International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Degraded Document Image Binarization Techniques

Jyoti Rani†* and Davinder Parkash†

†Department of Electronics & Communication Engineering, HCTM Kaithal, Haryana, India

Accepted 15 Dec 2015, Available online 20 Dec 2015, **Vol.5, No.6 (Dec 2015)**

Abstract

Document Image Binarization is performed in the preprocessing stage for document analysis and it aims to segment the foreground text from the document background. A fast and accurate document image binarization technique is important for the ensuing document image processing tasks such as optical character recognition (OCR) and Document Image Retrieval (DIR). This research area has been studied for decades; many techniques have been reported and applied on different commercial document analysis applications. However, there are still some unsolved problems need to be addressed due to the high inter/intra-variation between the text stroke and the document background across different document images. Image binarization is the method of separation of pixel values into dual collections, black as foreground and white as background. Thresholding has found to be a well-known technique used for binarization of document images. Thresholding is further divide into the global and local thresholding techniques.

Keywords: *Documents, Binarization, Local thresholding, Global Thresholding, binary image*

Nomenclature

OCR : Optical Character Recognition
DIR : Document Image Retrieval
MS : Multi Spectrum
PCA : Principle Component Analysis
LDA : Linear Discriminate Analysis
GSA : Gravitational Search Algorithm
TCM : Texton Co-occurance Matrix
DIBCO : Document Image Binarization Contest

1. Introduction

There is huge amount of textual information that is embedded within images. For example, more and more documents are digitalized everyday through scanner, camera and other equipment. Many digital images contain texts, and a large amount of textual information is embedded in web images. It will be very useful to convert the characters from image format to textual format by using optical character recognition (OCR). This converted text information is very important for documents images, document image retrieval and so on. However, in many cases, the document images cannot be directly fed to an OCR system due to the following reasons:

- The original document papers suffer from different kinds of degradation including smear, ink-bleeding through and intensity variation, especially for historical documents.
- The process of obtaining digital images from the real world is not perfect. There are many factors that may cause image distortion, such as incorrect focal length, camera shaking/object movement, low resolution, etc.

The web images in the internet are often susceptible to certain image degradation such as low resolution and small size, which is specially designed for faster network transmission rate, computer-generated-character artifacts, and special effects on images to attract visual attention.

1.1 Binarization

The main aim of image segmentation is to group image pixels according to constituent regions or objects. On document images this problem consists of two classes: foreground and background. The binarized image should be perceptually similar. For the binarization of documents global and adaptive binarization methods exist. The single threshold value is used for every pixel by global method, adaptive methods define local regions in which separate threshold values are calculated. Current binarization methods use gray value images as input. Color images can be converted with the standard conversion:

$$I(x, y) = 0.3R(x, y) + 0.59G(x, y) + 0.11B(x, y),$$

*Corresponding author: **Jyoti Rani**

where R, G and B are the Red, Green and Blue channel of the color image. For a gray value image I(x, y) with intensity values between 0 and 1 and a threshold T(x, y) each image pixel is classified in foreground (labeled as 1) and background (labeled as 0) resulting in the thresholded image $I_{th}(x, y)$:

$$I_{th}(x, y) = (1 \text{ if } I(x, y) > T(x, y)$$
$$0 \text{ if } I(x, y) \leq T(x, y)$$

where T(x, y) = Tg = constant if a global threshold is applied. Adaptive methods have the characteristic that the value of T depends on the local gray value characteristics. Global thresholds are suitable for images with a bimodal gray value distribution, where adaptive methods can handle documents with e.g. non-uniform illumination. Recent developments (see DIBCO and H-DIBCO) show that binarization methods estimate the background or combine multiple binarization methods to achieve a better segmentation. The methods presented comprise Niblack, Sauvola and a color segmentation method. In the following, state of the art methods of image binarization are categorized in global and adaptive methods, methods based on background estimation and methods that use a combination of different binarization methods.

1.1.1 Global Methods

Documents which are digitized with a defined setup (such as scanner which uses a constant illumination) and a defined minimum contrast between background and foreground (no faded out text) can use a pre-defined constant threshold value T.

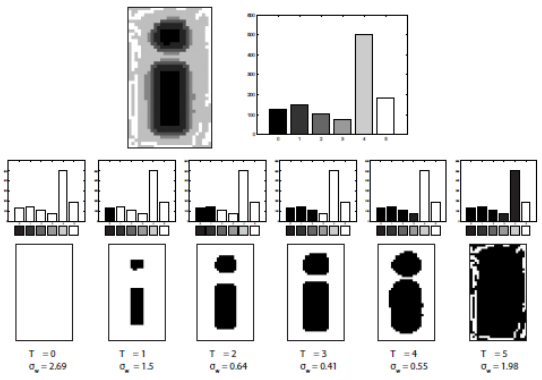
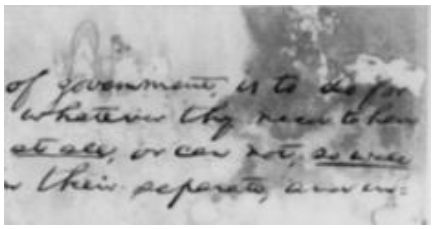
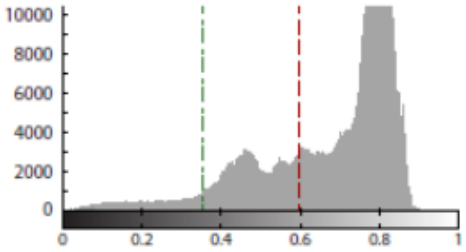


Figure 1.1: Gray value image of the character i and the corresponding gray value histogram.

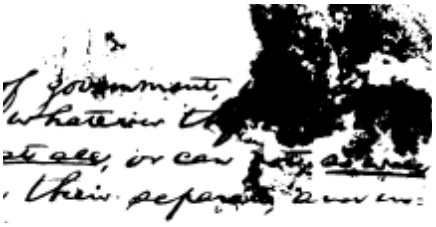
The results of all possible thresholds and the associated intra-class variance are shown to illustrate the output result of Otsu's method.



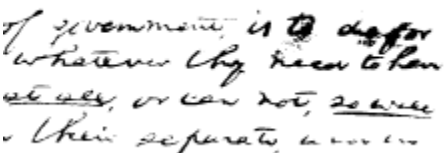
(a)



(b)



(c)



(d)

Figure 1.2: (a) Image of dataset (b) histogram with Otsu threshold (dashed) and manual threshold (dashed-dotted) (c) Otsu threshold image (d) manually thresholded image

A global threshold, which analysis the distribution of the gray values, is introduced by Otsu. Otsu's thresholding method assumes a bimodal histogram and minimizes the intra class variance. Global methods can be used for e.g. scanned documents which have constant illumination with a uniform background. Historical and degraded documents need adaptive algorithms due to the low contrast of faded out text and the presence of background clutter. Figure 1.1 shows a gray value image of the character i and the corresponding histogram. The image consists of 6 different gray values. To illustrate the methodology of Otsu, the image is thresholded at every possible gray value T and the binarized results with the associated intra class variance σ_w values are shown. It can be seen that at threshold T = 3 the classification into foreground and background leads to the smallest intra-class variance $\sigma_w = 0.41$, which will be the final result of Otsu's method. Figure 1.2 (a) shows an image of dataset with background variations and the corresponding histogram. It can be seen in Figure 1.2 (c) that the result of Otsu's method classifies parts of the background as foreground values due to the gray value distribution. But it is shown in Figure 1.2 (d) that a manual global threshold value leads to a better binarization result. Thus, Otsu determines the optimal global threshold value only for images with a bimodal distribution by definition.

1.1.2 Adaptive Methods

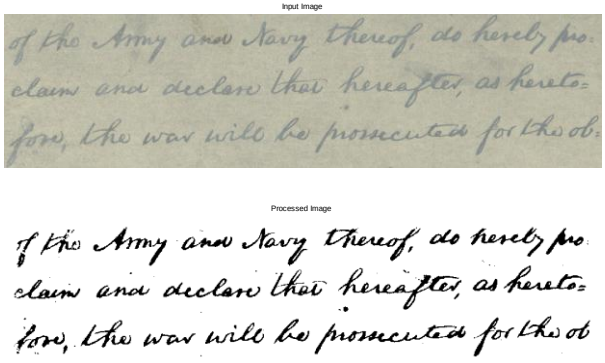
Adaptive methods define local regions R(x,y) and calculate a separate threshold value T(x, y) for each region. Current techniques make a rectangular subdivision of the gray value image depending on the character size. Niblack defines a threshold T based on the mean m and variance s within a local rectangular window by:

$$T = m + k \cdot s$$

where k is a negative parameter defining the amount of the print object boundary taken as a part of the given object (constant over the entire image). According to Gatos and Wolf et. al , the window size has to cover at least 1-2 characters and in k is set to -0.2 and the window size is 15 × 15 pixels. Milewski and Govindaraju state that document images with noise result in noise, jagged edges and broken character segments. An adaption of Niblack's algorithm is published by Sauvola and Pietikainen. In comparison to Niblacks method Sauvola can handle background noise. The proposed binarization algorithm is an adaption of Sauvola where the contrast and the mean gray level of the image is normalized. The main application of the method is multimedia documents and video frames. The threshold value using the normalized mean gray level is calculated by:

$$T = m - k\alpha\,(m - M),\ \alpha = 1 - s/R,\ R = \max(s)$$

where m is the mean gray value, s is the standard deviation, M is the minimum gray level of the image and R is the maximum of the standard deviation of all local windows. k is set to 0.5. To avoid salt and pepper noise pixels can also be divided into a third class which represents homogeneous regions. Homogeneous regions are classified into foreground and background based on their boundries.



1.1.3 Methods based on Background Estimation

A different class of adaptive algorithms specially used for ancient manuscripts are methods which estimate the background. A background estimation allows to compensate a variable background intensity caused by non-uniform intensity, shadows, smudge and low contrast. Gatos *et al.* use a Sauvola threshold for a rough foreground estimation. Based on the result of Sauvola they calculate a background surface estimation where foreground pixels are interpolated by a mean value of the surrounding background pixel. For the final threshold value the background image is subtracted from the original image to estimate the pixel contrast, and an adaptive threshold function based on a sigmoid function is defined. To get the result, a pre-processing is done by applying an adaptive low-pass Wiener filter. As a post-processing a shrink and swell filter is applied to remove noise and correct gaps, breaks or holes.

Lu *et al.* estimated the document background using a one dimensional polynomial smoothing (Savitzky-Golay smoothing). Afterwoads a global polynomial smoothing is applied to avoid the estimation of text regions (foreground) as in Su *et al.* and Gatos *et al.* The background image is compensated using the background image and thresholded as described in Su *et al.* using the text stroke edge image based on the contrast image. The proposed method is the winner of DIBCO 2009. Su et al use a normalized gradient image called contrast image - which is based on the local maximum and minimum of a 3 × 3 window. They state that the normalization compensates for the effect of the image contrast or brightness variation. To estimate the foreground & background similar to Gatos *et al.*, a simple threshold method (Otsu) is applied to the contrast image. The final threshold is defined by (m + s/2) where m is the mean value of the estimated foreground and s is the standard deviation within a local window. The window size is based on the mean stroke width which is determined using a horizontal projection and counting the distances between two stroke edges.

1.1.4 Binarization using Multi-spectral Images

An alternative to methods is mentioned to use multi-spectral imaging and to exploit information in the non-visible wavelengths of the reflected and emitted light of historical documents. Based on the assumption that the optoelectronic transfer function of the imaging system is linear, the cameras response r of an image pixel is given by the following equation

$$r = \text{integration}(\, I(\lambda)R(\lambda)O(\lambda)S(\lambda)\,)$$

where λ is the wavelength, I(λ) is the illumination energy that reaches the observed object, R(λ) is the color reflectance of the object, O(λ) describes the properties of the optical system and S(λ) is the responsiveness of the cameras sensor. Depending on the filters and illumination used, different spectral representations of cultural heritage objects (manuscripts) can be obtained. Figure 1.3 illustrates a possible Multi Spectrum (MS) acquisition setup. The imaging techniques used for the acquisition of ancient manuscripts are known as UV fluorescence or reflectography and IR reflectography. Based on the

properties of the writing material (for e.g. iron-gall ink) and the writing carrier (for e.g. parchment) the irradiated UV light is either reflected (UV reflectography) or absorbed resulting in a light source emitting radiation in the VIS part of the electromagnetic spectrum (UV fluorescence). Iron-gall inks used in ancient manuscripts have the property that they do not fluorescence in contrast to the parchment used as writing carrier. Thus, UV fluorescence can be used to enhance the contrast between the writing and the carrier material by exploiting optical properties of different materials. In Pentzien *et al.* it is stated that IR radiation is less scattered than visible light. By observing the reflected IR radiation (IR reflectography) it is possible to differentiate between different text layers (e.g. palimpsest text vs. newer text). Figure 1.4 shows a test panel where patterns are drawn with different writing materials. The patterns are covered with a painting layer (hide glue) such that the patterns are not visible to the human eye. It can be seen in Figure 1.4(b) that a global binarization (Otsu) applied to the painting area can only segment the patterns in the area which is not covered by the additional painting layer. If the panel is captured in the IR range of the light, the contrast of the drawn patterns is visible and can be segmented using a global binarization (see Figure 1.4 d) without any further contrast enhancement. Thus additional information is revealed in the multispectral images, which can be exploited for the binarization of document images. This shows the possibility to use the information within different wavelengths to enhance the result of the binarization. Hollaus uses Multi Spectral Imaging (MSI) to enhance the contrast of images by applying statistical analysis like Principle Component Analysis (PCA) and Linear Discriminate Analysis (LDA).

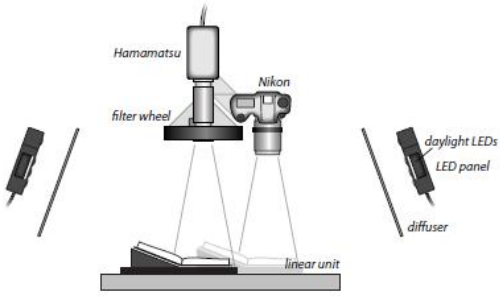


Figure 1.3: Schematic setup of the CVL MS acquisition system with UV illumination

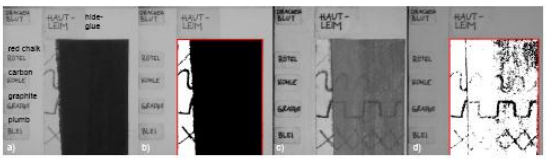


Figure 1.4: Test panel with different writing materials which are covered by hide glue. (a) RGB image (b) Global threshold of the test pattern (c) IR image (d) Global threshold of the IR image

1.2 Challenges on Degraded Document

Image Binarization though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem. This can be explained by the fact that the modeling of the document foreground or background is very difficult due to various types of document degradation such as uneven illumination, image contrast variation, bleeding-through, and smear. The recent Document Image Binarization Contests (DIBCO) held under the framework of the International Conference on Document Analysis and Recognition (ICDAR) 2009 and 2011 and the Handwritten Document Image Binarization Contest (H-DIBCO) held under the framework of the International Conference on Frontiers in Handwritten Recognition (ICFHR) show recent efforts on this issue.

Table 1.1: Document Image Binarization Method

Methods	Advantages	Disadvantages
Global Thresholding	Fast, Produce good results on clean documents	Fail on degraded images
Local Thresholding	Works on degraded document	Sensitive to window size
Background Subtraction	Produce good results when foreground varies	Performance decreased when background non uniform
Image Contrast	Produce good results when background varies	Performance decreased when foreground non uniform
Energy Based	Simple but effective	Need to tune a few parameters

These contests partially reflect the current efforts on this task as well as the common understanding that further efforts are required for better document image binarization solutions. Many practical document image binarization techniques have been applied on the commercial document image processing systems. These techniques perform well on the documents which do not suffer from serious document degradation. However, the degraded document image binarization is not fully explored and still needs further research.

2. Literature Review

Abdenour Sehad *et al.*(2013) has present a capable scheme for binarization of ancient and degraded document images, grounded on texture qualities. The suggested technique is an adaptive threshold-based. It has been calculated by using a descriptor centred on a co-occurrence matrix and the scheme is verified

objectively, on DIBCO dataset degraded documents furthermore subjectively, utilizing a set of ancient degraded documents offered by anational library. The outcomes are acceptable and assuring, present an improvement to classical approaches.

Konstantinos Ntirogiannis *et al.*(2013) has analysed that document image binarization is of incredible value in the document image examination and recognition pipeline as it disturbs further phases of the recognition procedure. The assessment of a binarization technique helps in examining its algorithmic conduct, and also confirming its adequacy, by giving qualitative and quantitative sign of its execution. A pixel-based binarization assessment approach for recorded handwritten/machine-printed document image has been proposed. .In the proposed assessment procedure, the review and accuracy assessment measures are fittingly adjusted utilizing a weighting plan that decreases any potential assessment unfairness. Extra execution measurements of the proposed assessment plan comprise of the rate rates of broken and missed content, false alerts, foundation commotion, character amplification, and combining.

Bolan Su *et al.* (2012) has studied a document image binarization structure that makes utilization of the Markov Random Field model. Structure isolates the document image pixels into three classes i.e. document background text, document foreground text, and uncertain pixels established binarization method. Uncertain pixels are belong to foreground and background categories by incorporating MRF model and boundary information.

P.Subashini, N.Sridevi (2011) presented comparison of various binarization algorithms by measuring their performance by evaluation metrics. In this paper, a new binarization method based on Particle Swarm Optimization (PSO) is proposed. A total of four different binarization methods such as Otsu, Niblack and Kittler Met along with proposed PSO method are considered and evaluated by evaluation metrics. From the experimental result, we can infer that proposed PSO method shows good result when compared with other methods, since their PSNR, SNR measures are higher and the MSE is lower. The higher value of PSNR means that the quality of the binarized image is better. According to the results, Proposed PSO method had the best overall performance with F-measures of 49.6418 which is higher when compared to other methods.

J. Sauvola, M. PietikaKinen A new method is presented for adaptive document image binarization, where the page is considered as a collection of subcomponents such as text, background and picture. The problems caused by noise, illumination and many source type-related degradations are addressed. Two new algorithms are applied to determine a local threshold for each pixel.

The performance evaluation of the algorithm utilizes test images with ground-truth, evaluation metrics for binarization of textual and synthetic images, and a weight-based ranking procedure for the nal result presentation. The proposed algorithms were tested with images including different types of document components and degradations. The results were compared with a number of known techniques in the literature. The benchmarking results show that the method adapts and performs well in each case qualitatively and quantitatively.

K. Ntirogiannis , B. Gatos b, I. Pratikakis c There are many challenges addressed in handwritten document image binarization, such as faint characters, bleed-through and large background ink stains. Usually, binarization methods cannot deal with all the degradation types effectively. Motivated by the low detection rate of faint characters in binarization of handwritten document images, a combination of a global and a local adaptive binarization method at connected component level is proposed that aims in an improved overall performance. Initially, background estimation is applied along with image normalization based on background compensation. Afterwards, global binarization is performed on the normalized image. In the binarized image very small components are discarded and representative characteristics of a document image such as the stroke width and the contrast are computed. Furthermore, local adaptive binarization is performed on the normalized image taking into account the aforementioned characteristics. Finally, the two binarization outputs are combined at connected component level. Our method achieves top performance after extensive testing on the DIBCO (Document Image Binarization Contest) series datasets which include a variety of degraded handwritten document images.

Conclusion

This paper has focused on the degraded document binarization technique. Document binarization is a key application of vision processing. The main objective of this paper is to evaluating the short comings of algorithms for degraded image binarization. It has been found that each technique has its own benefits and limitations; no technique is best for every case. The main limitation of existing work is found to be noisy and low intensity images. In near future we will propose a new algorithm which will use more reliable methodology to enhance the work. We will propose a new algorithm which will use nonlinear enhancement as a pre-processing technique to improve the results further.

References

Sayali Shukla, Ashwini Sonawane, Vrushali Topale, Pooja Tiwari (May,2014) Improving Degraded Document Images Using Binarization Technique, International Journal Of Scientific & Technology Research Volume 3, Issue 5.

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

Aroop Mukherjee, Soumen Kanrar (November 2010) Enhancement of Image Resolution by Binarization IJCA, Volume 10.

P.Subashini, N.Sridevi (September 2011) An Optimal Binarization Algorithm Based on Particle Swarm Optimization International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-4.

Yudong Zhang , Lenan WU (2011). Fast Document Image Binarization Based on an Improved Adaptive Otsu's Method and Destination Word Accumulation Journal of Computational Information Systems.

Chitrakala Gopalan, D.Manjula (Nov. 02, 2010) Sliding window approach based Text Binarisation from Complex Textual images (IJCSE) International Journal on Computer Science and Engineering Vol. 02, 309-313.

Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, IEEE (April 2013) Robust Document Image Binarization Technique for Degraded Document Images IEEE Transactions On Image Processing, Vol. 22, No. 4.

Qiang Chena, Quan-sen Suna, Pheng Ann Hengb, De-shen Xia (2007) A double-threshold image binarization method based on edge detector Pattern Recognition Society. Published by Elsevier Science Ltd.

Tao Chen, Mikio Takagi Dept. of electronics engg. Instiute of Industrial Science 'Image Binarization by Back Propagation Algorithm.

Bolan Su, Shijian Lu, and Chew Lim Tan (April 2013) Document Image Binarization using Background Estimation and Stroke Edges IEEE Transactions On Image Processing, Vol. 22, No. 4.

Abdenour Sehad, Youcef Chibani, Mohamed Cheriet and Yacine Yaddaden (April 2014) Co-occurrence matrix for ancient degraded document image binarization IEEE Transactions on Image Processing, Vol. 23, No. 4.

M. M. Mokji, S.A.R. Abu Bakar (October 2007) Adaptive Thresholding Based On Co-Occurrence Matrix Edge Information Journal Of Computers, Vol. 2, No. 8.

O. Imocha Singh, Tejmani Sinam (August 2012) Local Contrast and Mean based Thresholding Technique in Image Binarization International Journal of Computer Applications, Volume 51– No.6.

J. Sauvola and M. Pietikainen (2000) Adaptive document image binarization, Pattern Recognit., vol. 33, no. 2, pp.225 -236.

Om Prakash Vermaa, Rishabh Sharma, Deepak Kumar ([ICCCS-2012]) Binarization Based Image Edge Detection Using Bacterial Foraging Algorithm 2nd International Conference on Communication, Computing & Security.

K. Ntirogiannis, B. Gatos, I. Pratikakis (2012) combined approach for the Binarization of handwritten document images Pattern Recognition Letters, Elsevier.

Ioannis Pratikakis , Basilis Gatos and Konstantinos Ntirogiannis (2013) ICDAR 2013 Document Image Binarization Contest (DIBCO 2013), 12th International Conference on Document Analysis and Recognition.

T.Romen Singh , Sudipta Roy, O.Imocha Singh, Tejmani Sinam, (November 2011)A New Local Adaptive Thresholding Technique in Binarization IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2.

Bolan Su, Shijian Lu (April 2013) Binarization of Historical Document Images Using the Local Maximum and Minimum. IEEE Transactions On Image Processing, Vol. 22, No. 4.

Prashali Chaudhary, B.S. Saini (Jun-2014) An Effective And Robust Technique For The Binarization Of Degraded Document Images IJRET, Volume: 03 Issue: 06

Prof. S. P. Godse, Samadhan Nimbhore, Sujit Shitole, Dinesh Katke, Pradeep Kasar (May 2014) Recovery of badly degraded Document images using Binarization Technique International Journal of Scientific and Research Publications, Volume 4, Issue 5.

B. Gatos, I. Pratikakis, and I.J. Perantonis (2006) Adaptive degraded document image binarization Pattern Recognition, 39(3):317–327.

I. Blayvas, A. Bruckstein, and R. Kimmel (2006). Efficient computation of adaptive threshold surfaces for image binarization. Pattern Recognition, 39(1):89–101.

C.Wolf, J.M. Jolion, and F. Chassaing (2002) Text Localization, Enhancement and Binarization in Multimedia Documents International Conference on Pattern Recognition, (ICPR), 2:1037–1040.

B. Gatos, K. Ntirogiannis, and I. Pratikakis (Jul. 2009). ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 1375 – 1382.

Esmat Rashedi, Hossein Nezamabadi-pour, Saeid Saryazdi (2009). A Gravitational Search Algorithm. In Proceedings of the Information Science 179 (2009) 2232-2248. Published by Elsevier Science Ltd.

Document Image Binarization with automatic Parameter Tuning by Nicholas R. Howe.