Automatic Evaluation of Document Binarization Results*

E. Badekas and N. Papamarkos[1]

[1]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 document binarization techniques have many parameters that can initially be specified. Usually, subjective document binarization evaluation, employs human observes for the estimation of the best parameter values of the techniques. Thus, the selection of the best values for these parameters is crucial for the final binarization result. However, there is not any set of parameters that guarantees the best binarization result for all document images. It is important, the estimation of the best values to be adaptive for each one of the processing images. This paper proposes a new method which permits the estimation of the best parameter values for each one of the document binarization techniques and also the estimation of the best document binarization result of all techniques. In this way, document binarization techniques can be compared and evaluated using, for each one of them, the best parameter values for every document image.

1 Introduction

Document binarization is an active area in image processing. Many binarization techniques have been proposed and most of them have parameters, the best values of which must initially be defined. Although, the estimation of the parameters values is a crucial stage, it is usually missed or heuristic estimated because there is no automatic parameter estimation process exists for document binarization techniques, until now.

In this paper, a Parameter Estimation Algorithm (PEA), which can be used to detect the best values for the parameter set (PS) of every document binarization technique, is proposed. The estimation is based on the analysis of the correspondence between the different document binarization results obtained by the application of a specific binarization technique to a document image, using different PS values. The proposed method is based on the work of Yitzhaky and Peli [1] which is used for edge detection evaluation. In their approach, a specific range and a specific step for each one of the parameters is initially defined. The best values for the PS are then estimated by comparing the results obtained by all possible combinations of the PS values. The best 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 in order to estimate the best parameter

* This paper was partially supported by the project Archimedes of TEI Serron.

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 1005 – 1014, 2005.

value an adaptive convergence procedure is applied. Specifically, in each iteration of the adaptive procedure, the parameters' ranges are redefined according to the estimation of the best and second best binarization result obtained. The adaptive procedure terminates when the ranges of the parameters values cannot be further reduced and the best PS values are those obtained from the last iteration.

For document binarization, it is important to lead to the best binarization result comparing the binary images obtained by a set of independent binarization techniques. For this purpose, we introduce a new technique that using the PEA leads to the evaluation of the best binarization results obtained by a set of independent binarization techniques. Specifically, for every independent binarization technique the best PS values are first estimated by using the PEA. Next, the best document binarization results obtained are compared using the Yitzhaky and Peli method and the final best binarization result is achieved.

2 Obtaining the Best Binarization Result

When we binarize a document image, we do not know initially the optimum result, that is, which is the ideal result that we must obtain. This is a major problem in comparative evaluation tests. In order to have comparative results, it is important to estimate a ground truth image. By estimating the ground truth image we can compare the different results obtained, and therefore, we can estimate the best of it. This Estimated Ground Truth (EGT) image, can be selected from a list of Potential Ground Truth (PGT) images as proposed by Yitzhaky and Peli [1].

Consider N document binary images D_j $(j = 1,...,N)$ obtained by the application of 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. After this, the independent binarization results are compared with the EGT image using the ROC analysis or a Chi-square test.

The entire procedure is described in the following where with "0" and "1" are considered the background and foreground pixels, respectively.

Stage 1 For every pixel, it is counted how many binary images consider this as foreground pixel. The results are stored to a matrix $C(x, y)$, $x = 0,..,K-1$ and $y = 0,..,L-1$. The values of the matrix will be between 0 and *N*.

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

Stage 3 For each PGT_i image, four average probabilities are defined which they assigned to pixels that are:

- Foreground in both PGT_i and D_j images:

$$TP_{PGT_i} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{K \cdot L}\sum_{k=1}^{K}\sum_{l=1}^{L} PGT_{i_1} \cap D_{j_1} \tag{1}$$

- Foreground in PGT_i image and background in D_j image:

$$FP_{PGT_i} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{K\cdot L}\sum_{k=1}^{K}\sum_{l=1}^{L} PGT_{i_1} \cap D_{j_0} \tag{2}$$

- Background in both PGT_i and D_j images:

$$TN_{PGT_i} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{K\cdot L}\sum_{k=1}^{K}\sum_{l=1}^{L} PGT_{i_0} \cap D_{j_0} \tag{3}$$

- Background in PGT_i image and foreground in D_j image:

$$FN_{PGT_i} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{K\cdot L}\sum_{k=1}^{K}\sum_{l=1}^{L} PGT_{i_0} \cap D_{j_1} \tag{4}$$

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

$$TPR_{PGT_i} = \frac{TP_{PGT_i}}{P} \tag{5}$$

$$FPR_{PGT_i} = \frac{FP_{PGT_i}}{1-P} \tag{6}$$

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

Stage 5 This stage is used to obtain the EGT image, which is selected to be one of the PGT_i images. There are two measure methods that can be used:

The ROC analysis

It is a graphical method which is using a diagram constituted of two curves (CT-ROC diagram). The first curve (the ROC curve) constituted of N points with coordinates $(TPR_{PGT_i}, FPR_{PGT_i})$ and each one of the points is assigned to a PGT_i image. The points of this curve are the correspondence levels of the diagram. A second line, which is considered as diagnosis line, is used to detect the Correspondence Threshold (CT). This line has two points with coordinates (0,1) and *(P,P)*. The PGT_i point of the ROC curve which is closest to the intersection point of the two curves is the CT level and defines which PGT_i image will be then considered as the EGT image.

The Chi-square test

For each PGT_i, the $X^2_{PGT_i}$ value is calculated, according to the relation:

$$X^2_{PGT_i} = \frac{(sensitivity - Q_{PGT_i})\cdot(specificity - (1-Q_{PGT_i}))}{(1-Q_{PGT_i})\cdot Q_{PGT_i}} \tag{7}$$

A histogram from the values of $X^2_{PGT_i}$ is constructed (CT-Chi-square histogram). The best CT will be the value of i that maximizes $X^2_{PGT_i}$. The PGT_i image in this CT level will be then considered as the EGT image. Fig.1 shows examples of a CT ROC Diagram and a CT Chi-square histogram, for $N = 9$. In both cases the CT level is equal to five.

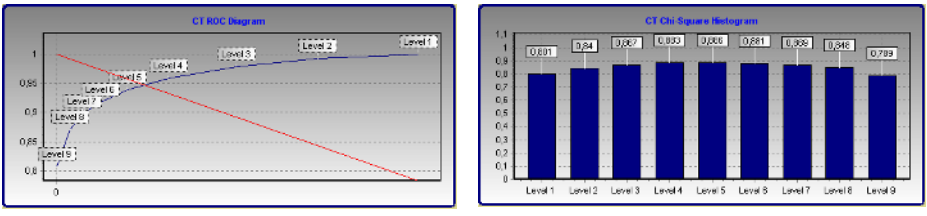

Fig. 1. A CT ROC diagram (left) and a CT Chi-square histogram (right)

Stage 6 For each D_j image, four probabilities are calculated (as in Stage 3), which they assigned to pixels that are: (a) foreground in both D_j and EGT images $TP_{D_j,EGT}$, (b) foreground in D_j image and background in EGT image $FP_{D_j,EGT}$, (c) background in both D_j and EGT images $TN_{D_j,EGT}$, (d) background in D_j image and foreground in EGT image $FN_{D_j,EGT}$.

Stage 7 Stages 4 and 5 are repeated to compare each binary image D_j with the EGT image, using the probabilities calculated in stage 6 rather than the average probabilities calculated in Stage 3. According to the Chi-square test, the maximum value of $X^2_{D_j,EGT}$ indicates the D_j image which is the estimated best document binarization result. Sorting the $X^2_{D_j,EGT}$ values, the D_j images are sorted according to their quality.

3 Parameter Estimation Algorithms

In the first stage of the proposed evaluation system it is necessary to estimate the best PS values for each one of the independent document binarization techniques. This estimation is based on the method of Yitzhaky and Peli [1] proposed for edge detection evaluation. However, in order to increase the accuracy of the estimated best PS values we improve this algorithm by using a wide initial range for every parameter and an adaptive convergence procedure. That is, the parameters' ranges are redefined according to 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 best PS values are those obtained from the last iteration. It is important to notice that this is an adaptive procedure because it is applied to every processing 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 of the cases $St_1 = St_2 = 3$.

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

$$L_1 = \frac{e_1 - s_1}{St_1 - 1}, \qquad L_2 = \frac{e_2 - s_2}{St_2 - 1} \tag{8}$$

Stage 4 In each step, the values of parameters $P_1,\ P_2$ are updated with the relations:

$$P_1(i) = s_1 + i \cdot L_1, \quad (i = 0, ..., St_1 - 1) \tag{9}$$

$$P_2(i) = s_2 + i \cdot L_2, \quad (i = 0, ..., St_2 - 1) \tag{10}$$

Stage 5 Apply the binarization technique to the processed document image using all the possible combinations of (P_1, P_2). Thus, N binary images $D_j, j = 1, ..., N$ 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:

$$[s_1^{'}, e_1^{'}] = \begin{cases} \text{if } P_{1B} \neq P_{1S} \text{ then } \begin{cases} \text{if } P_{1B} > P_{1S} \text{ then } [s_1^{'}, e_1^{'}] = [P_{1S}, P_{1B}] \\ \text{if } P_{1B} < P_{1S} \text{ then } [s_1^{'}, e_1^{'}] = [P_{1B}, P_{1S}] \end{cases} \\ \text{if } P_{1B} = P_{1S} = A \text{ then } [s_1^{'}, e_1^{'}] = [\frac{s_1 + A}{2}, \frac{e_1 + A}{2}] \end{cases} \tag{11}$$

$$[s_2^{'}, e_2^{'}] = \begin{cases} \text{if } P_{2B} \neq P_{2S} \text{ then } \begin{cases} \text{if } P_{2B} > P_{2S} \text{ then } [s_2^{'}, e_2^{'}] = [P_{2S}, P_{2B}] \\ \text{if } P_{2B} < P_{2S} \text{ then } [s_2^{'}, e_2^{'}] = [P_{2B}, P_{2S}] \end{cases} \\ \text{if } P_{2B} = P_{2S} = A \text{ then } [s_2^{'}, e_2^{'}] = [\frac{s_2 + A}{2}, \frac{e_2 + A}{2}] \end{cases} \tag{12}$$

Stage 8 Redefine the steps $St_1^{'}, St_2^{'}$ for the ranges that will be used in the next iteration according to the relations:

$$St_1^{'} = \begin{cases} \text{if } e_1^{'} - s_1^{'} < St_1 \text{ then } St_1^{'} = St_1 - 1 \\ \text{else } St_1^{'} = St_1 \end{cases} \quad (13)$$

$$St_2^{'} = \begin{cases} \text{if } e_2^{'} - s_2^{'} < St_2 \text{ then } St_2^{'} = St_2 - 1 \\ \text{else } St_2^{'} = St_2 \end{cases} \quad (14)$$

Stage 9 If $St_1^{'} \cdot St_2^{'} > 3$ 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 or equal to 3 ($St_1^{'} \cdot St_2^{'} \leq 3$). The best 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 best PS values are compared through a similar to the Section 2 procedure. That is, the final best document binarization result is obtained as follows:

Stage 1 Estimate the best 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 best 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 Experimental Results

The proposed evaluation technique is used to compare and estimate the best document binarization result produced by seven independent binarization techniques: Otsu [2], Fuzzy C-Mean (FCM) [3], Niblack [4], Sauvola and Pietaksinen's [5-6], Bernsen [7], Adaptive Logical Level Technique (ALLT) [8-9] and Improvement of Integrated Function Algorithm (IIFA) [10-11]. It should be noticed that we use improvement versions for the ALLT and IIFA, proposed by Badekas and Papamarkos [12].

Fig. 2 shows a document image coming from the old Greek Parliamentary Proceedings. For the specific image, the initial range for each parameter and the best

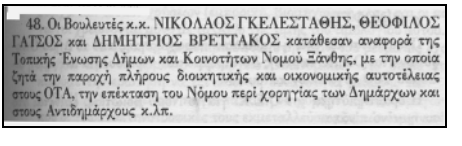
48. Οι Βουλευτές κ.κ. ΝΙΚΟΛΑΟΣ ΓΚΕΛΕΣΤΑΘΗΣ, ΘΕΟΦΙΛΟΣ ΓΑΤΣΟΣ και ΔΗΜΗΤΡΙΟΣ ΒΡΕΤΤΑΚΟΣ κατάθεσαν αναφορά της Τοπικής Ένωσης Δήμων και Κοινοτήτων Νομού Ξάνθης, με την οποία ζητά την παροχή πλήρους διοικητικής και οικονομικής αυτοτέλειας στους ΟΤΑ, την επέκταση του Νόμου περί χορηγίας των Δημάρχων και στους Αντιδημάρχους κ.λπ.

Fig. 2. Initial gray-scale document image

Table 1. The initial ranges and the estimated best *PS* values

	Technique	Initial ranges	Best *PS* values
1.	Niblack	$W \in [3,15]$, $k \in [0.2,1.2]$	$W = 14$ and $k = 0.67$
2.	Sauvola	$W \in [3,15]$, $k \in [0.1,0.6]$	$W = 14$ and $k = 0.34$
3.	Bernsen	$W \in [3,15]$, $L \in [10,90]$	$W = 14$ and $L = 72$
4.	ALLT	$a \in [0.1,0.4]$	$a = 0.10$
5.	IIFA	$T_p \in [10,90]$	$T_p = 10$

Table 2. The five iterations that applied in order to detect the best *PS* values for the binarization techniques of Niblack, Sauvola and Bernsen

Iterations	Niblack	Sauvola	Bernsen
First	1. W=3, k=0.2 2. W=3, k=0.7 3. W=3, k=1.2 4. W=9, k=0.2 5. W=9, k=0.7 (1st) 6. W=9, k=1.2 7. W=15, k=0.2 8. W=15, k=0.7 (2nd) 9. W=15, k=1.2	1. W=3, k=0.1 2. W=3, k=0.35 3. W=3, k=0.6 4. W=9, k=0.1 5. W=9, k=0.35 (1st) 6. W=9, k=0.6 7. W=15, k=0.1 8. W=15, k=0.35 (2nd) 9. W=15, k=0.6	1. W=3, L=10 2. W=3, L=50 3. W=3, L=90 4. W=9, L=10 5. W=9, L=50 (1st) 6. W=9, L=90 7. W=15, L=10 8. W=15, L=50 9. W=15, L=90 (2nd)
Second	1. W=9, k=0.45 2. W=9, k=0.7 3. W=9, k=0.95 4. W=12, k=0.45 5. W=12, k=0.7 (1st) 6. W=12, k=0.95 7. W=15, k=0.45 8. W=15, k=0.7 (2nd) 9. W=15, k=0.95	1. W=9, k=0.22 2. W=9, k=0.35 3. W=9, k=0.48 4. W=12, k=0.22 5. W=12, k=0.35 (1st) 6. W=12, k=0.48 7. W=15, k=0.22 8. W=15, k=0.35 (2nd) 9. W=15, k=0.48	1. W=9, L=50 2. W=9, L=70 3. W=9, L=90 4. W=12, L=50 5. W=12, L=70 (1st) 6. W=12, L=90 7. W=15, L=50 8. W=15, L=70 (2nd) 9. W=15, L=90
Third	1. W=12, k=0.58 2. W=12, k=0.7 3. W=12, k=0.82 4. W=14, k=0.58 5. W=14, k=0.7 (1st) 6. W=14, k=0.82 7. W=16, k=0.58 8. W=16, k=0.7 (2nd) 9. W=16, k=0.82	1. W=12, k=0.28 2. W=12, k=0.35 3. W=12, k=0.42 4. W=14, k=0.28 5. W=14, k=0.35 (1st) 6. W=14, k=0.42 7. W=16, k=0.28 8. W=16, k=0.35 (2nd) 9. W=16, k=0.42	1. W=12, L=60 2. W=12, L=70 3. W=12, L=80 4. W=14, L=60 5. W=14, L=70 (1st) 6. W=14, L=80 (2nd) 7. W=16, L=60 8. W=16, L=70 9. W=16, L=80
Fourth	1. W=14, k=0.64 (1st) 2. W=14, k=0.7 (2nd) 3. W=14, k=0.76 4. W=16, k=0.64 5. W=16, k=0.7 6. W=16, k=0.76	1. W=14, k=0.32 2. W=14, k=0.35 (2nd) 3. W=14, k=0.38 4. W=16, k=0.32 5. W=16, k=0.35 (1st) 6. W=16, k=0.38	1. W=13, L=70 2. W=13, L=75 3. W=13, L=80 4. W=14, L=70 (2nd) 5. W=14, L=75 (1st) 6. W=14, L=80
Fifth	1. W=14, k=0.64 2. W=14, k=0.67 (1st) 3. W=14, k=0.7 (2nd)	1. W=14, k=0.34 (1st) 2. W=14, k=0.36 3. W=16, k=0.34 (2nd) 4. W=16, k=0.36	1. W=14, L=70 2. W=14, L=72 (1st) 3. W=14, L=74 (2nd)

PS values obtained are given in Table 1. The best PS values for all binarization techniques are obtained using five iterations. Tables 2 and 3 give all the PS values obtained during the five iterations and also the best and second best PS values that are estimated in each iteration. The Otsu's technique has no parameters to define and FCM is used with a value of fuzzyfier *m* equal to 1.5. The results obtained by the application of the independent techniques using their best PS values, are compared using the algorithm described in Section 2. Fig.3 shows the binary images obtained by ALLT and Bernsen's technique which are estimated as the best binarization results using the Chi-square test and the ROC analysis, respectively, in order to obtain the EGT image. The corresponding diagrams for these two cases, which are constructed according to the proposed technique to compare the independent binarization techniques, are given in Fig. 4.

Table 3. The five iterations that applied in order to detect the best *PS* values for the ALLT and IIFA

Iterations	ALLT	IIFA
First	1. a=10 (1st) 2. a=25 (2nd) 3. a=40	1. Tp=10 (2nd) 2. Tp =50 (1st) 3. Tp =90
Second	1. a=10 (1st) 2. a=18 (2nd) 3. a=26	1. Tp=10 (2nd) 2. Tp =30 (1st) 3. Tp =50
Third	1. a=10 (1st) 2. a=14 (2nd) 3. a=18	1. Tp=10 (2nd) 2. Tp =20 (1st) 3. Tp =30
Fourth	1. a=10 (1st) 2. a=12 (2nd) 3. a=14	1. Tp=10 (1st) 2. Tp =15 (2nd) 3. Tp =20
Fifth	1. a=10 (1st) 2. a=11 (2nd) 3. a=12	1. Tp=10 (1st) 2. Tp =12 (2nd) 3. Tp =14

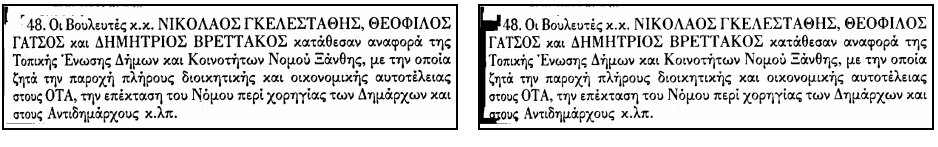

Fig. 3. Binarization result of ALLT (left) and Bernsen's technique (right)

The proposed technique is applied to a large number of document images. For each document image, the binarization results obtained, by the application of the independent binarization techniques, are sorted according to the ordering quality results 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 Fig. 5 is constructed using these values. It is obvious that the minimum value of this histogram is assigned to the binarization technique which has the best performance. The Sauvola's technique gives, in most of the cases, the best document binarization result. These conclusions agree with the evaluation test that has been made by Sezgin and Sankur [13].

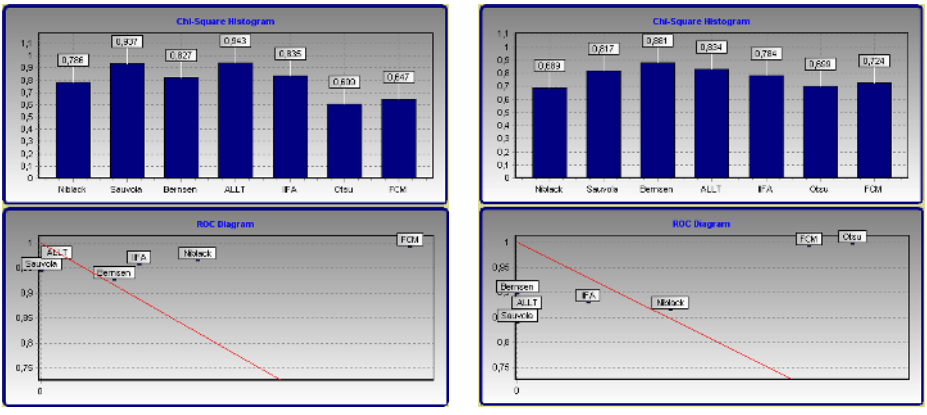

Fig. 4. The Chi-square histogram and the ROC diagram constructed using the *EGT* image calculated from the CT Chi-square histogram (left) and the CT ROC diagram (right)

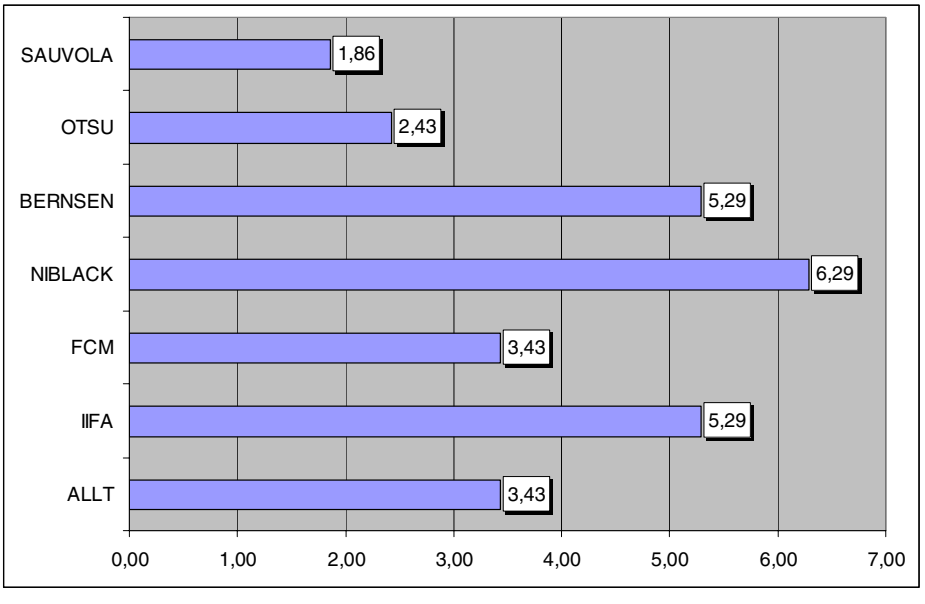

Fig. 5. The histogram constructed by the mean rating values. Sauvola's technique is the binarization technique with the best performance in the examined document image database

6 Conclusions

This paper proposes a method for the estimation of the best PS values of a document binarization technique and the best binarization result obtained by a set of independent document binarization techniques. It is important that the best PS values are adaptively estimated according to the processing document image. The proposed

method is extended to produce an evaluation system for independent document binarization techniques. The estimation of the best 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. Many of them came from standard document databases such as the old Greek Parliamentary Proceedings. The entire system is implemented in visual environment using Dephi 7.

References

1. 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.
2. N. Otsu, A thresholding selection method from gray-level histogram, IEEE Trans. Systems Man Cybernet. SMC-8 (1978) 62-66.
3. Z. Chi, H. Yan, and T. Pham, Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition, World Scientific Publishing, 1996.
4. W. Niblack, An Introduction to Digital Image Processing, Englewood Cliffs, N.J. Prentice Hall, (1986) 115-116.
5. J. Sauvola, T. Seppanen, S. Haapakoski, and M.Pietikainen, Adaptive Document Binarization, ICDAR Ulm Germany (1997) 147-152.
6. J. Sauvola and M. Pietikainen, Adaptive Document Image Binarization, Pattern Recognition 33 (2000) 225–236.
7. J. Bernsen, Dynamic thresholding of grey-level images, Proc. Eighth Int. Conf. Pattern Recognition, Paris (1986) 1251-1255.
8. 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.
9. Y. Yang and H. Yan, An adaptive logical method for binarization of degraded document images, Pattern Recognition 33 (2000) 787-807.
10. 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.
11. O.D. Trier and T. Taxt, Improvement of 'Integrated Function Algorithm' for binarization of document images, Pattern Recognition Letters 16 (1995) 277-283.
12. E. Badekas and N. Papamarkos, "A system for document binarization", 3rd International Symposium on Image and Signal Processing and Analysis ISPA 2003, Rome, Italy.
13. M. Sezgin and B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging 13(1) (2004) 146–165.