EFFICIENT & ADAPTIVE APPROACH FOR DEGRADED DOCUMENT BINARIZATION

SHIVANI GOYAL AND NARESH KUMAR GARG
Giani Zail Singh-PTU Campus/Computer Science & Engineering, Bathinda, India
er.shivanigoyal14@gmail.com, naresh2834@rediffmail.com

ABSTRACT: The digitization of historical documents in libraries is most important as they preserve the content and make the documents available through electronic media. The historical documents suffer from degradation due to damaged background, stained or damaged paper and similar other factors, so the results are not as desired. Binarization techniques can improve these recognition results. Here, the problem arises for selecting the correct threshold. The main goal or aim is to effectively binarize the document images that suffer from strain & smear, uneven background, holes & spots and various illumination effects by applying Adaptive Binarization Techniques. In this proposed work standard estimation E_{mean} divided by 4 and E_{std} multiplying by three for evaluating the results. In post processing, area value less than 15 for connected components and threshold value 0.4 is used. These parameter values are estimated after analysing the results. It has been found that each technique has its own benefits and limitations. None of the technique is best for each and every case. Accuracy value is 96.8% for printed images and 98.4% for handwritten images of DIBCO 2009 dataset. The new proposed technique gives better results as compared with previous techniques in term of the performance evaluated by using Parameters such as PSNR, Precision, Recall, F-Measure, Accuracy, Sensitivity and Specificity.

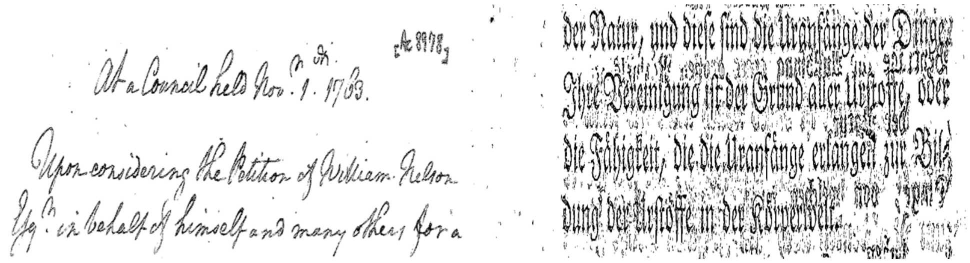
KEYWORDS: Binarization, Degraded Documents, Global & Local, Historical Documents, Illumination, Image Contrast and Threshold.

INTRODUCTION

Binarization is the preliminary step as a good binarization sets the base for further document image analysis and it refers to the conversion of the gray-scale [1] image to a binary image in document image analysis system. With the help of binarization techniques text can be differentiated from the background. The simplest way to binarize an image is to choose a threshold value. Then all pixels with values above this threshold are organized as white, while all other pixels as black. So, it is used as a text locating procedure. For binarization to be performed either global or local procedure is used for thresholding.

Historical Manuscripts

Historical manuscript [3] [4] collections are valuable resources of knowledge of the time past. There is a large collection of printed and handwritten historical documents have invaluable knowledge about the history, culture and religion. As the conservation of the material is of great concern. Therefore, only a small group of people has an access to such collection. Due to age and lack of preservation facilities these historical documents have deteriorated. The digitization of these documents can help in preserving the knowledge by storing them in multimedia format for future reference and these rare documents will be available to a large number of people.



[A 8978]

At a Council held Nov.r 1. 1763.

Upon considering the Petition of William Nelson
Esq.r in behalf of himself and many others for a

(a)

der Natur, und dieſe ſind die Uranfänge der Dinge.
Ihre Vereinigung iſt der Grund aller Urſtoffe, oder
die Fähigkeit, die die Uranfänge erlangen zur Bil-
dung der Urſtoffe in der Körperwelt.

(b)

Fig. 1. Types of İmages (a) Historical Manuscript (b) Degraded Document

İn the above fig. 1. (a) shows historical manuscript and (b) shows degraded document of images. These images are collected from Standard dataset of images DIBCO 2009 [18].

Degraded Documents

There are several reasons for the degradation of documents examples of degraded influence may include the appearance of variable background intensity caused by non-uniform intensity, as well as low contrast. A reliable result can be obtained by using Adaptive Degraded Document image binarization [11]. It is mainly performed by clearing up any needless objects appearing in the document, hiding background, removing noise and filling feasible breaks, gaps or holes in the forefront and finally, improving the quality of the character strokes before converting it to an editable text.[2] [8] [16]

Binarization Techniques

An image of up to 256 gray levels [9] is converted to a black and white image in image binarization technique. Binarization [12] is a pre-processing step that separates the pixels of the text from the background pixels. To get an image binarized, the simplest way is to choose a threshold value and organize all pixels with values higher than this as white while all other pixels as black. The proper selection of threshold is most important and is widely researched area. The threshold should be selected in such a way that most of the text information is retained by it & background is suppressed.

Binarization techniques[8] are used as a text locating techniques. Basically, there are two methods of thresholding: Global Thresholding, Local or Adaptive Thresholding [15]. The main focus of applying Binarization Techniques to digitize the historical document [10] is to suppress the background noise and to retain text information without any distortion.

Global Approach

In global approach a single threshold value is selected for whole image and is processed with this value. It mainly results in separation of foreground and background intensity. Whereas, in poor contrast, variable intensity of foreground & background. The method fails to binarize the image.

Local Approach

But with global thresholding the problem is that the changes in illumination across the image may cause some parts to be brighter i.e. in the light and some parts darker i.e. in the shadow in such a way that has nothing to do with the objects that are shown in the image. Such uneven illumination can be dealt at smallest amount in piece by determining thresholds locally. In local approach, threshold value used for each pixel or subordinate image is calculated. This moves towards the used case of historical document. The degradation such as shadows & non-uniform illumination, holes & strains are handled as they are dealt with in adaptive manner.

The rest of the paper is categorized in further different sections as follow:

- Section- II “Literature Survey” it provides overview of all the related work had been done in this area.
- Section- III “Research Methodology” includes detailed study about new proposed technique i.e. efficiently used for degraded Documents for Printed and Handwritten Documents. Here, it explains the precise algorithm
- Section- IV “Result & Discussion” gives all the aspects about results in form of figures, graphs & tables obtained and also discussion on that specified results
- Section-V “Conclusion & Future Work” includes all the details regarding future work that can be done on digitized form of images
- Acknowledgment
- References

LITERATURE SURVEY

BOLAN SU, 2013 [6] proposes that method used is easy, vigorous, involving minimum parameter modifications. Here, they achieve accuracies of 93.5%, 87.8%, and 92.03% for DIBCO-2009, 2011, 2010. It presents an adaptive image contrast based document image binarization technique that is tolerant for those types of document degradation i.e. uneven illumination and document smear. Y.H. CHIU et al. 2012 [17] proposes a two-step parameter free window based method to binarize the degraded document images. In the initial step, an incremental method is used to establish a correct window size ahead of which no significant increase in the local deviation of pixel intensities is observed. In the next step, based on the resolute window size, a noise suppressing method delivers the last binarized image by different two binarized images which are formed by two adaptive thresholding which includes the value of local mean gray and gradient. Experimental results show that the proposed technique is competitive when compared to the existing adaptive binarization methods and achieves better results in precision, accuracy and F-measure. CHUCHE FUNG 2010 [7] is performed on different binarization techniques such as Otsu, Sauvola, Ni-Black algorithm and different evaluation measurements have been taken. This technique result in noise reduction and in selection of automatic optimal binarization algorithm. K.NTIROGIANNIS, B.GTOS, 2009 [13] it improves the adaptive logical level techniques by making the window variable to extract the essential features as character stroke width (SW) as some character have different SW so there will be different SW values. It uses skeleton & the counterpoints of the local adaptive binarization output, avoiding the image block division and run length histogram. Modified threshold offers robustness and enhances performance. Proposed technique is based on local averaging it incorporates the acceleration similar to [7] for rapid performance. B. GATOS, S.J. PERANTONIS AND I. PRATIKAKIS, 2008 [5] proposes the binarization of the document by new approach which is a combination of various binarization techniques and various factors and found that it has 92.3% f-measure which is better than previous techniques, so it proves that the new adaptive technique for binarization and degraded document is better than previous techniques. NTOGAS, NIKOLAOS, VENTZAS, DIMITRIOS, 2008 [14] This proposed the comparison of different techniques of binarization (i) Bernsen (ii) Ni-Black (iii) Otsu (iv) Sauvola The proposed method is implemented in 5 steps: (i) & (ii) the image is captured and converted to jpeg or tiff format (iii) de noising of image is done by best filters after their comparisons basically mean, median, wiener filters for spatial domain and Butterworth, Gaussian filter for frequency domain (iv) binarization techniques are applied (v) refinement of the image after implementation. B.GATOS AND I.PRATIKIT, 2008 [4] In this to get the binarization of the document by new approach which is a combination of various binarization methodology and various factor (i) efficient preprocessing that has been performed by using wiener filter to remove noise and make the image smooth (ii) binarization that has been performed by applying different binarization techniques on the image such as local, global, and many more these binarization result are combined to produce binary image (iii) edge information of the gray level image was combined with the binary result of the previous step. According to that criteria and smoothing algorithm in order to fill the text area (iv) enhancement of image by mathematical morphology. Comparison has been made with any other algorithm and found that it has 91.9% f-measure which is better than previous techniques. Therefore, it proves that the new adaptive technique for binarization and degraded document is better than previous techniques. At last, this survey provides the overview of all the related work had been done by the researchers to remove degradation from the deteriorated images.

RESEARCH METHODOLOGY

Proposed Algorithm

Bolan Su [6] has implemented a technique of his known as Bolan Su by his name. The new proposed technique is implemented by modifying a few parameters for evaluating better results from previous techniques. The implemented algorithm of the new proposed technique is as follows:

Step 1: Load the images to which the proposed technique has to be applied.
Step 2: Firstly, apply the following normalized equation to construct the image contrast on degraded image

$$C_a\,(i,j) = \frac{\alpha C\,(i,j)+(1-\alpha)\big(I_{max}\,(i,j)-I_{min}(i,j)\big)}{\big(I_{max}\,(i,j)+I_{min}(i,j)\big)} \qquad (1)$$

Where, $\big(I_{max}\,(i,j) + I_{min}(i,j)\big)$
This equation is named as normalized in which 0 or 1 is considered.

$$\alpha = (\tfrac{Std}{128})^{\gamma}, \qquad \gamma=2^{-10}$$

Step 3: Now, Compute canny edge detection by using threshold value i.e. 0.4 & width is approximate 10 to 12.
***Step 4*:** After this, applied Ostu on step 2 and gets the result in Binary map.
Step 5: Canny edge detection and step 4 results values are same means 0 (Zero) then considered these values otherwise, discard the values.
Step 6: Now, Combined Binary map resulted after step 5 in which, if values founds as Zero (edge) and precede original image intensity otherwise, if not zero then consider as 1 (background). The new image is constructed with edge intensity known as E.
Step 7: Compute Edge mean as $\boldsymbol{Emean}$ and Edge Standard Deviation $\boldsymbol{Estd}$ from edge image E.
Step 8: and then compute the Local threshold by using below given formula,

$$S(x,y) = \left\{\begin{array}{c|c} 1 & I(x,y) \le \frac{Emean}{4} + Estd * 3 \\ 0 & otherwise \end{array}\right\} \qquad (2)$$

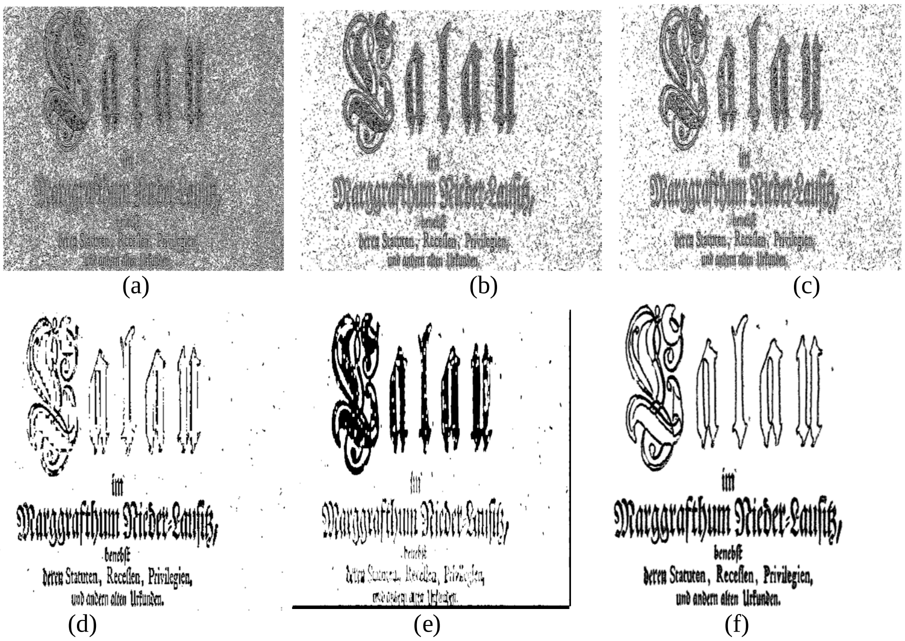
Step 9: At last, Applied Post Processing technique for removing smaller objects area which is less than 15 is used.

Post Processing is used by finding connected components and their area. These parameter values are estimated after analyzing the results that obtain better accuracy than previous parameters.

RESULTS & DISCUSSION

A procedure is followed and experimental results are designed to display the effectiveness and robustness of the new proposed technique. Firstly, examine the performance of the proposed technique on public datasets to select parameters. Then the technique that is proposed is experienced and compared with previous techniques i.e. Bernsen, Sauvola and B. Gatos over on standard DIBCO dataset of 2009 [18]. The binarization performance is evaluated by using Parameters such as PSNR, Precision, Recall, F-Measure, accuracy, Sensitivity, Specificity.

Experimental Results of Printed Documents


Salau
im
Marggrafthum Nieder-Lauſitz,
benebſt
deren Statuten, Receſſen, Privilegien,
und andern alten Urkunden.

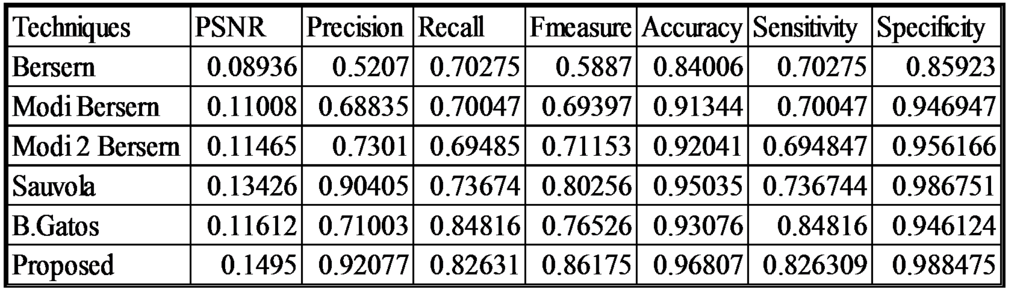
(a) (b) (c)

(d) (e) (f)

Fig. 2 Comparison Results of Printed Image (a) p02bersen (b) p02bersen1modi (algo1) (c) p02bersen2modi (algo2) (d) Local Method (Sauvola method) (e) B.Gatos Techniques (f) Proposed Technique

İn the above fig. 2 (a) shows the original image after applying Bernsen technique, fig. 2 (b) shows the image after applying Bernsen techniques with modified algorithm method, fig. 2 (c) shows the image after applying Bernsen second modified algorithm method, fig. 2 (d) shows the Sauvola Method with Local techniques, fig. 2 (e) image is the result with B.Gatos method, fig. 2 (f) image shown as result with new proposed technique. İn above images show that the new proposed technique gives better result of an image than others**.**

Table 1. Average Results Of Printed Dibco Dataset Images



Techniques	PSNR	Precision	Recall	Fmeasure	Accuracy	Sensitivity	Specificity
Bersern	0.08936	0.5207	0.70275	0.5887	0.84006	0.70275	0.85923
Modi Bersern	0.11008	0.68835	0.70047	0.69397	0.91344	0.70047	0.946947
Modi 2 Bersern	0.11465	0.7301	0.69485	0.71153	0.92041	0.694847	0.956166
Sauvola	0.13426	0.90405	0.73674	0.80256	0.95035	0.736744	0.986751
B.Gatos	0.11612	0.71003	0.84816	0.76526	0.93076	0.84816	0.946124
Proposed	0.1495	0.92077	0.82631	0.86175	0.96807	0.826309	0.988475

Above TABLE 1 show the overall average of all parameters showing the best result of the proposed technique.

GRAPH OF AVERAGE RESULT DATASET OF PRINTED IMAGES

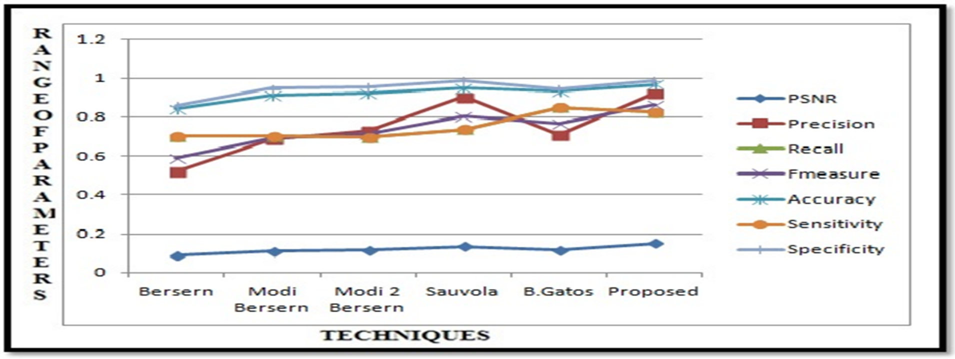
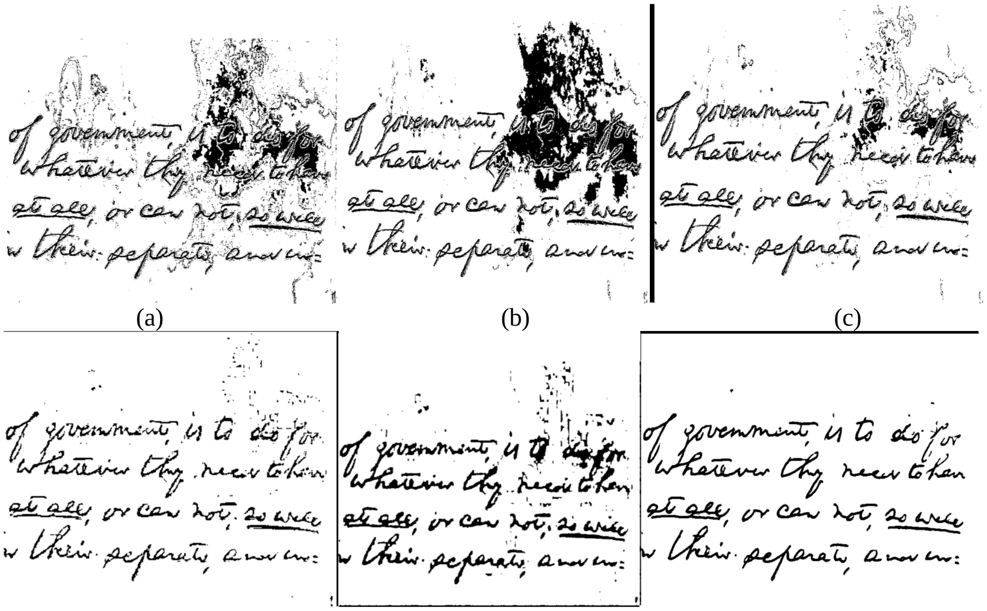


Fig. 3 Average Results of Graphical Interpretation of the all Printed DIBCO dataset images with all Parameters and Technique Implementation.

Above graph shows the overall result of average values as provided in TABLE I. This graph shows about graphical representation of all parameters average in which best results obtained from proposed technique.

Experimental Results of Handwritten Document

İn the above fig. 4 (a) shows the original image after applying Bernsen technique, fig. 4 (b) shows the image after applying Bernsen techniques with modified algorithm method, fig. 4 (c) shows the image after applying Bernsen second modified algorithm method, fig. 4 (d) shows the Sauvola Method with Local techniques, fig. 4 (e) image is the result with B.Gatos method, fig. 4 (f) image shown as result with new proposed technique.İn above images show that the new proposed technique gives better result of an image than others**.**



(a) (b) (c)

(d) (e) (f)

Fig. 4 Comparison Results of Handwritten Image (a) H04bersen (b) H04 bersen1modi (algo1) (c) H04 bersen2modi (algo2) (d) Local Method (Sauvola method) (e) B.Gatos Techniques (f) Proposed Technique

Table 2. Average Result Of Handwritten Dibco Dataset Images

Techniques	PSNR	Precision	Recall	Fmeasure	Accuracy	Sensitivity	Specificity
Bersen	0.13225	0.56709	0.79978	0.62491	0.94228	0.799776	0.951491
Modi 1 Bersen	0.13491	0.56411	0.78627	0.62359	0.94473	0.786269	0.954255
Modi 2 Bersen	0.15372	0.72409	0.75698	0.71733	0.96906	0.756979	0.974136
Sauvola	0.16181	0.84319	0.66598	0.7327	0.97401	0.665978	0.986604
B.Gatos	0.14266	0.61331	0.87705	0.67445	0.95952	0.87705	0.968686
Proposed	0.18804	0.86211	0.77246	0.86765	0.98404	0.77246	0.990037

Above TABLE II show the overall average of all parameters showing the best result of the proposed technique.

GRAPH OF AVERAGE RESULT DATASET OF HANDWRITTEN IMAGES

The graph shows the overall result of average values as provided in TABLE II. This graph shows about graphical representation of all parameters average in which best results obtained from proposed technique.

The proposed method is compared with the equivalent traditional techniques of handwritten and printed documents in the form of figures, tables and through graphs. The new proposed technique gives better results of documents for removing degradations.

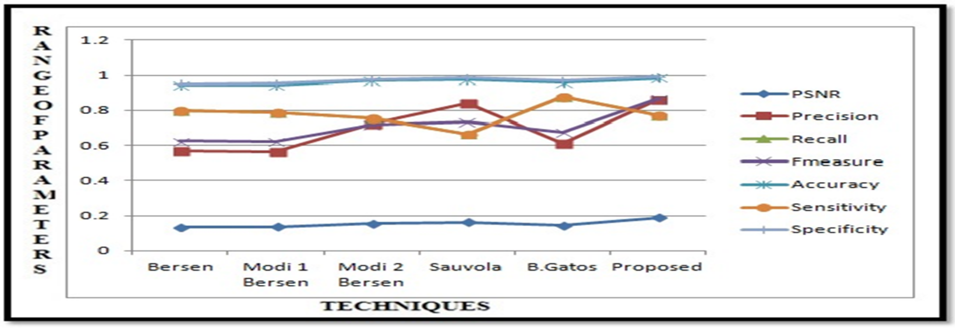


Fig. 5 Average results of graphical interpretation of the all Handwritten DIBCO dataset images with all parameters and Technique Implementation

CONCLUSION & FUTURE WORK

This paper presents binarization technique that is tolerant for dissimilar types of document degradation such as uneven enlightenment and document smear. The new proposed technique is simple and in which various parameters are involved. Moreover, it works on different types of documents i.e. handwritten and printed documents of degraded documented images. The new proposed technique making use of the local image contrast is evaluated based on the local maxima and minima. Accuracy value is 96.8% for printed images and 98.4% for handwritten images of DIBCO 2009 dataset [18]. The new proposed technique has been tested on the various images of dataset DIBCO 2009 [18] that experimentally shows the proposed method results are better than previous document binarization methods in term of the performance evaluated by using Parameters such as PSNR, Precision, Recall, F-Measure, Accuracy, Sensitivity and Specificity. The upcoming area of development will be on Novel degradation method for video sequences and also suggest for looking into the effective implementation of any types of Document that degraded in quality.

ACKNOWLEDGMENT

I would like to thank Naresh Kumar Garg Associate Professor, Department of Computer Science and Engineering, Giani Zail Singh, Punjab Technical University Campus, Bathinda who is the co-author of this paper for introducing the present topic and for his inspiring intellectual guidance, constructive criticism and valuable suggestion throughout the research work and also thankful to DIBCO dataset 2009 [18] for providing the degraded images on printed and handwritten documents.

REFERENCES

Benjamin Perret, Sébastien Lefèvre, Christophe Collet, and Éric Slezak, “Hyperconnections and Hierarchical Representations for Grayscale and Multiband Image Processing”, IEEE Transactions on Image Processing, Vol. 21, pp: 14-27, **2011**.

B.Gatos, I. Pratikakis and S.J. Perantonis, “Adaptive Degraded Document Image Binarization”, Pattern Recognition, Vol. 39(3), pp: 317 – 327, **2006**.

B. Gatos, I. Pratikakis and S.J. Perantonis, “An adaptive binarization technique for low quality historical documents”, IARP Workshop on Document Analysis Systems, Lecture Notes in Computer Science (3163), pp: 102 - 113, **2004**.

B. Gatos, I. Pratikakis and S.J. Perantonis, “Efficient Binarization of Historical and Degraded Document Images ’’, IEEE Transactions on Image Processing, Vol. 7, pp: 447 - 454, **2008**.

B. Gatos, I. Pratikakis and S.J. Perantonis, “Improve Document Image Binarization by Using a Combination of Multiple Binarization Techniques and Adapted Edge Information”, Proceedings of the 19th International Conference on Pattern Recognition, pp: 1 - 4, **2008**.

Bolan Su, el.at, “Robust Document Image Binarization Technique for Degraded Document Images”, IEEE Transactions on Image Processing, Vol. 22, pp: 1408-1417, **2013.**

Fung, C.C. and Chamchong, R., “A Review of Evaluation of Optimal Binarization Technique for Character Segmentation in Historical Manuscripts”, IEEE 3rd International Conference on Knowledge Discovery and Data Mining, pp: 236 – 240, **2010**

Jagroop Kaur, Dr.Rajiv Mahajan, “A Review of Degraded Document Image Binarization Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, **2014**.

J. Bernsen, “Dynamic Thresholding of Grey-Level Images”, 8th International Conference on Pattern Recognition, France-Paris, ICPR, pp: 1251 – 1255, **1986**.

J. He, Q.D.M. Do, A.C. Downton, and J.H. Kim., “A Comparison of Binarization Methods for Historical Archive Documents”, International Conference on Document Analysis and Recognition, pp: 538 – 542, **2005**.

Jiang Duan, Mengyang Zhang, Qing Li, “A Multi-stage Adaptive Binarization Scheme for Document Images”, Proceedings of the Second International Joint Conference on Computational Sciences and Optimization(CSO), Sanya, Hainan, China, IEEE, Vol. 1, pp: 867 - 869, **2009**.

J.J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen, “Adaptive Document Binarization", International Conference on Document Analysis and Recognition (ICDAR), Vol. 1, pp: 147 – 152, **1997**.

K. Ntirogiannis ,B. Gatos and I. Pratikakis, “A Modified Adaptive Logical Level Binarization Techniques For Historical Document Images”, 10^{th} International Conference on Document analysis and Recognition(ICDAR), Barcelona, Spain, IEEE, pp: 1171 – 1175, **2009**.

Ntogas Nikolaos, Ventzas Dimitrios, “A Binarization method for. Historical Manuscripts”, 12th WSEAS International Conference on Comunications, Heraklion, Greece, pp: 23 – 25, **2008**.

W. Niblack., “An Introduction to Digital Image Processing”, N.J.:Prentice Hall, pp: 115-116, **1986**.

Yahia S. Halabi, Zaid SA, Faris Hamdan, Khaled Haj Yousef, “Modeling Adaptive Degraded Document Image Binarization and Optical Character System”, Euro Journals Publishing, Inc., Vol.28, pp: 14 - 32, **2009**.

Y.H. Chiu et al. “Parameter-free based two-stage method for binarizing degraded document images” in Pattern Recognition, Vol. 45(12), pp: 4250-4262, **2012**.

http://users.iit.demokritos.gr/~bgat/DIBCO2009/benchmark/