

ESTIMATION OF PROPER PARAMETER VALUES FOR DOCUMENT BINARIZATION

E. Badekas and N. Papamarkos

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

ABSTRACT

Most of the existing document-binarization techniques deal with many parameters that require a priori setting of their values. Due to the unknown of the ground-truth images, the evaluation of document binarization techniques is subjective and employs human observers for the estimation of the appropriate parameter values. The selection of the appropriate values for these parameters is crucial and influences to the final binarization. However, there is no predetermined set of parameters that guarantees optimal binarization for all document images. This paper proposes a new technique that allows the estimation of proper parameters values for each one of the document binarization techniques. The proposed approach is based on a statistical performance analysis of a set of binarization results, which are obtained by applying various binarization techniques with different parameter values. The proposed statistical performance analysis can also depict the best document binarization result obtained by a set of document binarization techniques.

KEY WORDS

Binarization, Thresholding, Document Processing, Segmentation, Parameter Evaluation and Detection

1. Introduction

Document binarization is an active research area in image processing. Many binarization techniques for gray-scale, and more recently, for color document images have been proposed [1-23]. Most of these techniques include parameters, which require appropriate initial settings of their values. It is obvious that the selected values of the parameter set (PS) may be appropriate to specific document images and possibly to other similar documents. Thus, estimation of appropriate PS values will be necessitated again for distinctive images. This paper proposes a new technique, which can be used to adaptive estimation of appropriate PS values for various document binarization techniques.

In general, both global [1-5] and local [6-14] binarization techniques exist for document binarization. The global binarization techniques are suitable for converting any gray-scale image into a binary form but

are inappropriate for complex document images, and perform even worse with degraded document images. In cases of complex and degraded documents, local binarization techniques give better binarization results. This category includes the techniques of Bernsen [6], Chow and Kaneko [7], Eikvil [8], Mardia and Hainsworth [9], Niblack [10], Taxt [11], Yanowitz and Bruckstein [12], Sauvola and Pietikainen [13-14] and Gatos et al. [31]. In the context of document binarization, the most powerful techniques are probably those that go beyond the image gray-scale values to incorporate the structural characteristics of the characters [15-18], [29]. Methods that are based on stroke analysis, such as the stroke width (SW) and other characters' geometrical characteristics belong to this category. The Adaptive Logical Level Technique (ALLT) and its improved versions [15-16], [29] as well as the Improvement of Integrated Function Algorithm (IIFA) [17-18], [29] are two of the most powerful techniques that utilize these characteristics. Finally, there are binarization techniques that are based on generic clustering approaches, such as the Fuzzy C-means algorithm (FCM) [19] and the Kohonen neural network based techniques, proposed by Papamarkos et al. [20-23].

Most of the binarization techniques, especially those categorized as local thresholding algorithms have PS values that must be defined prior to their application to some document image. Clearly, different values of the PS lead to different binarization results, which mean that there is not a safe set of appropriate PS values for all types of document images.

In this paper, a Parameter Estimation Algorithm (PEA) is proposed which can be used to detect the proper PS values of every document binarization technique. The estimation is based on the correlation analysis between the different document binarization results obtained by applying some specific binarization technique to a document image, using different PS values. The proposed method is based on the work of Yitzhaky and Peli [28], which is used for edge detection evaluation. In their approach, a specific range and a specific step for each one of the parameters are initially defined. The proper values for the PS are then estimated by comparing the results obtained by all possible combinations of the PS values. The proper PS values are estimated using a Receiver

Operating Characteristics (ROC) analysis and a Chi-square test. In order to improve this algorithm, we use a wide initial range for every parameter and apply an adaptive convergence procedure, in order to estimate the proper parameter values. Specifically, at each iteration of the adaptive procedure, the ranges of the parameters are redefined according to the estimation of the best and the second best binarization result obtained. The adaptive procedure terminates when the ranges of the parameter values cannot be reduced further and the proper PS values are those obtained from the last iteration.

The proposed technique was extensively tested using a variety of documents most of which come from the old Greek Parliamentary Proceedings and from the University of Washington database [30]. Binarization results of independent techniques are compared with the proposed evaluation technique and according to their performance in an OCR application. The mean rating values calculated for each binarization technique, from the proposed technique and the human-assessment experiment, are similar with a variation of ± 0.5 . All the experiments presented, confirm the effectiveness of the proposed technique. Sauvola and Pietikainen's technique seems to work properly, in most of the cases, with the specific document database.

2. Obtaining the best Binarization Result

For the body of your document, use Times New Roman font, 10-point type size, single-spaced. The whole document should be fully justified (not only left-justified). Headings should be 12-point, upper- and lower-case, bold or 10pt upper case, bold. Subheadings should be 10-point upper- and lower-case.

When a document image is converted to a binary form, the ideal result is not known a priori. This is a major problem in comparative evaluation tests. In order to have comparative results, it is important to estimate a ground truth image and consequently using this image as a reference. Doing this, we can compare the different binarization results obtained, and therefore, we can estimate the best of those. This ground truth image, known as Estimated Ground Truth (EGT) image, is selected from a list of Potential Ground Truth (PGT) images as proposed by Yitzhaky and Peli [28] for edge detection evaluation.

Consider N document binary images D_j , $j=1, \dots, N$ obtained by applying one or more document binarization techniques to a gray-scale document image of size $K \times L$. In order to get the best binary image it is necessary to obtain the EGT image first. In turn, the independent binarization results are compared with the EGT image using the Chi-square test.

The entire procedure is described below, the background and foreground pixels are represented by "0" and "1", respectively.

Stage 1 For every pixel, the number of binary images that consider it as foreground pixel is calculated. The

results are stored to a matrix $C(x, y)$, $x=0, \dots, K-1$ and $y=0, \dots, L-1$. It is obvious that the values of this matrix will be between 0 and N .

Stage 2 N PGT_i , $i=1, \dots, N$ binary images are produced using the matrix $C(x, y)$. Every PGT_i image is considered as the image that has as foreground pixels all the pixels with $C(x, y) \geq i$.

Stage 3 PGT_{i_0} and PGT_{i_1} are defined respectively as the background and foreground pixels in PGT_i image, while D_{j_0} and D_{j_1} represent the background and foreground pixels in D_j image. For each PGT_i image, four probabilities are defined:

- Probability that a pixel is a foreground pixel in both PGT_i and D_j images:

$$TP_{PGT_i, D_j} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L PGT_{i_1} \cap D_{j_1} \quad (1)$$

- Probability that a pixel is a foreground pixel in PGT_i image and background pixel in D_j image:

$$FP_{PGT_i, D_j} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L PGT_{i_1} \cap D_{j_0} \quad (2)$$

- Probability that a pixel is a background pixel in both PGT_i and D_j images:

$$TN_{PGT_i, D_j} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L PGT_{i_0} \cap D_{j_0} \quad (3)$$

- Probability that a pixel is a background pixel in PGT_i image and foreground pixel in D_j image:

$$FN_{PGT_i, D_j} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L PGT_{i_0} \cap D_{j_1} \quad (4)$$

According to the above definitions, for each PGT_i the average value of the four probabilities resulting from its match with each of the individual binarization results D_j , is calculated:

$$TP_{PGT_i} = \frac{1}{N} \sum_{j=1}^N TP_{PGT_i, D_j} \quad (5)$$

$$FP_{PGT_i} = \frac{1}{N} \sum_{j=1}^N FP_{PGT_i, D_j} \quad (6)$$

$$TN_{PGT_i} = \frac{1}{N} \sum_{j=1}^N TN_{PGT_i, D_j} \quad (7)$$

$$FN_{PGT_i} = \frac{1}{N} \sum_{j=1}^N FN_{PGT_i, D_j} \quad (8)$$

Stage 4 In this stage, the sensitivity TPR_{PGT_i} and specificity $(1 - FPR_{PGT_i})$ values are calculated according to the relations:

$$TPR_{PGT_i} = \frac{TP_{PGT_i}}{P} \quad (9)$$

$$FPR_{PGT_i} = \frac{FP_{PGT_i}}{1-P} \quad (10)$$

where $P = TP_{PGT_i} + FN_{PGT_i}$, $\forall i$

Stage 5 This stage is used to obtain the EGT image. The EGT image is selected to be one of the PGT_i images. For each PGT_i , the $X_{PGT_i}^2$ value is calculated, according to the relation:

$$X_{PGT_i}^2 = \frac{(\text{sensitivity} - Q_{PGT_i})(\text{specificity} - (1 - Q_{PGT_i}))}{(1 - Q_{PGT_i}) \cdot Q_{PGT_i}} \quad (11)$$

where $Q_{PGT_i} = TP_{PGT_i} + FP_{PGT_i}$. A histogram from the values of $X_{PGT_i}^2$ is constructed (CT-Chi-square histogram). The best CT will be the value of i that maximizes $X_{PGT_i}^2$. The PGT_i image in this CT level will be then considered as the EGT image. An example of a CT Chi-square histogram is shown in Figure 1 for $N = 9$. The detected CT level in this example is the fifth.

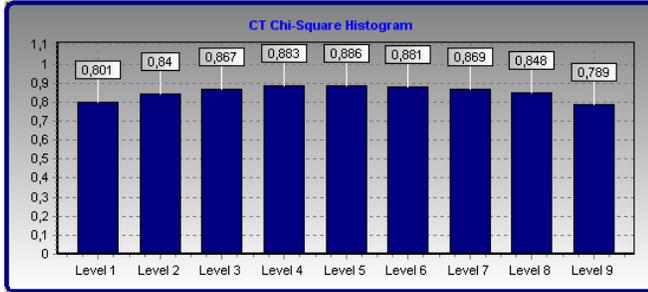


Figure 1. An example of a CT Chi-square histogram. The fifth level is the CT level.

Stage 6 For each image D_j , four probabilities are defined:

- Probability that a pixel is a foreground pixel in both D_j and EGT images:

$$TP_{D_j,EGT} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L D_{j_k} \cap EGT_l \quad (12)$$

- Probability that a pixel is a foreground pixel in D_j image and background pixel in EGT image:

$$FP_{D_j,EGT} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L D_{j_k} \cap EGT_0 \quad (13)$$

- Probability that a pixel is a background pixel in both D_j and EGT images:

$$TN_{D_j,EGT} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L D_{j_0} \cap EGT_0 \quad (14)$$

- Probability that a pixel is a background pixel in D_j image and foreground pixel in EGT image:

$$FN_{D_j,EGT} = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=1}^L D_{j_0} \cap EGT_l \quad (15)$$

Stage 7 Stages 4 and 5 are repeated to compare each binary image D_j with the EGT image, using the relations (12)-(15) rather than the relations (5)-(8) which calculated in Stage 3. According to the Chi-square test, the maximum value of $X_{D_j,EGT}^2$ indicates the D_j image which is the estimated best document binarization result. Sorting the values of the Chi-square histogram, the binarization results are sorted according to their quality.

3. Parameter Estimation Algorithm

In the first stage of the proposed evaluation system it is necessary to estimate the proper PS values for each one of the independent document binarization techniques. This estimation is based on the method of Yitzhaky and Peli [28] proposed for edge detection evaluation. However, in order to increase the accuracy of the estimated proper PS values, we improve this algorithm by using a wide initial range for every parameter and an adaptive convergence procedure. That is, the ranges of the parameters are re-determined in line with the estimation of the best and second best binarization result obtained in each iteration of the adaptive procedure. This procedure terminates when the ranges of the parameters values cannot be further reduced and the proper PS values are those obtained during the last iteration. It is important to note that this is an adaptive procedure and is applicable to every document image.

The stages of the proposed parameter estimation algorithm, for two parameters (P_1, P_2), are as follows:

Stage 1 Define the initial range of the PS values. Consider as $[s_1, e_1]$ the range for the first parameter and $[s_2, e_2]$ the range for the second one.

Stage 2 Define the number of steps that will be used in each iteration. For the two parameters case, let St_1 and St_2 be the numbers of steps for the ranges $[s_1, e_1]$ and $[s_2, e_2]$, respectively. In most cases $St_1 = St_2 = 3$.

Stage 3 Calculate the lengths L_1 and L_2 of each step, according to the following relations:

$$L_1 = \frac{e_1 - s_1}{St_1 - 1}, \quad L_2 = \frac{e_2 - s_2}{St_2 - 1} \quad (16)$$

Stage 4 In each step, the values of parameters P_1, P_2 are updated according to the relations:

$$P_1(i) = s_1 + i \cdot L_1, \quad i = 0, \dots, St_1 - 1 \quad (17)$$

$$P_2(i) = s_2 + i \cdot L_2, \quad i = 0, \dots, St_2 - 1 \quad (18)$$

Stage 5 Apply the binarization technique to the document image using all the possible combinations of (P_1, P_2) . Thus, N binary images $D_j, j = 1, \dots, N$ are produced, where N is equal to $N = St_1 \cdot St_2$.

Stage 6 Examine the N binary document results, using the algorithm described in Section 2, to estimate the best and the second best document binarization results. Let (P_{1B}, P_{2B}) and (P_{1S}, P_{2S}) be the parameters' values

obtained from the best and the second best binarization results, respectively.

Stage 7 Redefine the ranges for the two parameters as $[s'_1, e'_1]$ and $[s'_2, e'_2]$ that will be used during the next iteration of the method, according to the relations:

$$[s'_1, e'_1] = \begin{cases} \text{If } P_{1B} \neq P_{1S} \text{ then } \begin{cases} \text{If } P_{1B} > P_{1S} \text{ then } [s'_1, e'_1] = [P_{1S}, P_{1B}] \\ \text{If } P_{1B} < P_{1S} \text{ then } [s'_1, e'_1] = [P_{1B}, P_{1S}] \end{cases} \\ \text{If } P_{1B} = P_{1S} = A \text{ then } [s'_1, e'_1] = \left[\frac{s_1 + A}{2}, \frac{e_1 + A}{2} \right] \end{cases}$$

$$[s'_2, e'_2] = \begin{cases} \text{If } P_{2B} \neq P_{2S} \text{ then } \begin{cases} \text{If } P_{2B} > P_{2S} \text{ then } [s'_2, e'_2] = [P_{2S}, P_{2B}] \\ \text{If } P_{2B} < P_{2S} \text{ then } [s'_2, e'_2] = [P_{2B}, P_{2S}] \end{cases} \\ \text{If } P_{2B} = P_{2S} = A \text{ then } [s'_2, e'_2] = \left[\frac{s_2 + A}{2}, \frac{e_2 + A}{2} \right] \end{cases}$$

Stage 8 Adjust the steps St'_1, St'_2 with the ranges that will be used in the next iteration according to the relations:

$$St'_1 = \begin{cases} \text{If } e'_1 - s'_1 < St_1 \text{ then } St'_1 = St_1 - 1 \\ \text{else } St'_1 = St_1 \end{cases}$$

$$St'_2 = \begin{cases} \text{If } e'_2 - s'_2 < St_2 \text{ then } St'_2 = St_2 - 1 \\ \text{else } St'_2 = St_2 \end{cases}$$

Stage 9 If $St'_1 \cdot St'_2 \geq 3$ go to Stage 3 and repeat all the stages. The iterations terminate when the calculated new steps for the next iteration have a product less to 3 ($St'_1 \cdot St'_2 < 3$). The proper PS values are those estimated during the Stage 6 of the last iteration.

4. Comparing the results of different binarization techniques

The proposed evaluation technique can be extended to estimate the best binarization results by comparing the binary images obtained by independent techniques. The algorithm described in Section 2 can be used to compare the binarization results obtained by the application of independent document binarization techniques. Specifically, the best document binarization results obtained from the independent techniques using the estimated proper PS values are compared through the procedure described in Section 2. That is, the final best document binarization result is obtained as follows:

Stage 1 Estimate the proper PS values for each document binarization technique, using the PEA described in Section 3.

Stage 2 Obtain the document binarization results from each one of the independent binarization techniques by using their proper PS values.

Stage 3 Compare the binary images obtained in Stage 2 and estimate the final best document binarization result by using the algorithm described in Section 2.

5. The binarization techniques incorporated in the evaluation system

In order to achieve satisfactory document binarization results, a number of powerful binarization techniques are included in the proposed evaluation system. Two of them are global, three are local and two are based on structural characteristics of the characters. In particular, the incorporated binarization techniques are:

1. Otsu's technique [1]
2. Fuzzy C-Mean (FCM) [19]
3. Niblack's technique [10]
4. Sauvola and Pietikainen's technique [13-14]
5. Bernsen's technique [6]
6. Adaptive Logical Level Technique (ALLT) [15-16], [29]
7. Improvement of Integrated Function Algorithm (IIFA) [17-18], [29].

6. Experimental Results

Experiment 1

This experiment demonstrates the application of the proposed technique to a large number of document images obtained from the old Greek Parliamentary Proceedings and the University of Washington database [30]. The goal of this experiment is to evaluate the seven independent binarization techniques and to decide which of them gives the best results. For each document image, the best binarization result of each independent binarization technique is obtained. These results are rated and sorted according to their chi-square test values obtained by the proposed evaluation method. The rating value for a document binarization technique can be between 1 (best) and 7 (worst). The mean rating value for each binarization technique is then calculated and the histogram shown in Figure 2 is constructed using these values.

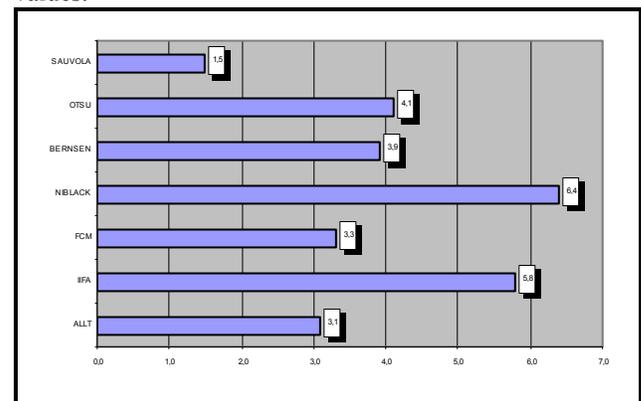


Figure 2. The histogram constructed by the mean rating values.

It is obvious that the minimum value of this histogram is assigned to the binarization technique which has the best performance for the specific document image database.

The mean chi-square values obtained for each binarization technique are presented in the histogram shown in Figure 3. According to the evaluation results it is concluded that the Sauvola and Pietikainen's technique gives, in most of the cases, the best document binarization results. This conclusion agrees with other evaluation tests such as the test performed by Sezgin and Sankur [27]. It should be noticed that the Niblack's technique was used without any post-processing step and this has as a result the technique to achieve the worst mean rating value.

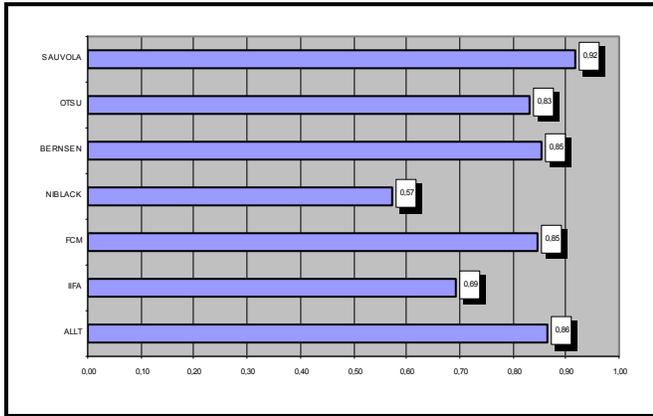


Figure 3. The histogram constructed by the mean chi-square values.

Experiment 2

In order to prove the effectiveness of the proposed evaluation technique, a human-assessment experiment has been performed in which a group of people was asked to compare the visual results obtained in Experiment 1 by the independent binarization techniques. In particular, these results were printed and handed out to 20 persons, asking them to rank the images according to their quality. The mean rating values obtained in this experiment were similar to the values obtained in the previous experiment, with a variation of ± 0.5 . The corresponding histogram constructed in this human-assessment experiment is shown in Figure 4.

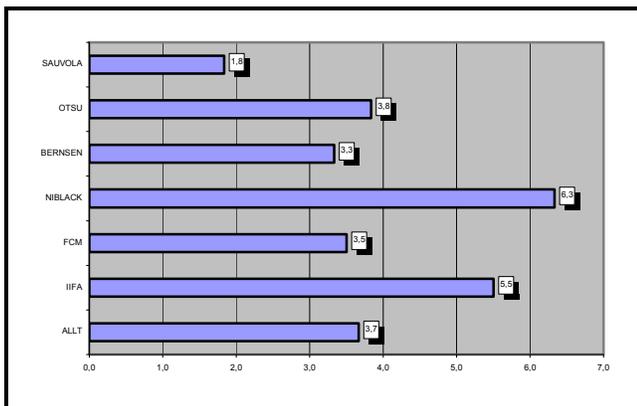


Figure 4. The histogram constructed by the mean rating values in the experiment based on human-assessment.

7. Conclusions

In this paper a new method is proposed for the estimation of the proper PS values of document binarization techniques and the best binarization result obtained by a set of independent document binarization techniques. The proposed method is extended to produce an evaluation system for independent document binarization techniques. The estimation of the proper PS values is achieved by applying an adaptive convergence procedure starting from a wide initial range for every parameter. The entire system was extensively tested with a variety of document images. The mean rating values obtained for each binarization technique in the human-assessment experiment, are similar with the mean rating values obtained by the proposed evaluation technique. Sauvola and Pietikainen's technique gives, in most of the cases, the best document binarization results.

Acknowledgements

This work is co-funded by the European Social Fund and National Resources-(EPEAEK-II) ARCHIMIDES 1, TEI Serron

References

- [1] N. Otsu, A thresholding selection method from gray-level histogram, IEEE Trans. Systems Man Cybernet. SMC-8, 1978, 62-66.
- [2] J. Kittler and J. Illingworth, Minimum error thresholding, Pattern Recognition 19 (1), 1986, 41-47.
- [3] S.S. Reddi, S.F. Rudin and H.R. Keshavan, An optimal multiple Threshold scheme for image segmentation, IEEE Tran. On System Man and Cybernetics 14 (4), 1984, 661-665.
- [4] J.N. Kapur, P.K. Sahoo and A.K. Wong, A new method for gray-level picture thresholding using the Entropy of the histogram, Computer Vision Graphics and Image Processing 29, 1985, 273-285.
- [5] N. Papamarkos and B. Gatos, A new approach for multithreshold selection, Computer Vision Graphics and Image Processing 56 (5), 1994, 357-370.
- [6] J. Bernsen, Dynamic thresholding of grey-level images, Proc. Eighth Int. Conf. Pattern Recognition, Paris, 1986, 1251-1255.
- [7] C.K. Chow and T. Kaneko, Automatic detection of the left ventricle from cineangiograms, Computers and Biomedical Research 5, 1972, 388-410.
- [8] L. Eikvil, T. Taxt and K. Moen, A fast adaptive method for Binarization of document images, Proc. ICDAR, France, 1991, 435-443.

- [9] K.V. Mardia and T.J. Hainsworth, A spatial thresholding method for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, 10 (8), 1988, 919-927.
- [10] W. Niblack, *An Introduction to Digital Image Processing*, Englewood Cliffs, N.J. Prentice Hall,, 1986, 115-116.
- [11] T. Taxt, P.J. Flynn, and A.K. Jain, Segmentation of document images, *IEEE Trans. Pattern Anal. Mach. Intell.* 11 (12), 1989, 1322-1329.
- [12] S.D. Yanowitz and A.M. Bruckstein, A new method for image segmentation, *Computer Vision, Graphics and Image Processing* 46 (1), 1989, 82-95.
- [13] J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen, Adaptive Document Binarization, *ICDAR Ulm Germany*, 1997, 147-152.
- [14] J. Sauvola and M. Pietikainen, Adaptive Document Image Binarization, *Pattern Recognition* 33, 2000, 225-236.
- [15] M. Kamel and A. Zhao, Extraction of binary character / graphics images from gray-scale document images, *CVGIP: Graphical Models Image Process.* 55 (3), 1993, 203-217.
- [16] Y. Yang and H. Yan, An adaptive logical method for binarization of degraded document images, *Pattern Recognition* 33, 2000, 787-807.
- [17] J.M. White and G.D. Rohrer, Image segmentation for optical character recognition and other applications requiring character image extraction, *IBM J. Res. Dev.* 27 (4), 1983, 400-411.
- [18] O.D. Trier and T. Taxt, Improvement of 'Integrated Function Algorithm' for binarization of document images, *Pattern Recognition Letters* 16, 1995, 277-283.
- [19] Z. Chi, H. Yan, and T. Pham, *Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition*, World Scientific Publishing, 1996.
- [20] N. Papamarkos, A neuro-fuzzy technique for document binarization, *Neural Computing & Applications*, 12 (3-4), 2003, 190-199.
- [21] N. Papamarkos, C. Strouthopoulos and I. Andreadis, "Multithresholding of color and gray-level images through a neural network technique", *Image and Vision Computing*, 18, 2000, 213-222.
- [22] N. Papamarkos and A. Atsalakis, "Gray-level reduction using local spatial features", *Computer Vision and Image Understanding*, 78, 2000, 336-350.
- [23] N. Papamarkos, A. Atsalakis and C. Strouthopoulos, "Adaptive Color Reduction", *IEEE Trans. on Systems, Man, and Cybernetics-Part B*, 32 (1), 2002, 44-56.
- [24] O.D. Trier and T. Taxt, Evaluation of binarization methods for document images, *IEEE Trans. Pattern Anal. Mach. Intelligence* 17 (3), 1995, 312-315.
- [25] O.D. Trier and A.K. Jain, Goal-Directed Evaluation of Binarization Methods, *IEEE Trans. Pattern Anal. Mach. Intelligence* 17 (12), 1995, 1191-1201.
- [26] G. Leedham, C. Yan, K. Takru and J. H. Mian, Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images, *Proc. of 7th ICDAR (2) Scotland*, 2003, 859-865.
- [27] M. Sezgin and B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, *Journal of Electronic Imaging* 13(1), 2004, 146-165.
- [28] Y. Yitzhaky and E. Peli, A Method for Objective Edge Detection Evaluation and Detector Parameter Selection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (8), 2003, 1027-1033.
- [29] E. Badekas and N. Papamarkos, A system for document binarization, *3rd International Symposium on Image and Signal Processing and Analysis ISPA 2003*, Rome, Italy
- [30] UW: English Document Image Database, University of Washington, Seattle, 1993.
- [31] B. Gatos, I. Pratikakis and S.J. Perantonis, Adaptive degraded document image binarization, *Pattern Recognition* 39, 2006, 317-327.