An Ancient Degraded Images Revamping Using Binarization Technique

Kaveri Jagtap, Chandraprabha. A. Manjare

Abstract— Revamping of ancient degraded document images is a grueling task due their foreground text and background which is degraded due to uneven illumination, dust, water marks, smear, strain, ink bleed and low contrast etc. The proposed Binarization technique addresses this problem by using adaptive image contrast which is a combination of the local image gradient and local image contrast that is stoic to text and background variation. In the proposed technique, for an input ancient degraded document image an adaptive contrast map is first constructed. The contrast map is then binarized and combined with Canny's edge map to recognize the text stroke edge pixels. The text of document is further segmented by a local threshold that is concluded based on the intensities of detected text stroke edge pixels within a local window. Dataset of different languages like Modi, Marathi and English are used as input in handwritten and printed form. Modi, Marathi, English database are from year 1908, 1957, 1922.The proposed system is simple, required minimum parameter tuning, and give the superior performance compared with other techniques.

Index Terms— ***Document Image Processing, Document Analysis, Pixel Classification, Degraded Document Image Binarization, Adaptive Image Contrast.***

I. INTRODUCTION

Ancient document are of enormous importance to us, as it contain more valuable information. Some of ancient languages like Modi, Pandulipi are on the way of vanish or already vanished. So to preserve such languages that represent our culture it is necessary to restore documents that represent ancient languages. So to revamp such ancient document proposed system follows Binarization technique.
Several document image binarization techniques has been studied for long period, the thresholding of degraded document images is still an unsolved dispute as many degraded document do not have a clear bimodal pattern. Handwritten text within the degraded documents often shows Degradation of such ancient document is mainly due to the high inter/intra variation between the text stroke and the document background, Variation due to noise, uneven illumination, Degradation due to smear, smudge, shading, ink bleed through a certain amount of variation in terms of the stroke connection, stork brightness, and stroke width and document background.

Manuscript Received on November 2014

Kaveri Jagtap, Department of Electronics and Telecomm Engineering, JSPMs,Jayawantrao Sawant College of Engg, Hadapsar, Pune 28, India .

Chadraprabha. A. Manjare, Department of Electronics and Telecomm Engineering, JSPMs,Jayawantrao Sawant College of Engg, Hadapsar, Pune 28, India.

Different types of document degradations tend to cause the document thresholding error and make degraded document image binarization a big challenge.

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).Binarization technique is aimed to be used as a first stage in various document analysis, processing and retrieval tasks.

The proposed method is capable of handling different types of degraded document images with minimum parameter tuning. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradations. In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum algorithm. At the same time, the parameters used in the algorithm can be adaptively estimated.

The rest of this paper is organized as follows. Section II shows the reviews of current state-of-the-art binarization techniques. Proposed document binarization technique is described in Section III. Then experimental results are reported in Section IV to demonstrate the superior performance of our framework. Finally, conclusions are presented in Section V.

II. RELATED WORK

J. Sauvola and M. Pietikainen-2000 [1], represented method 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. Researcher has used Hybrid approach and took document region class properties into consideration. 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 results were compared with a number of known techniques like Bernsen, Eikvil. The benchmarking results show that the method adapts and performs well in each case qualitatively and quantitatively. The algorithm of

this method tolerated with severe noise, up to 45%.

T. Lelore and F. Bouchara-2009 [5] in his research presents approach for the binarization of seriously degraded manuscript. Introduce a technique based on a Markov Random Field (MRF) model of the document. Depending on the available information, the model parameters (clique potentials) are learned from training data or computed using heuristics. The observation model is estimated thanks to Expectation Maximization (EM) algorithm which extracts text and paper's features. The performance of the proposition is evaluated on several types of degraded document images where considerable background noise or variation in contrast and illumination exist. Comparison with Gatos et algorithm, Sauvola's algorithm, Wolf algorithm, Otsu's method is given, shows that bad behavior of Sauvola's and Wolf's cannot deal with low contrast and simultaneously dark regions and bright regions on the document. The global threshold of the Otsu's method can explain the poor results of the character recognition.

Patvardhan, C., A. K. Verma, and C. Vasantha Lakshmi-2012 [6] researcher has studied that images may contain difficult background i.e. shading or a denoising. Binarization method of document images creates them suitable for OCR using discrete curvelet transform. Curvelet transform is used for eliminate difficult image background, white Gaussian noise and gives improved binarized document image. The Curvelet transform also helps to enhanced in text shape still in the occurrence of noise. This method is capable to eliminate high frequency Gaussian noise and low frequency complex backgrounds and shows better performance. Researcher measure PSNR, FM at different noise strength, 17 and 93.5are average PSNR and FM for different curvelet transform methods.

Rabeux, Vincent, et al. in 2013-[9] research has an approach to expect the outcome of binarization algorithms on a known document image according to its situation of degradation. Document shaving degradation which result in binarization errors. To characterize the degradation of a document image by using different features based on the strength, amount and position of the degradation. These characteristics allow us to build calculation models of binarization algorithms that are very accurate according to R2 values and p-values. The prediction models are used to select the best binarization algorithm for a given document image. Repeated random sub-sampling cross-validation shows that the models are accurate gives max percentage error equals 11%.

Wagdy, M., Ibrahima Faye, and DayangRohaya-2013 [10], researcher has implemented a quick and proficient document image clean-up and binarization technique depends on retinex hypothesis and global thresholding. This technique joins of local and global thresholding with concept of retinex theory which can efficiently improve the degraded and poor quality document image. Then, quick global threshold is utilized to change over the document image into binary form. The new method conquers the limitations of the related global threshold techniques.

III. PROPOSED METHOD

This section describes the proposed document image binarization techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

A. Preprocessing

In the preprocessing Adaptive Contrast Map is applied to the input image. Adaptive Contrast Map, combine the local image contrast with the local image gradient of the input image. It detects many non-stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background. The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly.

Adaptive local image contrast as follows

$$Ca(i, j) = \alpha C(i, j) + (1-\alpha)(Imax(i,j)-Imin(i,j)) \quad (1)$$

C(i, j) denotes the local contrast and (Imax(i, j) − Imin(i, j)) refers to the local image gradient that is normalized to [0, 1] , α is the weight between local contrast and local gradient that is controlled based on the document image statistical information.

Ideally, the image contrast will be assigned with a high weight(i.e. large α) when the document image has significant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results.

Otherwise, the local image gradient will be assigned with a high weight. The proposed binarization technique relies more on image gradient and avoid the over normalization problem of our previous method

We model the mapping from document image intensity variation to α by a power function as follows:

$$\alpha = (Std /128)\ \gamma \quad (2)$$

Where

Std denotes the document image intensity standard deviation, and γ is a pre-defined parameter the local image gradient will play the major role in Equation 1 when γ is large and the local image contrast will play the major role when γ is small. So we get stock edge pixel detected image at the end of this module.

B. Text stroke edges detection

Detect the text stroke edge pixel candidate by using Otsu's global thresholding method.

Otsu's preserve the maximum number of background pixels and also generate less background pixel misclassification at first stage

As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels.

The binary map can be further improved through the combination with the edges by Canny's edge detector, because Canny's edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges.

The combination of Otsu's global thresholding and canny edge map helps to extract the text stroke edge pixels accurately.

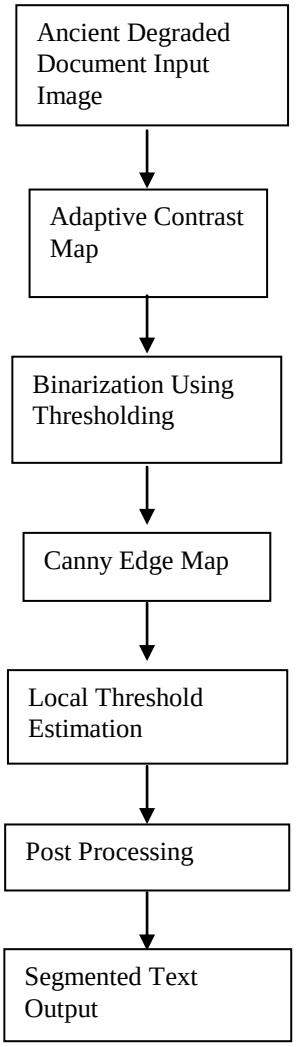

Fig.1. Block Diagram of Proposed System

C. Segmentation

The text can then be extracted from the document background pixels once the high contrast stroke edge pixels are detected properly.

Two characteristics can be observed from different kinds of document images:

1) The text pixels are close to the detected text stroke edge pixels.

2) There is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels.

The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

$$R(x, y) = 1 \; I(x, y) \leq \; Emean + Estd/2$$
$$= 0 \;\; otherwise \qquad (3)$$

Where,

Emean and Estd are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighbourhood window W, respectively.

Since we do not need a precise stroke width, we just calculate the most frequent distance between two adjacent edge pixels is done in horizontal direction and used it as the estimated stroke width. The edge image is scanned horizontally row by row and the edge pixel candidates are selected. The histogram is constructed that the frequency of the distance between two adjacent candidate pixels. The stroke edge width can then be approximately estimated by using the most frequently occurring distances of the adjacent edge pixels.

D. Post processing

The binarization result is further improved by post processing. The isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. The neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text). One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators.

IV. TESTING AND RESULT

1) Input Degraded Image

Fig.2. shows the Degraded Input Image English Database

2) Local Image Gradient image

This step gives Stock edge pixel detected image. Fig.3. shows the Contrast Gradient Output Image for English Database.

3) Local Image Contrast Image

Local Image Contrast suppresses the background Variation. Fig.4. shows the Local Image Contrast Image for English Database.

4) Adaptive Contrast Image

This step adds effect of both Local Image Gradient and Local mage Contrast and Fig. 5. Give background variation suppressed stroke edge detected image.

5) OSTU's Threshold Image

OSTU's Threshold detects the text stroke edge pixels. Fig 6. shows the OSTU's Threshold image for English Database.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

6) Complemented Canny Edge Detector Image

Canny edge Mark the edges close to real edge location but detect the wide range of edges including non-stoke edge. Fig.7. shows the Complement of Canny edge Detector Image for English Database.

7) Combination of OSTU's and Canny Edge Detection

Combination of OSTU's and canny keep only those pixels that appear within both high contrast image and canny edge map gives the stroke edge pixel extracted image.(Fig.8.)

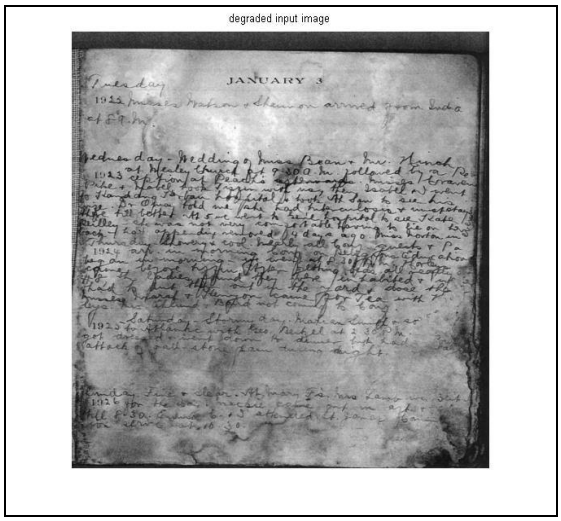

Fig. 2. Degraded Input Image English Database

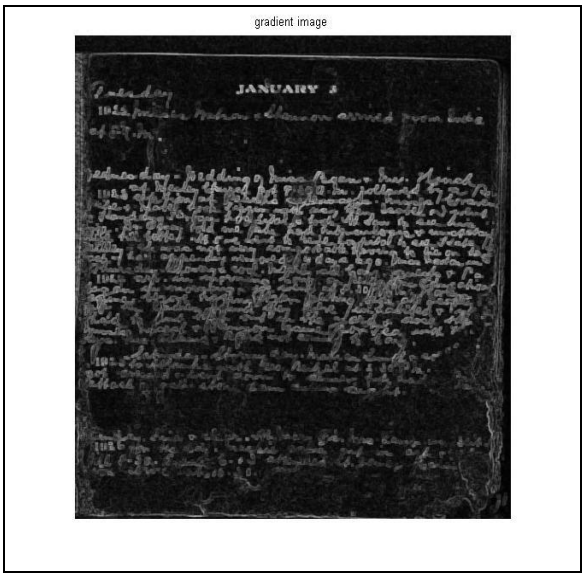

Fig 3.Contrast Gradient Output Image for English Database

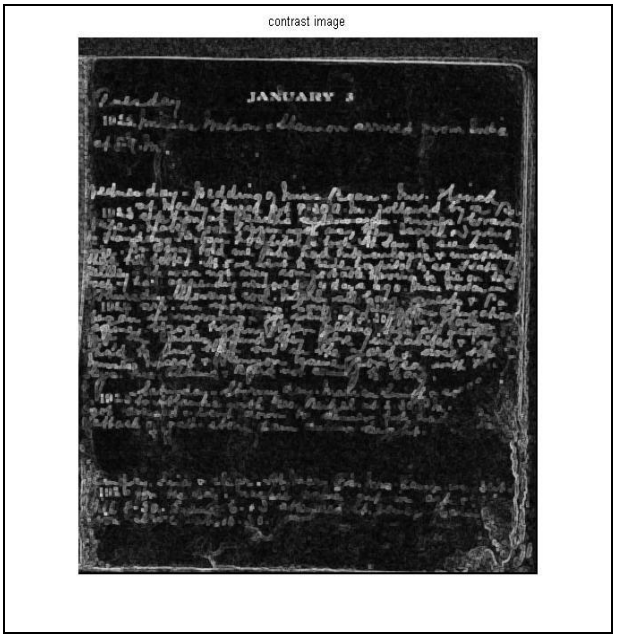

Fig 4.Local Image Contrast Image for English Database

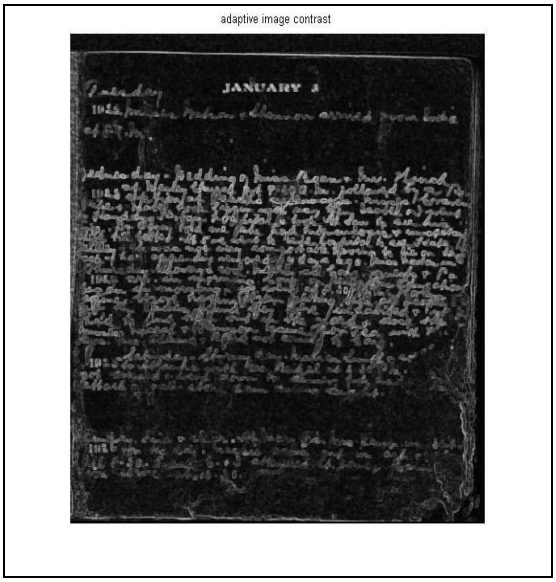

Fig 5. Adaptive Contrast map for English Database

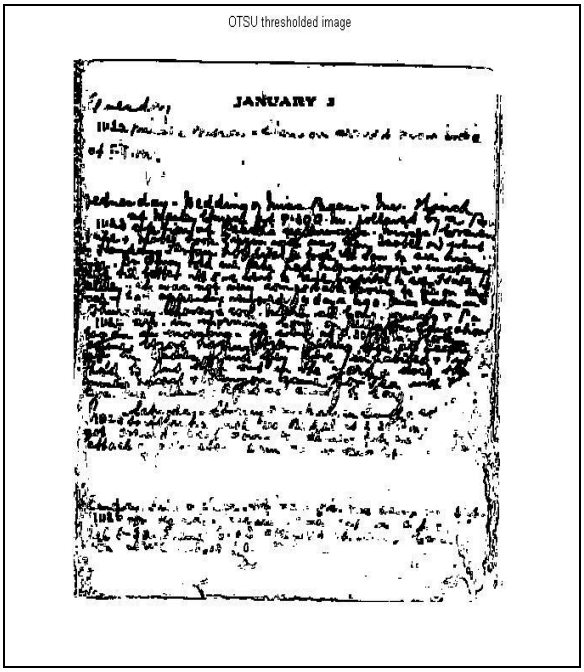

Fig 6. OSTU's Threshold image for English Database

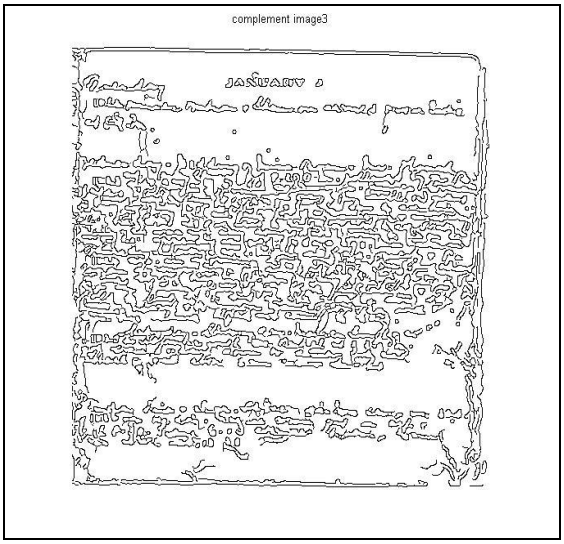

Fig 7.Complement of Canny edge Detector Image for English Database

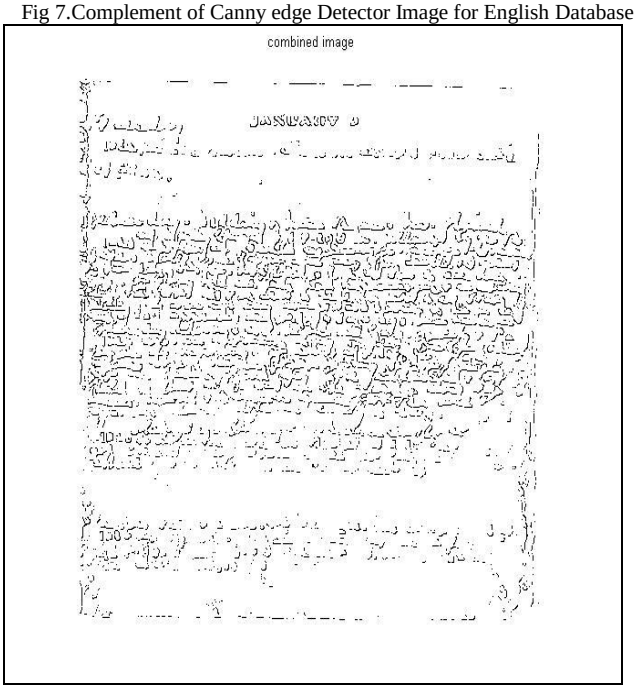

Fig 8. Stroke Edge Pixel Extracted image for English database

8) Time Evaluation for Different Database:

From Execution, I have conclude that Execution time required for handwritten database is more as compared to execution time required for Printed database, this is because writing style of humans which is vary due to Variation in curves of handwritten text, Size of handwritten data.Slant variation in handwritten data- few people write left hand, while some people write with right hand a slant of less than 90 degree is right hand slant a slant of more than 90 degree is a left hand slant. Average time in minute for handwritten data is 2.47 minute while for printed data is 2.22 minute.

1 TIME EVALUATIONS FOR DIFFERENT DATABASE

Database	Execution Time in minute (Average)
English – handwritten	2.51
English - printed	2.23
Modi – handwritten	2.48
Marathi - handwritten	2.43
Marathi - printed	2.21

V. CONCLUSION

Proposed Binarization technique is simple, required minimum parameter tuning. Execution time required (till second module) for handwritten database is more than required for printed database. Proposed system represents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear, sludge, etc. Proposed Method is suitable for all type, all Languages database not restricted to particular images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. It also conclude that the proposed algorithm eliminate the problem of over normalization.

REFERENCES

[1] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognition, vol. 33, no. 2, pp. 225–236, 2000.
[2] C. Wolf and D. Doermann, "Binarization of Low Quality Text using a Markov Random Field Model," International Conference on Pattern Recognition, pp. 160–163, 2002.
[3] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Iterative multimodel subimage binarization for handwritten character segmentation," IEEE Transactions on Image Processing, vol. 13, pp. 1223–1230, September 2004
[4] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, "Document image segmentation using a 2D conditional random field model," International Conference on Document Analysis and Recognition, pp. 407–411, September 2007.
[5] T. Lelore and F. Bouchara, "Document image binarisation using markov field model," International Conference on Document Analysis and Recognition, pp. 551–555, July 2009.
[6] Patvardhan, C., A. K. Verma, and C. Vasantha Lakshmi. "Document image denoising and binarization using Curvelet transform for OCR applications."Engineering (NUiCONE), 2012 Nirma University International Conference on. IEEE, 2012.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

[7] Papavassiliou, Vassilis, et al. "A Morphology Based Approach for Binarization of Handwritten Documents." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on.IEEE, 2012.
[8] Su, Bolan, Shijian Lu, and Chew Lim Tan. "A learning framework for degraded document image binarization using Markov random field." Pattern Recognition (ICPR), 2012 21st International Conference on.IEEE, 2012.
[9] Rabeux, V., et al. "Quality evaluation of ancient digitized documents for binarization prediction." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.IEEE, 2013.
[10] Wagdy, M., Ibrahima Faye, and DayangRohaya. "Fast and efficient document image clean up and binarization based on retinex theory." Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on.IEEE, 2013
[11] A. KEFALI, Toufik SARI, Mokhtar SELLAMI, "Evaluation of several binarization techniques for old Arabic documents images"
[12] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 1375–1382.
[13] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Handwrit. Recognit., Nov. 2010, pp. 727–732.
[14] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," Int. J. Document Anal. Recognition, vol. 13, no. 4, pp. 303–314, Dec. 2010.
[15] Bolan Su, Shijian Lu Member and Chew Lim Tan, "A Robust Document Image Binarization Technique for Degraded Document Images", in IEEE Transaction on Image Processing, Vol. 22 No. 4, April 2013.
[16] Manju Joseph, Jijina K.P, "An Improved Contrast Image Based Document Imaged Binarization Technique for Degraded Document Images", Vol. 04, Issue 04, April 2014.
[17] Prashali Chaudhary and B.S. Saini, "An Effective and Robust Technique For The Binarization Of Degraded Document Images", vol. 03, Issue. 06, June 2014.
[18] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images," in Proc. Int. Conf. Document Anal. Recognit., Sep. 1991, pp. 435–443.
[19] J. Parker, C. Jennings, and A. Salkauskas, "Thresholding using an illumination model," in Proc. Int. Conf. Doc. Anal. Recognit., Oct. 1993, pp. 270–273.