left[E(p_T) + E\left(\sum_{i=T+1}^{255} p_i\right)\right]$$

$$S_b(T) = \log\left(\sum_{i=0}^{T} p_i\right) + \left(1/\sum_{i=0}^{T} p_i\right)\left[E(p_T) + E\left(\sum_{i=0}^{T-1} p_i\right)\right]$$

where $E(x)=-x\log(x)$ and T is the threshold value.

4. Palumbo at al. [Palumbo et al 1986] local algorithm consists in measuring the local contrast of five 3x3 neighborhoods organized in a center-surround scheme. The central 3x3 neighborhood Acenter of the pixel is supposed to capture the foreground while the four 3x3 neighborhoods, called in ensemble Aneigh, in diagonal positions to Acenter, capture the background. Parker's [Parker 1991] local method first detects edges and then the area between edges is filled. First, for the eight-connected neighborhood of each pixel the negative of the brightest neighbor, D, is found. Then it is broken up to regions r x r, and for each region, the sample mean and standard deviations are calculated. Both values are smoothed, and then bilinearly interpolated to give two new images, M and S, originating from the mean values and standard deviations. Then for all pixels (x,y), if $M(x,y) \geq m0$ or $S(x,y) < s0$, then the pixel is regarded as part of a flat region and remains unlabeled; else, if $D(x,y) < M(x,y) + kS(x,y)$, then (x,y) is labeled foreground; else (x,y) remains unlabeled. The resulting binary image highlights the edges. This is followed by pixel aggregation and region growing steps to locate the remaining parts of the print objects. Adaptive Local Level Thresholding (ALLT) [Yang and Yan 2000] is a local thresholding technique. Firstly, they analyze connection characteristics of the character stroke from the run-length

histogram for selected image regions and various inhomogeneous gray-scale backgrounds. Then, they propose a modified logical thresholding method to extract the binary image adaptively from the degraded gray-scale document image with complex and inhomogeneous background. Thus, it can adjust the size of the local area and logical thresholding level adaptively according to the local run-length histogram and the local gray-scale inhomogeneity. Here, the local area was set to *15x15*. Gatos et al. [Gatos et al 2006] local method claims to deal with degradations which occur due to shadows, non-uniform illumination, low contrast, large signal-dependent noise, smear and strain, so it looks appropriate for the cases we experiment. They follow several 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 while incorporating image up-sampling and finally a post-processing step in order to improve the quality of text regions and preserve stroke connectivity.

5. Liu and Li [Liu and Li 1993] proposed a 2-D Otsu thresholding method, which claim to perform much better than the 1-D Otsu method does, when images are corrupted by noise. Their method calculates the local average gray level within a limited window. They constructed a 2-D histogram, in which the x-axis and the y-axis are the gray value and the local average gray level, respectively. The optimal threshold is selected at the maximum between-class variance. Mardia and Hainsworth [Mardia and Hainsworth 1988] is a local method that performs an initial binarization using Otsu's [Otsu 1979] method. Then several steps are iterated until convergence is reached. First, the estimated mean μ and the number of pixels ni in both print and background of the current binary image are calculated. Then, a threshold t is calculated based on these values, and for each pixel a weighted mean, G, of the pixel and its eight neighbors is computed. If $G \leq t$ then the pixel is classified as "foreground", otherwise as "background". Vonikakis et al. [Vonikakis et al 2008] presents a local method whose main objective is to adopt the characteristics of the OFF-ganglion cells of the Human Visual System (HVS) and employ them in the text binarization process. OFF-ganglion cells have an antagonistic center-surround receptive field. This characteristic is also present in the artificial center-surround cells that are employed by the proposed method. Since the HVS simultaneously processes many spatial scales, four receptive field sizes, ranging from 3×3 to 15×15 pixels, are used in order to extend the performance of the proposed method from fine to coarse spatial scales. Additionally, a new activation function for the proposed OFF center-surround cells is introduced. This activation function exhibits constant responses for a document subjected to uneven illumination. Finally, the output of the OFF center-surround cells is segmented with the Otsu technique, delivering good results at various illumination levels.

2.2 Hybrid Algorithms

The Improved IGT [Kavallieratou 2005] is a hybrid approach, a combination of global applied to the whole document image, followed by local thresholding only for the areas they need it. It is based on the global IGT method and consists of the following steps: (i) apply IGT to the document image calculating a global threshold, (ii) detect the areas with remaining noise, and (iii) re-apply IGT to each detected area calculating a local threshold for each area. The IGT consists of two procedures that are applied alternately several times. Firstly, the average color value of the image is calculated and then subtracted from the image (the corresponding pixels are forced to background). In the second part of the algorithm, histogram stretching is performed, thus the remaining pixels will expand and take up all of the grayscale tones. The procedure is repeated till the difference between successive thresholds is small enough.

3 Experimental sets

The evaluation of the binarization methods was made on synthetic images. That is, starting from a clean document image (*doc*), which is considered as the ground truth image, noise of different types is added (*noisy* images). This way, during the evaluation, it is able to decide, objectively, for every single pixel if its value is correct comparing it with the corresponding pixel in the original image. Two sets of images were combined by using image mosaicing techniques.

The *doc* set consists of ten document images in pdf format, including tables, graphics, columns and many of the elements that can be found in a document. A short description of each document is given in Table 1. The *noisy* set consists of fifteen old blank images, taken from a digitized document digitized archive of the 18th century. These include most kinds of problems that can be met in old documents: presence of stains and strains, background of big variations and uneven illumination, ink seepage etc. Their description as well as their size is shown in Table 2. Samples of both sets are shown in Figure 1 and 2.

image	Description
doc_1	only text, variation in columns, variation in type and size of fonts
doc_2	only text, two columns, variation in type and size of fonts
doc_3	two columns, table
doc_4	two columns
doc_5	single column, figure
doc_6	single column, figure, formula
doc_7	printed and handwrittten text
doc_8	single column, figure
doc_9	single column, formulas
doc_10	single column, figure and graphics

Table 1: Description of doc images.

The images of the first set are all of size A4. In order to check if the relation of the size of the two images during the synthesis affects the result, we selected *noisy* images of different sizes in the second set. A wide area from less than 4% to around 350% of the size_of_noisy /size_of_doc ratio is covered. A relation of 4 % means that the *noisy* image is only 0.04 times the *doc* size (much smaller), while 350% means that the *noisy* image is 3.5 times the *doc* size (between A0 and A1).

image	description	Size
noise_1	uneven illumination, ink seepage, stains	1912x2281
noise_2	uneven illumination, ink seepage	1912x2218
noise_3	uneven illumination, ink seepage	1912x2219
noise_4	ink seepage, stains, strains	1188x889
noise_5	stains, strains, stripes	1218x1405
noise_6	uneven illumination, ink seepage, stains	1661x2335
noise_7	uneven illumination, stains	1701x2340
noise_8	uneven illumination, stains, ink seepage	2453x3502
noise_9	uneven illumination, stains	2552x3509
noise_10	background variation, stains	2552x3510
noise_11	background variation, stains	2507x3510
noise_12	uneven illumination, strains	2317x3419
noise_13	uneven illumination, strains, ink seepage	2552x3510
noise_14	uneven illumination, strains	2544x3510
noise_15	background variation, stains	949x595

Table 2: Description and size of noisy images.

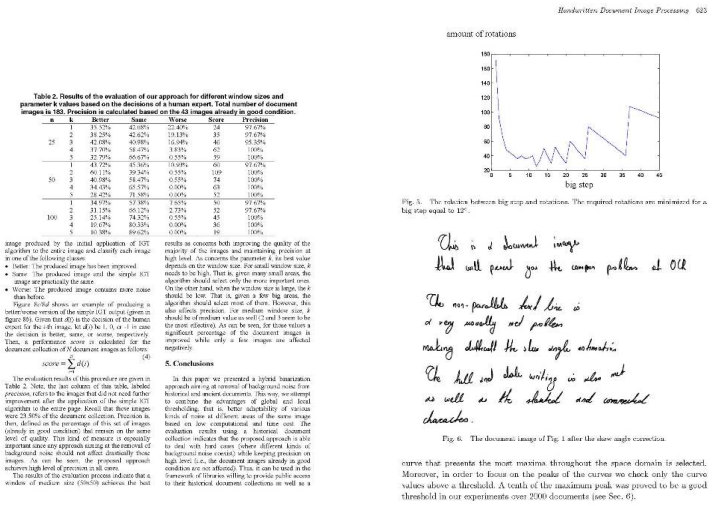


Table 2. Results of the evaluation of our approach for different window sizes and parameter k values based on the decisions of a human expert. Total number of document images is 183. Precision is calculated based on the 43 images already in good condition.

n	k	Better	Same	Worse	Score	Precision
25	1	33.33%	42.08%	23.40%	54	97.67%
	2	31.23%	42.62%	19.13%	35	97.67%
	3	42.08%	40.98%	16.94%	40	95.35%
	4	37.70%	58.47%	3.83%	65	100%
	5	32.79%	66.67%	0.55%	59	100%
50	1	41.72%	49.36%	10.93%	60	97.67%
	2	60.11%	39.34%	0.55%	109	100%
	3	40.98%	58.47%	0.55%	74	100%
	4	34.43%	65.57%	0.00%	63	100%
	5	28.42%	71.58%	0.00%	52	100%
100	1	34.93%	57.38%	7.65%	50	97.67%
	2	31.15%	66.12%	2.73%	52	97.67%
	3	25.14%	74.32%	0.55%	45	100%
	4	10.47%	89.53%	0.00%	30	100%
	5	10.38%	89.62%	0.00%	19	100%

image produced by the initial application of ICI algorithm to the entire image and classify each image in one of the following classes:

- Better : The produced image has been improved
- Same : The produced image and the simple ICI image are practically the same
- Worse: The produced image contains more noise than before.

Figure 8c-8d shows an example of producing a better/worse version of the sample ICI output (given in figure 8b). Given that $d(i)$ is the decision of the human expert for the i-th image, let $d(i)$ be 1, 0, or -1 in case the decision is better, same, or worse, respectively. Then, a performance score is calculated for the document collection of N document images as follows:

$$score = \sum_{i=1}^{N} d(i) \quad (4)$$

The evaluation results of this procedure are given in Table 2. Note, the last column of this table, labeled *precision*, refers to the images that did not need further improvement after the application of the simple ICI algorithm to the entire page. Recall that these images were 23.50% of the document collection. Precision is, then, defined as the percentage of this set of images (already in good condition) that remain in the same level of quality. This kind of measure is especially important since any approach aiming at the removal of background noise should not affect drastically these images. As can be seen, the proposed approach achieves high level of precision in all cases.

The results of the evaluation process indicate that a window of medium size (50x50) achieves the best results as concerns both improving the quality of the majority of the images and maintaining precision at high level. As concerns the parameter k, its best value depends on the window size. For small window size, k needs to be high. That is, given many small areas, the algorithm should select only the more important ones. On the other hand, when the window size is large, the k should be low. That is, given a few big areas, the algorithm should select most of them. However, this also affects precision: for medium window size, k should be of medium value as well (3 and 4 seem to be the most effective). As can be seen, for those values a significant percentage of the document images is improved, while only a few images are affected negatively.

5. Conclusions

In this paper we presented a hybrid binarization approach aiming at removal of background noise from historical and ancient documents. This way, we attempt to combine the advantages of global and local thresholding, that is, better adaptability of various kinds of noise at different areas of the same image based on low computational and time cost. The evaluation results using a historical document collection indicate that the proposed approach is able to deal with hard cases (where different kinds of background noise coexist) while keeping precision on high level (i.e., the document images already in good condition are not affected). Thus, it can be used in the framework of libraries willing to provide public access to their historical document collections as well as to

Handwritten Document Image Processing 625

amount of rotations

Fig. 5. The relation between big step and rotations. The required rotations are minimized for a big step equal to 12°.

This is a document image
that will permit you the comon problem of OCR
The non-parallels text line is
a very usually met problem
making difficult the skew angle estimation
The hand and dark writing is also not
as well as the skewed and connected
characters.

Fig. 6. The document image of Fig. 1 after the skew angle correction

curve that presents the most maxima throughout the space domain is selected. Moreover, in order to focus on the peaks of the curves we check only the curve values above a threshold. A tenth of the maximum peak was proved to be a good threshold in our experiments over 2000 documents (see Sec. 6).

Figure 1: Samples of doc images.

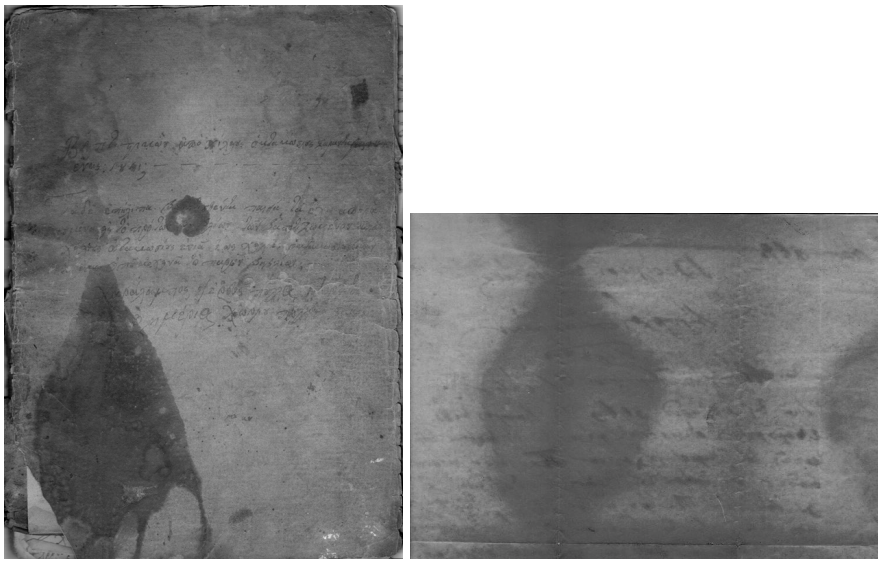
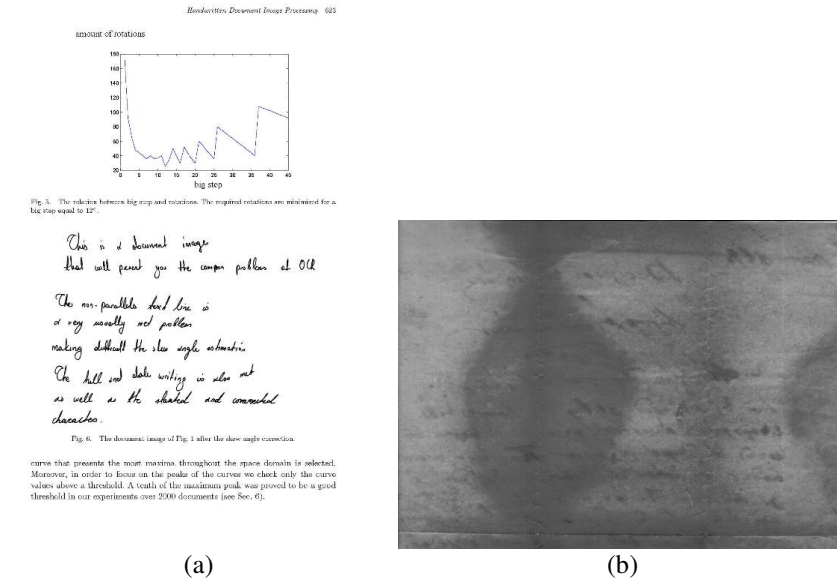
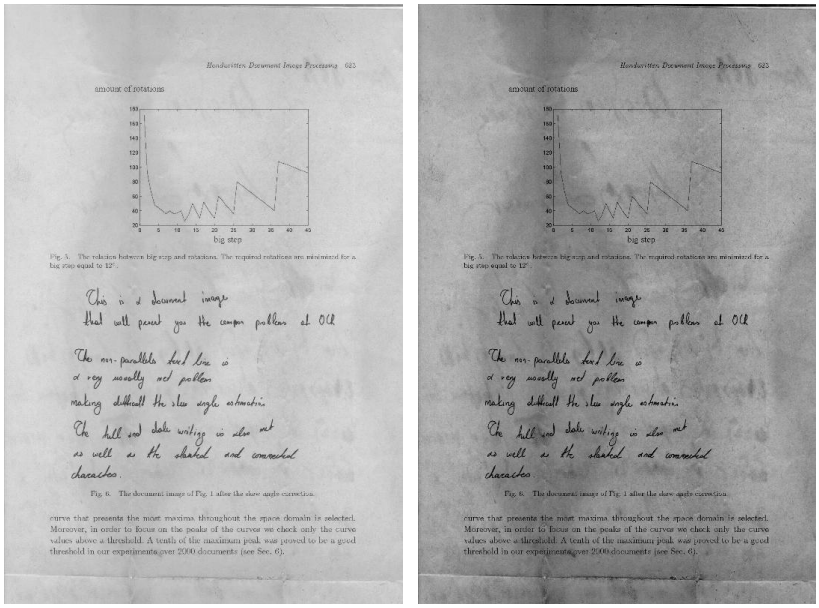


Figure 2: Samples of noisy images.

The two sets were combined by applying image mosaicing superimposing techniques for blending [Gottesfeld Brown 1992]. We built up two different sets of 150 document images each. In more detail, we used as target images the *docs* and resized all the *noisy* images to A4 size. Then, we used two different techniques for the blending: the maximum intensity and the image averaging approaches.



(a) (b)



(c) (d)

Figure 3: Construction procedure of the synthetic images: (a) doc image, (b) noisy image, (c) ave-int image and (d) max-int image.

In the first case, the maximum intensity technique (max_int), the new image was constructed by picking up for each pixel in the new image, the darkest corresponding pixel of the two images. This means that in case of foreground, the *doc* would have a lead over the *noisy*, but in the background we would have the one from the *noisy* image since it is almost always darker than the document background that is absolutely white. This technique has a good optical result as it can be seen in Figure 3 but it is not very natural as the foreground would be always the darkest, since it is not affected at all from the noise. This set permits us to check how much of the background can be detracted by a binarization method. However, in order to have a more natural result, we also used the image averaging technique (ave-int), where each pixel in the new image is the average of the two corresponding ones in the original images. In this case, the result presents a lighter background than that of the maximum intensity technique but the foreground is also affected by the level of noise in the image. The result is also shown in Figure 3 for the same images.

4 Experimental results

As we already mentioned, our intention is to be able to check for every pixel if it is right or wrong. Thus, we introduce a novel evaluation measure that we call *pixel error,* that is the total amount of pixels of the image that in the output image have wrong color: black if white in original document or white if black originally. Thus, the pixel error rate (PERR) will be:

$$PERR = \frac{pixelerror}{MxN} \tag{1}$$

In order to asses the utility of this metric, we used traditional measures of image quality description [Kite et al 2000]. More specifically, we used the square error (MSE), the signal to noise ratio (SNR) and the peak signal to noise ratio (PSNR). Let *x(i,j)* represent the value of the i-th row and j-th column pixel in the original *doc* x and let *y(i,j)* represent the value of the corresponding pixel in the output image y. Since it is all about black and white images, both values will be either 0 (black) or 255 (white). The local error is $e(i,j)=x(i,j)-y(i,j)$ and the total square error rate will be:

$$MSE = \frac{\sum_i \sum_j e(i,j)^2}{MxN} \tag{2}$$

Notice that if a pixel is right color the value of $e(i,j)^2$ will be *0*, while if the pixel is wrong color it will be 255^2 . Thus, taking into account the PERR definition, it will be:

$$PERR = MSE / 255^2 \Leftrightarrow MSE = PERR \cdot 255^2 \tag{3}$$

SNR [40] is defined as the ratio of average signal power to average noise power and for an *MxN* image is

$$SNR(DB) = 10\log_{10}\frac{\sum_i\sum_j x(i,j)}{\sum_i\sum_j \left(x(i,j)-y(i,j)\right)^2} = 10\log_{10}\frac{\sum_i\sum_j x(i,j)}{MSE} = 10\log_{10}\frac{\sum_i\sum_j x(i,j)}{PERR\cdot 255^2} \tag{4}$$

The peak measure, PSNR, depends on the word-length of the image pixels, and it is defined as the ratio of peak signal power to average noise power. For 8-bit images, as in our case, it is:

$$PSNR(DB) = 10\log_{10}\frac{255^2\cdot MN}{\sum_i\sum_j \left(x(i,j)-y(i,j)\right)^2} = 10\log_{10}\frac{255^2\cdot MN}{MSE} = 10\log_{10}\frac{MN}{PERR} \tag{5}$$

Thus, it is obvious that the three metrics MSE, SNR and PSNR depend on the PERR, however we will include them for reasons of completeness.

We applied all the methods described in section 2 to both sets described in section 3. The pixels that changed value (white-to-black or vice versa) were counted by comparing the output image with the original document image. It should be mentioned that the majority of the pixel errors are mostly white-to-black conversions with a max of 0.02‰ black-to-white conversions in both techniques.

Tables 3 and 4 show all the above mentioned metrics plus the PERR variation for max-int and ave-int techniques, respectively, in PERR ascending order. In the cases that there is no established name for a technique, we use the first author name of the corresponding paper. Next to each name in the tables 3 and 4 there is a code of the form C.S.DDDD, where C stands for the three main categories (1-global, 2-local, 3-hybrid), S corresponds to the sub-cases as they are described in section 2 and DDDD indicates the date of the paper. Moreover, the PERR values are also given in graphics of Figure 4, in order to have a visualization of the mean behavior of each algorithm and the change in their performance on each set.

	MSE	**SNR**	**PSNR**	**PERR**	**PERR variat.**
Sauvola (2.2.2000)	1105.647	17.9324	18.1326	1.700341	0.4167
Johansen (2.3.1982)	1176.348	17.66227	17.86199	1.80907	0.4898
Vonikakis (2.5.2008)	1712.938	16.28393	16.54482	2.634276	3.4973
Black Percent. (1.2.1962)	1772.267	15.51629	15.73644	2.725517	0.2751
Brink (2.3.1996)	1843.791	15.66916	15.92546	2.83551	1.3631
Histogr. peaks (1.2.1966)	1875.928	15.88174	16.01977	2.884933	5.1598
IIFA(1.1.1995)	2350.078	14.64512	14.99464	3.614115	2.7403
Li(2.5.1993)	2587.894	16.40223	16.72078	3.979844	52.695
Palumbo (2.4.1986)	2595.835	14.18978	14.54335	3.992057	3.5038
Gatos (2.4.2006)	2795.906	14.99625	15.36254	4.299741	15.996
ALLT (2.4.2000)	2922.703	14.58387	14.9179	4.494738	14.289
Reddi (1.1.1984)	4388.948	14.85988	15.36612	6.749631	161.52
Abutaleb (2.3.1989)	4404.042	11.2722	11.73721	6.772845	0.9628
Otsu (1.1.1979)	5842.581	13.26888	13.87394	8.98513	150.38
Kohonen SOM (2.1.2000)	6242.384	12.80606	13.44569	9.599975	157.17
Bernsen (2.2.1986)	6356.625	12.45118	13.08459	9.775664	138.09
Parker (2.4.1991)	8952.282	7.901008	8.661	13.76745	4.3483
IGT (3.1.2005)	9014.171	3.062373	4.19233	13.86262	0.1431
Riddler (1.5.1978)	9285.395	11.90665	12.82858	14.27973	314.35
K-means (1.3.1988)	11824.21	11.68115	12.30377	18.1841	683.19
Fuzzy C-means (1.3.1973)	13901.2	9.544554	11.21195	21.37825	721.94
Niblack (2.2.1986)	15288.62	5.023332	6.333367	23.51191	12.132
Lloyd (1.5.1985)	16567.28	5.726616	7.271567	25.47832	165.41
Kapur (2.3.1985)	18423.7	1.403792	8.922133	28.33326	1399.4
RosenfeldKak (1.2.1982)	18582.36	3.881317	5.582062	28.57725	49.992
Mardia (2.5.1988)	22771.88	2.429526	4.65036	35.0202	49.996
Ramesh (1.2.1995)	23270.71	2.789177	11.63249	35.78733	2026.4
Yen (1.4.1995)	23486.78	-1.63902	7.439523	36.11962	1494.8
Kittler (1.5.1985)	27002.61	2.016883	5.234863	41.5265	549.26
Pun (1.4.1980)	29081.33	0.61298	3.506799	44.72331	11.337

Table 3: The evaluation metrics for max-int technique.

	MSE	**SNR**	**PSNR**	**PERR**	**PERR variat.**
Johansen (2.3.1982)	1030.09	18.29947	18.49771	1.584145	0.39702
Li(2.5.1993)	1064.482	18.11257	18.31684	1.637035	0.39598
Reddi (1.1.1984)	1067.702	18.1055	18.31057	1.641987	0.407469
ALLT (2.4.2000)	1080.179	18.00511	18.20386	1.661175	0.374789
Gatos (2.4.2006)	1082.475	18.13241	18.39107	1.664706	0.950808
Vonikakis (2.5.2008)	1116.98	17.86367	18.08074	1.717771	0.416447
Otsu (1.1.1979)	1136.256	17.82513	18.03767	1.747414	0.653088
Fuz. C-means (1.3.1973)	1143.395	17.85714	18.04058	1.758393	0.521405
Bernsen (2.2.1986)	1148.187	17.75559	17.96667	1.765763	0.443693
Ramesh (1.2.1995)	1317.732	17.48979	17.65106	2.026501	1.453468
Palumbo (2.4.1986)	1388.759	16.88965	17.10351	2.135731	0.461838
Kohon. SOM (2.1.2000)	1479.509	17.33808	17.57842	2.275293	11.58626
Sauvola (2.2.2000)	1493.785	16.5649	16.80586	2.297247	0.77009
IGT (3.1.2005)	1592.008	16.22354	16.31476	2.448303	0.435930
Black Percen. (1.2.1962)	1626.66	15.93992	16.15941	2.501591	0.354353
Brink (2.3.1996)	1956.728	16.05018	16.15053	3.009194	3.234369
Kapur (2.3.1985)	1958.988	15.41409	15.69104	3.012669	1.64126
IIFA(1.1.1995)	2043.185	15.25285	15.58478	3.142154	1.827724
Yen (1.4.1995)	2080.253	15.2717	15.55781	3.199158	4.127165
Histog. peaks (1.2.1966)	2184.715	15.7486	15.91618	3.359808	15.59393
Abutaleb (2.3.1989)	4079.849	11.6206	12.08744	6.274278	1.246441
Parker (2.4.1991)	8455.937	8.187518	8.912703	13.00413	4.279259
K-means (1.3.1988)	9069.963	14.99826	15.19859	13.94842	992.2235
Kittler (1.5.1985)	14453.05	8.047554	9.957478	22.22692	639.8624
Niblack (2.2.1986)	15780.57	4.806632	6.192451	24.26846	12.07129
Riddler (1.5.1978)	15970.74	6.585206	8.091815	24.56092	213.2842
Rosenf.Kak (1.2.1982)	18277.45	3.958347	5.613259	28.10834	36.88286
Lloyd (1.5.1985)	19626.18	3.494714	5.314108	30.18251	46.62763
Mardia (2.5.1988)	19973.03	3.291748	5.205405	30.71592	34.67375
Pun (1.4.1980)	27847.65	0.950004	3.697339	42.82607	12.31683

Table 4: The evaluation metrics for ave-int technique.

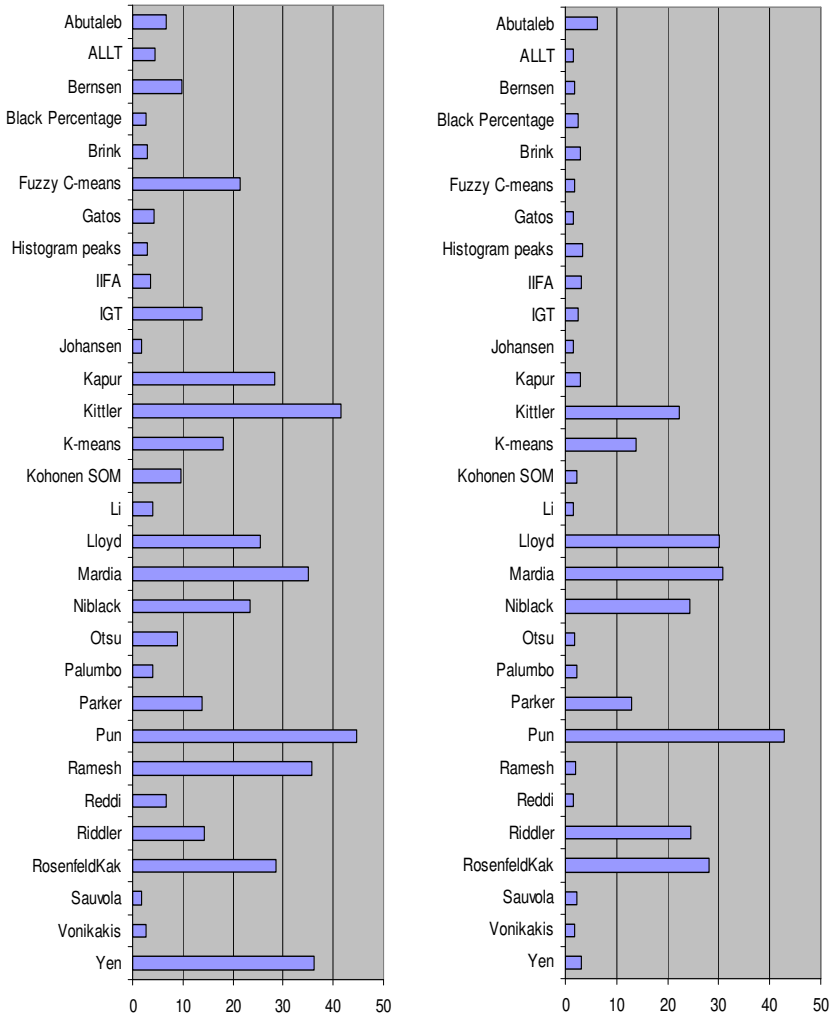


Figure 4: The mean PERR of each algorithm on (a) max-int set (b) ave-int set.

By comparing the tables 3 and 4 and looking carefully at the figure 4 where the algorithm performance is given in alphabetic order for the two experimental sets, we can make some remarks:

1) If we accept that the mean PERR gives a good estimation of the final image, in accordance with MSE, SNR and PSNR, the variation of the PERR gives a good indication of the algorithm stability. Thus, there are very stable methods in one (e.g Sauvola) or both (e.g Black Percentage) cases and others very unstable in both cases (e.g K-means).
2) The majority of the algorithms (21 out of 30) have a better performance on the test of ave-int technique. This way we can distinguish the methods in those that perform considerably better when there is clear outstanding of the

foreground (e.g Lloyd) and others in the opposite case (e.g Otsu). Moreover, there are some algorithms with good and very stable performance in every case (e.g. Vonikakis et al.).

3) Although in average local binarization methods perform slightly better than the global ones, there is a large variance. Hence, some global methods have a very good performance and some local ones are close to the worst case.
4) Among the global methods with very good performance, the ones based on histograms or classification techniques presented better results than the other global ones.
5) There is no obvious dependence of the algorithm performance on how recent the algorithm is. It is remarkable that the oldest algorithm we tested (black percentage), a simple and very old algorithm was the fourth better in the case of max-int.
6) It is quite surprising that algorithms specifically designed for applications of historical document images (IGT, Gatos) didn't perform better than those of general purpose.

In tables 5 and 6, the best algorithms for each *doc* or *noisy* image, respectively, on both experimental sets are given.

image	max-int	ave-int
doc_1	Sauvola	Gatos
doc_2	Sauvola	Johansen
doc_3	Sauvola	Gatos
doc_4	Sauvola	Johansen
doc_5	Sauvola	Gatos
doc_6	Sauvola	Gatos
doc_7	Sauvola	Johansen
doc_8	Sauvola	Johansen
doc_9	Sauvola	Johansen
doc_10	Sauvola	Johansen

Table 5: Best algorithm for each document image.

Examining the output images in more detail and taking into account the descriptions of tables 1 and 2, we realized that about half of the methods were giving their best results for doc_7 and doc_10. In some cases, doc_7 was first and doc_10 second and in other cases the opposite. In those methods, the worst cases were the doc_1 and doc_3 with variance in the order again. On the contrary, there was no obvious dependency on the noisy images neither on its size. Moreover, the noise_2 and noise_3 images that are very similar, in the majority of the algorithms, with very few exceptions were given very similar results and in many cases exactly the same PERR. However, we consider the remarks of this paragraph very preliminary and need further analysis.

image	max-int	ave-int
noise_1	Sauvola	Gatos
noise_2	Gatos	Gatos
noise_3	Gatos	Gatos
noise_4	Sauvola	Johansen
noise_5	Sauvola	Gatos
noise_6	Ramesh	Gatos
noise_7	Sauvola	Johansen
noise_8	Sauvola	ALLT
noise_9	Sauvola	Gatos
noise_10	Sauvola	Li
noise_11	Sauvola	Reddi
noise_12	Ramesh	Johansen
noise_13	Sauvola	Gatos
noise_14	Sauvola	Gatos
noise_15	Sauvola	Gatos

Table 6: Best algorithm for each noisy image.

5 Conclusion

A technique was proposed for the evaluation of binarization algorithms. This technique is appropriate for document images that are difficult to be evaluated by techniques based on segmentation or recognition of the text. In order to demonstrate the proposed method we tested 30 existing binarization algorithms of general and special purpose. The proposed methodology was presented on historical documents. We performed experiments on two document sets made by using two different techniques of image mosaicing and combining old blank document pages that include all the common problems of historical documents with noise-free pdf documents. This way, after the application of the binarization algorithms to the synthetic images, it is easy to evaluate the results by comparing the resulted image with the original document.

Although there is a slightly better performance of the local binarization methods vs. the global ones, the global ones based on histograms or classification techniques presented almost as good results as the local ones. There is no obvious dependence of the algorithm performance on how recent the algorithm is and novel algorithms, specialized on historical document images didn't perform better than those of general purpose.

Our future plan is to conduct more experiments in order to examine the binarization procedure with more algorithms and specific applications.

References

[Abutaleb 1989] Abutaleb, S.: “Automatic thresholding of gray-level pictures using two-dimensional entropy”; Comput. Vis. Graph. Image Process. 47, 22–32, 1989.

[Bernsen 1986] Bernsen, J.: “Dynamic thresholding of grey-level images”, Proc. 8th ICPR, pp 1251-1255, 1986.

[Brink and Pendock 1996] Brink, A.D., Pendock, N.E.: “Minimum Cross-Entropy Threshold Selection”; PR(29), pp. 179-188, 1996.

[Doyle 1962] Doyle, W.: “Operation useful for similarity-invariant pattern recognition”; J. Assoc. Comput. Mach, vol. 9, pp. 259-267, 1962.

[Duda and Hart 1973] Duda, R,and Hart, P.: “Pattern Classification and Scene Analysis”; Wiley, New York 1973.

[Gatos et al 2006] Gatos, B., Pratikakis, I.E., Perantonis, S.J.: “Adaptive degraded document image binarization”; Pattern Recognition (39), No. 3, pp. 317-327, 2006.

[Glasbey 1993] Glasbey, C. A.: “An analysis of histogram-based thresholding algorithms”; Graph. Models Image Processing 55, 532–537, 1993.

[Gottesfeld Brown 1992] Gottesfeld Brown, L.: “A survey of Image Registration Techniques”; ACM Computing Surveys, Vol 24, No 4,325-376, 1992.

[He et al 2005] He, J., Do, Q.D.M., Downton, A.C., Kim, J.H.: “A Comparison of Binarization Methods for Historical Archive Documents”; Proc. 8th ICDAR, pp. 538-542, 2005.

[Jain and Dubes 1988] Jain, A. K., and Dubes, R. C.: "Algorithms for Clustering Data"; Prentice Hall, 1988.

[Johannsen and Bille 1982] Johannsen, G., and Bille, J.: "A threshold selection method using information measures"; Proc. Sixth Intern. Conf. Pattern Recognition, pp. 140-143, Munich, Germany, 1982.

[Kapur et al 1985] Kapur, N., Sahoo, P.K,.and Wong, A.K.: "A new method for gray-level picture Thresholding using the Entropy of the histogram"; Computer Vision Graphics and Image Processing 29, 273-285, 1985.

[Kavallieratou 2005] Kavallieratou, E.: “A Binarization Algorithm Specialized on Document Images and Photos”; 8th Int. Conf. on Document Analysis and Recognition, 2005, pp.463-467.

[Kavallieratou 2008] Kavallieratou, E.: “An objective way to evaluate and compare binarization algorithms”; SAC’08, March 16-20, 2008, Fortaleza, Ceará, Brazil.

[Kite et al 2000] Kite, T.D., Evans, B.L., Daamera-Venkata, N., and Bovil, A.C.: “Image Quality Assesment Based on a Degradation Model”; in IEEE Trans. Image Processing, vol.9, pp.909-922, 2000.

[Kittler and Illingworth 1985] Kittler, J., Illingworth, J.: “On threshold selection using clustering criteria”; IEEE Trans. Systems Man Cybernet.15, 652–655, 1985.

[Lee et al 1990] Lee, S. U., Chung, S. Y., and Park, R. H.: “A comparative performance study of several global thresholding techniques for segmentation”; Graph. Models Image Process. 52, 171–190 , 1990.

[Leedham et al 2002] Leedham, G., Varma, S., Patankar, A., Govindaraju, V.:"Separating Text and Background in Degraded Document Images"; Proc. 8th IWFHR, pp. 244-249, September, 2002.

[Lins and da Silva 2007] Lins, R. D.. da Silva, J. M. M.: "A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents"; SAC'07, 2007, Seoul, Korea, pp.610-616.

[Liu and Li 1993] Liu, J. Z., and Li, W. Q.: "The automatic thresholding of gray-level pictures via two-dimensional Otsu method"; Acta Automatica Sin. 19, 101–105, 1993.

[Lloyd 1985] Lloyd, D. E.: "Automatic target classification using moment invariant of image shapes"; Technical Report, RAE IDN AW126, Farnborough, UK, Dec. 1985.

[Mardia and Hainsworth 1988] Mardia, K.V., and Hainsworth, T.J.: "A Spatial Thresholding Method for Image Segmentation"; IEEE Trans. Pattern Analysis Machine Intelligence, 10, pp. 919-927, 1988.

[Niblack 1986] Niblack, W.: "An Introduction to Digital image processing"; Prentice Hall, pp 115-116, 1986.

[Otsu 1979] Otsu, N.: "A threshold selection method from gray-level histograms"; IEEE Trans. Systems Man Cybernet., 9 (1), pp. 62-66, 1979.

[Palumbo et al 1986] Palumbo, P. W., Swaminathan, P., and Srihari, S. N.: "Document image binarization: Evaluation of algorithms"; Proc. SPIE 697, 278–286, 1986.

[Palumbo et al 1986] Palumbo, P., Swaninathan, P., and Shrihari, S.: "Document Image Binarization: Evaluation of Algorithms"; SPIE Applications of Digital Image Processing IX, vol. 697, pp. 278-285, 1986.

[Papamarkos and Atsalakis 2000] Papamarkos N., and Atsalakis, A.: "Gray-level reduction using local spatial features"; Computer Vision and Image Understanding, pp. 336-350, 2000.

[Parker 1991] Parker, J.R.: "Gray level thresholding in badly illuminated images"; IEEE Trans. Pattern Anal. Mac. Intell. 13 (8), 813–819, 1991.

[Prewitt and Mendelsohn 1966] Prewitt, J. M. S., and Mendelsohn, M. L.: "The analysis of cell images"; in Ann. New York Acad. Sci., vol. 128, pp. 1035-1053, 1966.

[Pun 1980] Pun, T.: "A new method for gray-level picture thresholding using the entropy of the histogram"; Signal Process. 2, pp. 223–237, 1980.

[Ramesh et al 1995] Ramesh, N., Yoo, J.H., Sethi, I.K.: "Thresholding based on histogram approximation"; IEE Proc.-Vis.Image Signal Process.,Vol.142, No.5 pp: 4147, 1995.

[Reddi et al 1984] Reddi, S.S., Rudin, S.F,.and Keshavan, H.R.: "An optimal multiple Threshold scheme for image segmentation"; IEEE Tran. On System Man and Cybernetics 14 (4), 661-665, 1984.

[Ridler and Calvard 1978] Ridler, T.W., and Calvard, S.: "Picture thresholding using an iterative selection method"; IEEE Transactions on Systems, Man, and Cybernetics 8:630-632, 1978.

[Rosenfeld and Kak 1982] Rosenfeld, A., and Kak, A. C.: "Digital Picture Processing"; 2nd ed. New York: Academic, 1982.

[Sahoo et al 1988] Sahoo, P. K., Soltani, S., Wong, A. K. C., and Chen, Y.: "A survey of thresholding techniques"; Comput. Graph. Image Process. 41, 233–260, 1988.

[Sauvola and Pietikainen 2000] Sauvola, J., Pietikainen, M.: "Adaptive document image binarization "; Pattern Recognition 33, 225–236, 2000.

[Sezgin and Sankur 2004] Sezgin, M., Sankur, B.: “Survey over image thresholding techniques and quantitative performance evaluation”; Journal of Electronic Imaging 13(1), 146–165, 2004.

[Trier and Taxt 1995] Trier, O.D., and Taxt, T.: “Improvement of ‘integrated function algorithm’ for binarisation of document images”; Pattern Recognition Letters, 16, pp. 277–283, 1995.

[Vonikakis et al 2008] Vonikakis, V., Andreadis, I., and Papamarkos, N.: "Robust Document Binarization with OFF Center-surround Cells"; Pattern Analysis & Applications, to appear.

[Weszka and Rosenfeld 1978] Weszka, J. S., and Rosenfeld, A.,: “Threshold evaluation techniques”; IEEE Trans. Syst. Man Cybern. SMC-8, 627–629, 1978.

[Yang and Yan 2000] Yang, Y., and Yan, H.: “An adaptive logical method for binarization of degraded document images”; Pattern Recognit, 33, pp. 787–807, 2000.

[Yen 1995] Yen, J.C., Chang, F.J., and Chang, S.: “A New Criterion for Automatic Multilevel Thresholding”; IP(4), No. 3, March pp. 370-378, 1995.