IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-6 E-ISSN: 2347-2693

Data Document Image Binarization for Preserving Historical: A Review

Bharti Bansinge[1*] and R.K.Pateriya[2]

[1*,2] *Department of Computer Science & Engineering,*
Maulana Azad National Institute of Technology Bhopal, M.P., India

bharti.bansinge1102@gmail.com, pateriyark@gmail.com

www.ijcseonline.org

Received: Jun /05/2015 Revised: Jun/09/2015 Accepted: Jun/18/2015 Published: Jun/30/ 2015

***Abstract*—** The basic requirement of physical document analysis system is to digitalize the physical document. Recently number of researcher presented numerous techniques that can vary in sensitivity, quality and some more control parameters. Document binarization plays an important role in preserving the historical documents. The document image binarization focuses on extracting the text and background of the image. In doing this the edge detection approach also play the crucial role. This paper presents general review on the various approaches of document binarization. Various edge detection approaches are also been discussed. In addition various available data sets for image binarization developed in Document Image Binarization Contest (DIBCO) 2009 and Handwritten Document Image Binarization Competition (H-DIBCO) 2011 has also discussed.

***Keywords*—** Document Digitization, Edge Detection, Gaussian Filter.

I. INTRODUCTION

All Document digitization is an old but still a challenging and mind hunting task [1]. In the real world appearance of printed documents may vary with quality of printing color and shadings which degrade significantly quality of different binaries documents. Whereas the quality of different pixel of single binaries document may vary with light and view angle of different portion of physical document. Earlier, one of the main goals of document digitization was to differentiate the pixel of document image on the basis of their quality and take a proper treatment for them. But it's a very ambiguous process.

Digitations of physical documents generally use a representation of 24-bit color, or may be 8 bits of grayscale [2]. In most of recent applications these representation techniques do not include all of the data present in the original physical document, but retain more than enough. Recently research leads to retain a single bit per pixel document. Loyalty of digitations of physical document is relatively low, since information loss takes place, which is more or less related to the symbolic content of the document.

Most of the historical documents are produced using monochrome ink on paper, and their meanings are incorporated exclusively for the distribution of ink, a bit pattern representing the document explicitly.

Of course, the deduction of correct digitations of a document from color or grayscale representation can be difficult. The physical deterioration of the document, image illumination conditions and limits of the unfavorable resolution can contribute to obscure the original pattern. Now these days many researchers proposed numerous algorithms for digitations of physical document towards this dispute. In fact standard document image binarization contest (DIBCO) held in-order to gather a deep research in this era. However, the results of these competitions show that there is always room for improvement in the quality of automatic binarization.

II. IMAGE BINARIZATION

Academic institutions, libraries and historical museums pile-up or keep documents in storage areas. Image binarization contributes to maintaining a safe and effective exploitation of research in the original condition through the years. These are important to collect the historical documents that have to be preserved. It is a problem that the deterioration of the documents.

DIBCO scan documents and allows open access to general public for study and research while the cultural heritage institutions and organizations create local or national digital libraries that have restricted access through the Internet. The lot of work is done on the basic techniques that are used to improve and restore the image, eliminate noise and focuses on binarization.

Dr. R.K.Pateriya is currently working as associate professor in Computer Science & Engineering Department in Maulana Azad National Institute of Technology Bhopal, M.P., India, Ph-0755-4051313.

E-mail: pateriyark@gmail.com

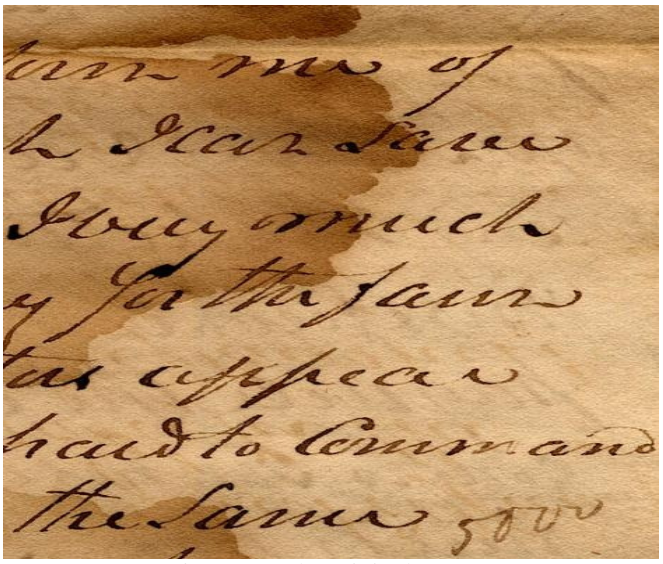


Fig.1. Sample Original Image

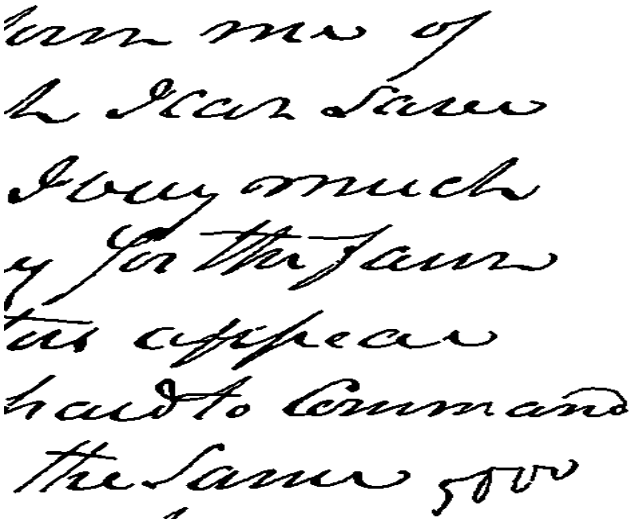


Fig.2. Sample image after binarization

III. DOCUMENT IMAGE BINARIZATION CLASSIFICATION

Image Binarization is primary step to perform the document image analysis by which the pixels of image can be distinguished in text and background. The optical character recognizer (OCR) and self-image recovery approaches are used for it. The changes within the background light of the image make it more challenging for researchers.

Now a days the adaptive image binarization is needed in which optimal thresholding of the image area from selected image is used. Thresholding approach is easy for the image segmentation from the gray scale image. Thresholding is capable of generating the binary images. Although the image binarization threshold that has been considered for a long period, of degraded images recorded, remains a problem yet to be solved. This may be seen from the way the offer submitted registration. Background is very problematic due to the different types of degradation of the images, for example, lighting degrees, the photo collection by contrast, dying over and swabs. Here we are trying to analysis on the strong and productive image binarization methods that have the ability to handle large archival purposes pictures. Generally severely degraded, this can be classified into three main types: binarization global, local binarization and methods of binarization hybrids.

The text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

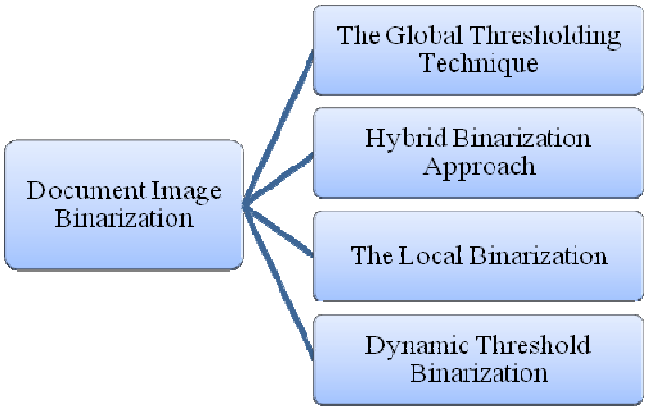


Fig.3. Classification of Image binarization

A. The Global Thresholding Technique

It calculates the optimal threshold for the entire image; these techniques require some calculations, and can work well in simple cases. But fails in complex backgrounds such as uneven color and poor backgrounds illuminated. These methods are generally not suitable for degraded document images because they have a clear pattern which separates foreground and background text.

B. The Local Binarization Techniques

Generally, these techniques are sensitive to background noise due to the great variation in the case of a poor document. Here different threshold approaches has been used for different target pixels which are based on the local pixels information.

C. *Hybrid Binarization Approach*

It is a combination of global and local threshold. The first step is to conduct a global threshold for the classification of the bottom of the document image and keep only the part that contains the foreground. The second step aims to improve the image obtained by the previous step, to get the result more clearly through the application of adaptive threshold technique.

D. *Dynamic Binarization Approach*

This method is used usually for binarizing images of poor quality, particularly with images with graph. However, due to the expense of the dynamic threshold, the method has a high computational complexity and slow speed.

IV. EDGE DETECTION TECHNIQUE

Digital image processing uses the computer algorithms to perform image processing. These are used widely in different digital imaging processes such as feature extraction, pattern recognition, segmentation, image morphology technology, etc. Edge detection is also independently developed in the field of image processing. Edge detection technique is basically image segmentation that divides the spatial scale, which is known as picture, in part or in large areas. The edges are the boundaries of excellence, and is therefore of fundamental importance in image processing problem. Usually occurs at the edges of the boundary between two different regions of the image. Edge detection allows the user to control the properties of the image where there is a change, more or less sudden in gray level or texture refers to the end zone in the picture and the beginning of another.

Edge detection is a facility by which the basic work of image processing like feature detection and its extraction can take place in the world of image processing. The objective of edge detection is to get the pixels or points of an image where the brightness of image going to be sharply change and may have some discontinuities. Edge detection also focused on reducing the size of image. It is important to minimize the time complexity. But still the structural properties remain same or it should be preserved.

The gray level image also has the edges but it comes under the local feature of an image. Here the neighbor pixels make the division according to its feature and create the region in order to separate the section by making the edge. It will depend on the uniform values on the different points of both sides of the edge in the image section.

It is hard to detect the edge in noisy images. Here, the results can be found but they have the blur and distortion in the results.

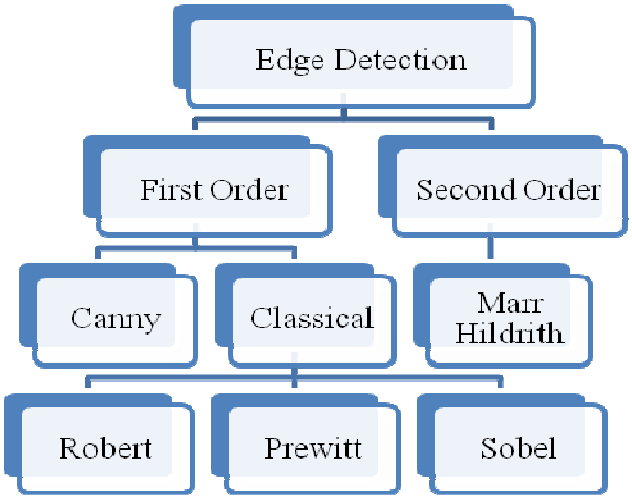


Fig.4. Edge Detection Classification

V. RELATED WORK

Otsu's method for document digitization is a parameter less global threshold method. In this method, presence of separate distributions for background and text has been assumed, and calculates a threshold value in such a manner that lead to minimize the difference between two distributions [4]. The limit for the two distribution of Otsu method was eliminated in [4], where the modes of degradation in the image histogram does one remove one applying recursively Otsu's method until only one mode remains in the picture. In another work, the overall limitation of the method is removed [5] and an adaptation method is introduced, which uses the same concept as the Otsu method, but local patches.

B. Gatos [6] introduced a binarization method in which initially a coarse binarization of the document image is obtained which is forwarded by rough background estimation. In the next step, local threshold values are calculated based on the estimated background and some parameters. These threshold values are used to calculate the final binarization that is post-processed to remove noise.

B. Gatos, K. Ntirogiannis and I. Pratikakis [7] presented their method in DIBCO'09. This comprises four steps. First, removing the background by polynomial fit of the lines. Second, detecting the contours of the race using Otsu's method on the gradient information. Third, then on local threshold by averaging the detected pixels in a local neighborhood window edges, and finally post-processing of results.

Fabrizio and Marcotegui [9] presented a method in DIBCO'9 based on morphological mapping operator shaft rocker [8]. To avoid salt and pepper noise associated with balancing mapping, are excluded from the analysis of pixel erosion and dilation which are too close. Pixels are then classified as text, background pixels and uncertain. The uncertain pixels are assigned to the text and background, depending on their class boundary.

Although images of the documents may suffer from degradation but we can assume that there are areas that could be described as actual text or background to enhance the quality of an image. This hypothesis has been the basis of many learning methods, which are based on a rough estimate of the text boxes and background and then try to learn their behavior to classify regions found in the confusion range [8, 10].In [11], a frame using a binarization method is presented to identify three classes; i.e. text, background, and uncertain pixels, the pixels of uncertainty reclassified using a classifier trained using the classes of the text and background.

VI. DOCUMENT IMAGE BINARIZATION CONTEST (DIBCO)

DIBCO 2009 was the first contest for the document binarization which was internationally organized. It was in a conference known as ICDAR 2009. The objective of the conference was to get the new researches in the area of document image binarization.

This conference has again organized in 2011. In both conferences the authors of the presented methods enrolled in the competition and downloaded representative samples with the corresponding ground truth. In next step, all registered participants were required to submit their executable binarization. After evaluating all methods of candidates, all test data and evaluation software (5 pictures handwritten with ground truth associated printed machine and 5) have become accessible to the public. (http://www.iit.demokritos.gr/~bgat/DIBCO2009/benchmark)

Image data set for printed media for ICDAR 2009

(A)	(B)
	

Table 1 Sample images of ICDAR 2009

Image data set for handwriting for ICDAR 2011

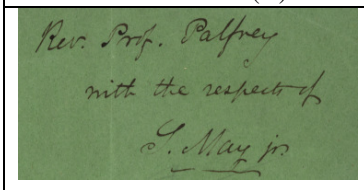
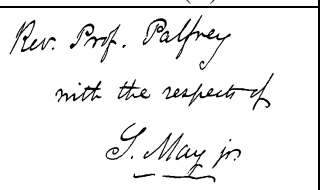
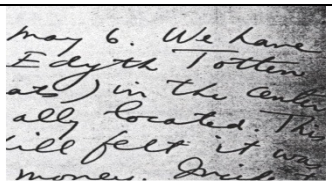
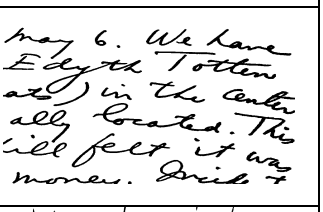
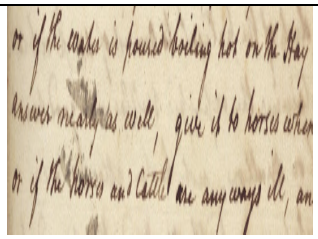
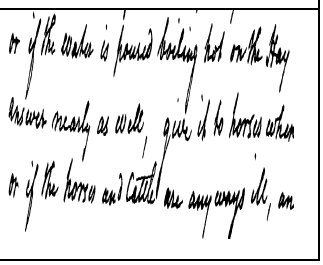
(A)	(B)
	
	
	

Table 2 Sample Images of ICDAR 2011

VII. CONCLUSION

In many applications, especially in the case of physically degraded documents, the smoothness of the edges and boundaries are highly important. Therefore, finding more precise techniques for document binarization is the necessity. In many applications, the continuity of the strokes and their topological aspects are more important than pixel-wise comparison with ground-truth binarized images. This paper throws some light on the basic

approaches regarding the document binarization techniques and the data set which can be used for the research work.

REFERENCES

[1] Reza Farrahi Moghaddamn, Mohamed Cheriet "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization", Elsevier transaction of Pattern Recognition**,2012**, pg no- **2419–2431**.

[2] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009), ICDAR'09**,2009**, pp. **1375–1382**.

[3] Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2011 document image binarization contest (DIBCO 2011), International Conference on Document Analysis and Recognition**,2011**, pp**. 1506–1510.**

[4] M. Sezgin, B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation", Journal of Electronic Imaging 13 (1),**2004**, pp.**146–168**.

[5] R. Farrahi Moghaddam, M. Cheriet, "A multi-scale framework for adaptive binarization of degraded document images", Pattern Recognition 43 (6),**2010**, pp. **2186–2198**.

[6] B. Gatos, I. Pratikakis, S.J. Perantonis, "Adaptive degraded document image Binarization", Pattern Recognition 39 (3),**2006**, pp. **317–327**.

[7] B. Gatos, K. Ntirogiannis, I. Pratikakis, DIBCO 2009: document image binarization contest, International Journal on Document Analysis and Recognition, **2010**,pp**. 1-10**.

[8] J. Fabrizio, B. Marcotegui, M. Cord, "Text segmentation in natural scenes using toggle-mapping", ICIP'09, **2009**, pp. **2373–2376**.

[9] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009), in: ICDAR'09,**2009,** pp. **1375–1382**.

[10] R. Hedjam, R. Farrahi Moghaddam, M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images", Pattern Recognition 44 (9),**2011**, pp.**2184–2196.**

[11] B. Su, S. Lu, C.L. Tan, "A self-training learning document binarization frame work", ICPR'10,**2010,** pp. **3187–3190**.

AUTHORS PROFILE

Bharti Bansinge received the B.E. degree in Computer Science from RGPV University, Bhopal in 2012. She is now the research fellow in Maulana Azad National Institute of Technology, Bhopal. Her current research interests include Document Binarization and Data Retrieval and Recognition.

Dr. R. K. Pateriya is currently associate professor in Maulana Azad National Institute of Technology, Bhopal. His current research interests include Web security and cloud computing.