CSIT (November 2014) 2(3):153–161
DOI 10.1007/s40012-014-0045-5

ORIGINAL ARTICLE

Efficient binarization technique for severely degraded document images

Brij Mohan Singh · Mridula

Received: 22 April 2014 / Accepted: 27 June 2014 / Published online: 10 September 2014

Abstract Degradations in document images appear due to shadows, non-uniform illumination, ink bleed-through and blur caused by humidity. Thresholding of such document images either result in broken characters or detection of false texts. Numerous algorithms exist that can separate text and background efficiently in the textual regions of the document; but portions of background are mistaken as text in areas that hardly contain any text. This paper presents a way to overcome these problems by a robust binarization technique that recovers the text from a severely degraded document images and thereby increases the accuracy of optical character recognition systems. The proposed document recovery algorithm efficiently removes degradations from document images. Proposed work is based on the fusion of two well known binarization methods: Gatos et al. and Niblack, using dilation and logical AND operations. The results of our proposed binarization approach are seen to be better when compared to five existing well known approaches proposed by Otsu, Gatos et al., Niblack, Souvola et al., and Bernsen using four evaluations measures: Execution time, F-measure, PSNR, and NRM.

Keywords OCR · Binarization · Degradation · Dilation

B. M. Singh (✉)
Department of CSE, College of Engineering Roorkee,
Roorkee 247667, Uttarakhand, India
e-mail: bmsingh1981@gmail.com

Mridula
Department of Earthquake Engineering, IIT Roorkee,
Roorkee 247667, Uttarakhand, India
e-mail: mridulaiitr@gmail.com

1 Introduction

There exist lots of handwritten and printed historical manuscripts in libraries and museums in the world including medieval manuscripts, books of the renaissance, author manuscripts, old newspapers, archives, etc. [1]. Now day's old and historical documents are archived and preserved in large quantities worldwide in digital form. A lot of research effort has been dedicated to optical character recognition (OCR) systems to convert paper information into digital. Numbers of algorithms are available for this purpose and many commercial OCR systems are now available in the market. But most of these systems work on good quality documents. There are not sufficient numbers of research works for OCR that can handle degraded or poor quality handwritten and printed documents.

Pre-processing is the most important steps of OCR systems. Pre-processing of poor quality documents is very important for the accuracy of the other phases of OCR. Binarization (thresholding) of document images is the first most important step in pre-processing of poor quality scanned documents to save all or maximum subcomponents such us text, background and image [2]. Binarization computes the threshold value that differentiate object and background pixels [3]. The main advantage of binary images is that it decreases computational load and increases efficiency of the systems. Binary images can be obtained from gray level images by binarization. Binarization of degraded document images is not an easy task. The binarization is become more difficult when there are a varying illumination, variance of gray levels within the background and the object, inadequate contrast, object shape, noise and size non-commensurate with the scene. Degradations appear frequently and may occur due to several reasons which range from the acquisition source type to environmental conditions.

Binarization technique can be grouped into two categories: Global and Local. Global binarization methods pick one threshold value for the entire document images which is often based on an estimation of the background level from the intensity histogram of the image. These methods are very fast and produce good results. These methods fail when there are non uniform illuminated document images. Local (adaptive) binarization method uses different values for each pixel according to the local area information. These methods achieve good results even on severely degraded documents, but they are often slow since the computation of image features from the local neighbourhood is to be done for each image pixel.

The most well-known global method was introduced by Otsu [4]. Trier Jain [5] evaluated goal directed binarization Methods and shown that all well known existing algorithms work good on constant illuminated document images but fails in case of varying illumination. In general, global thresholding performs well for documents where there is a clear separation between foreground and background. Unfortunately, real life document images frequently exhibit various kinds of degradations such as illumination contrast, stains and noise, which weaken any guarantee for such a separation. Gatos et al. [6] proposed a method based on combination of multiple binarization techniques but it fails in case of low illumination images and produces broken characters. A method proposed by Niblack [7] completely recovers text from degraded document images but it magnified background noise. Sauvola and Pietikainen [8] proposed a similar algorithm by making some assumptions based on the distribution of grey values associated with foreground and background pixels. Bernsen [9] proposed a local thresholding method based on mean and contrast information for the calculation of threshold over a local region. The other existing studies on Binarization algorithms are available in [10–16].

In this research work, we focused on the robust binarization in recognition of degraded machine-printed as well as handwritten Devanagari and Latin scripts documents. The proposed technique of binarization to remove degradations from degraded document images is based on the fusion of two methods using dilation and logical AND operations of image processing.

In the following sections, we present a detailed description of the proposed methodology, as well as experimental results that demonstrate the efficiency of the proposed approach of binarization to recover text from degraded images.

2 Proposed methodology

In the off-line OCR, handwritten and printed document images to be recognized are captured by a sensor, for example, a scanner or a camera. Pre-processing is always a necessity whenever the document to be processed is noisy, inconsistent or incomplete. Pre-processing significantly improves the effectiveness of document analysis techniques. The proposed methodology depends upon the fusion of two well known binarization methods to recover text from degraded document images. The architecture of proposed methodology for binarization is shown in Fig. 1.

2.1 Gatos et al. method

Gatos proposed an algorithm [6] of binarization that is based on the following distinct techniques.

2.1.1 Initial filtering

Noise reduction is essential step in preprocessing of poor quality grayscale document images due to presence of noisy text areas. Contrast enhancement and smoothing between background and text areas are very important issue before applying other steps of recovering text from degraded and poor quality document images. For the initial filtering, Gatos chosen the Wiener filter since it is a popular filter for document enhancement [17] which has the ability to reduce the noise, smooth the texture but, in the process, also blur the texture. It also finds the ideal balance between contrast enhancement and noise gain.

A low-pass adaptive wiener filter [17] based on statistics estimated from a local neighborhood around each pixel. The filtered grayscale image 'I' is obtained from source grayscale image 'Is' according to formula:

$$I(x,y) = \mu + \frac{(\sigma^2 - v^2)\,(I_s(x,y) - \mu)}{\sigma^2} \quad (1)$$

Where μ is the local mean, σ^2 is the variance at 3×3 neighbourhood around each pixel and v^2 is the average of all estimated variance for each pixel in the neighbourhood. An example result of wiener filtering is shown in Fig. 2.

2.1.2 Approximation of foreground and background regions

A binary image S(x,y) is obtained from filtered gray image I(x,y) by applying Sauvola et al. 's [8] method using $k = 0.2$ to calculate a approximation of foreground pixels. To extract the binary image S(x, y), an image is processed I(x, y), where 1's correspond to the rough estimated foreground regions.

An approximation of background surface B(x, y) of the image I(x, y) is computed using the valuation of S(x, y) image. For pixels that correspond to 0's at image S(x, y), the corresponding value at B(x, y) equals to I(x, y). For the remaining pixels, the valuation of B(x, y) is computed by a

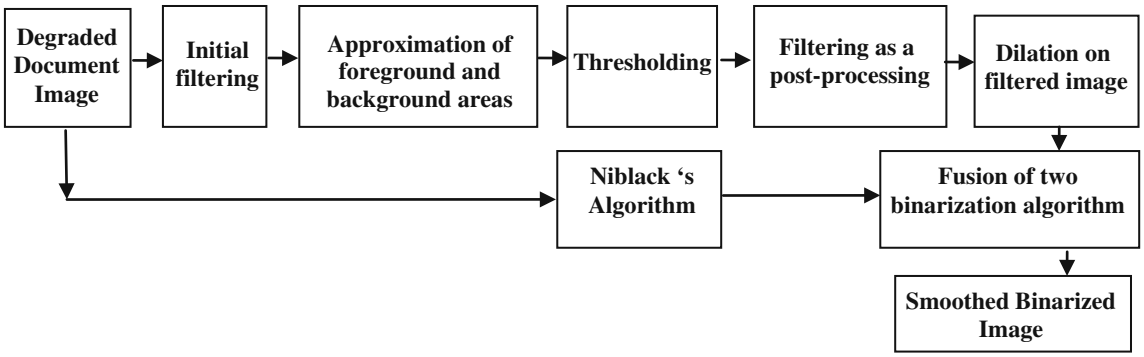
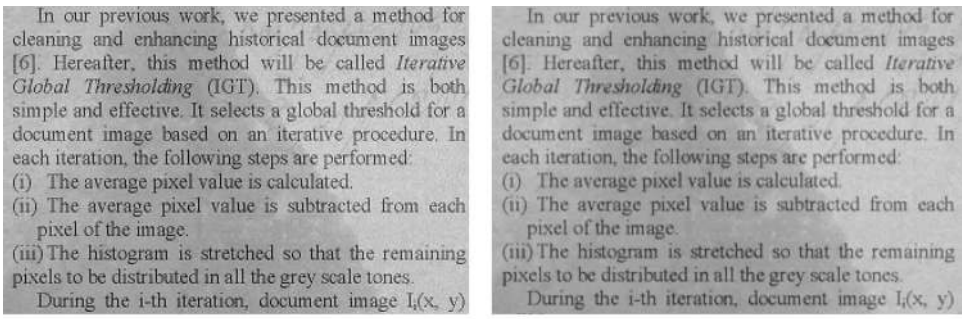


Fig. 1 An architecture of proposed methodology

Fig. 2 Wiener filtering



neighboring pixel interpolation that is described in the following equation

$$B(x,y) = \begin{cases} I(x,y), & S(x,y) = 0 \\ G(x,y), & S(x,y) = 1 \end{cases} \quad (2)$$

Where,

$$G(x,y) = \frac{\sum_{ix=x-d_x}^{x+d_x} \sum_{iy=y-d_y}^{y+d_y} (I(ix,iy)(1 - S(ix,iy)))}{\sum_{ix=x-d_x}^{x+d_x} \sum_{iy=y-d_y}^{y+d_y} (1 - S(ix,iy))} \quad (3)$$

The interpolation window size dx × dy is taken to cover at least two image characters.

2.1.3 Binarization

Binary image T(x,y) can be obtained using equation:

$$T(x,y) = \begin{cases} 1 & if\ B(x,y) - I(x,y) < T_d\ (B(x,y)) \\ 0 & otherwise \end{cases} \quad (4)$$

Where T_d is the threshold that must be changed according to the gray-scale value of the background surface B(x,y).

2.2 Niblack's method

Niblack's algorithm is also a local thresholding algorithm [7]. The threshold value for a pixel is decided by local mean and local standard deviation over a specific window size around each pixel. The local threshold T(x,y) for pixel (x,y) is calculated by formula:

$$T(x,y) = m(x,y) + k \cdot s(x,y) \quad (5)$$

where m(x,y) and s(x,y) are the local mean and local standard deviation of the pixels within the local window region. The local window is rectangular in nature and pixel (x,y) is the centre of it. The value of 'k' controls the amount of text region inside the local window. To conserve local details and handle local illumination level one requires small window size but if choose too small window size then it will not cover object and eliminate noise in the gray image. Window size of 15 × 15 and k = −0.2 was recommended by Trier and Jain [5].

2.3 Dilation on filtered image

Gatos et al. [6] proposed a binarization approach using all of the above mentioned steps. The observations of Gatos proposed approach, it removes degradations in document images such as shadows, non-uniform illumination, low contrast, large signal-dependent noise, smear and strain from images but produces broken characters, variable stroke width, in low resolution it completely removes the character or word from the document image and slows in

(a) Input Image (b) Output image of Gatos approach (c) Dilation on output of Gatos approach (d) Niblack's binarized output image (e) Final Binarized image obtained from our approach after AND operation on dilated image

Fig. 3 Steps of fusion of two methods

case of high resolution documents. It is a contrast independent binarization algorithm that efficiently extracts the text and eliminates the background.

So, to overcome these flaws, we applied a morphological dilation operation [17] on final filtered image, using circular structuring element with 7 × 7, radius 3, to search the text area on the binarized image.

2.4 Fusion of two binarization algorithms

The reason behind combining the Niblack's approach [7] with Gatos et al. [6] proposed approach is that Niblack's [7] method can distinguish the text from the background in the area close to the text and labels some part of background as text that is far from the text; it magnified background noise. The main advantage of using Niblack [7] method is that it completely recovers the text from badly illuminated degraded images but produces large noise in all degraded images and cannot separate background and foreground. The fusion of these two complementary binarization approaches is to get a reliable, robust and efficient binarization algorithm. We used an output of Niblack's [7] algorithm for AND operation with dilated output image of equation number 4 (binarized image). Logical AND operation is used to find the regions close to the text. The Fig. 3 shows the all steps of proposed approach to recover text from degraded image.

3 Evaluation approaches

Execution time, F-measure, peak signal-to-noise ratio (PSNR) and negative rate metric (NRM) were the standard performance measurements used to evaluate above mentioned algorithms.

The first evaluation measure is execution time: time spent by the system executing that task, including the time spent executing run-time or system services on its behalf.

The second measure is F-measure: the weighted harmonic mean of precision and recall. The equation of F-measure is as follows:

$$F_{alpha} = \frac{(1 + alpha) * precision * recall}{((alpha * precision) + recall)} \tag{6}$$

As the alpha value increases, the weight of recall increases in the measure. When using precision and recall, the set of possible labels for a given instance is divided into two subsets, one of which is considered "relevant" for the purposes of the metric. Recall is then computed as the fraction of correct instances among all instances that actually belong to the relevant subset, while precision is the fraction of correct instances among those that the algorithm believes to belong to the relevant subset.

The third evaluation measure is PSNR: the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.

$$PSNR = 10 \cdot \log\left(\frac{C^2}{MSE}\right) \tag{7}$$

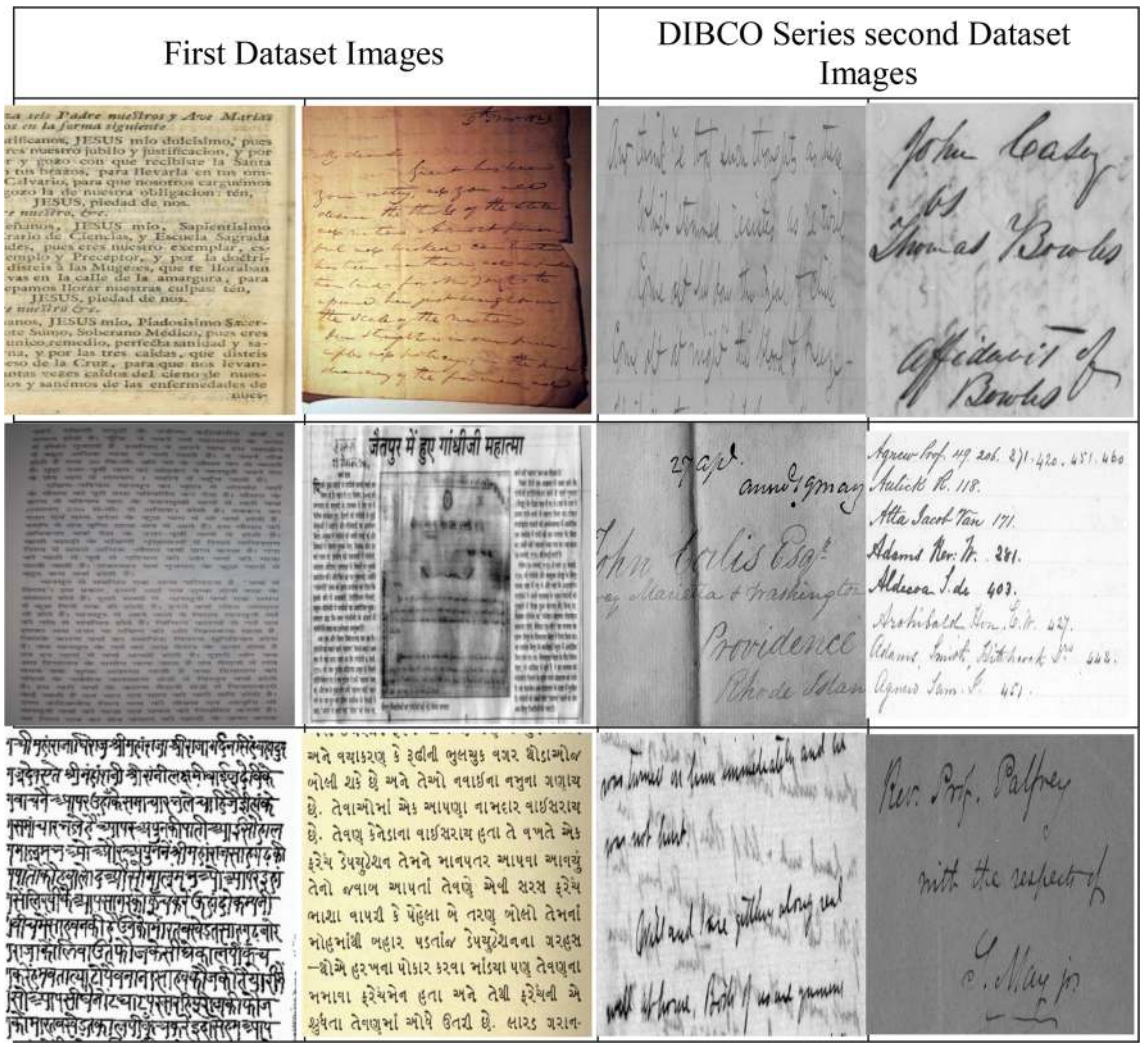


Fig. 4 Sample images of both datasets

Where, Mean Squared Error:

$$MSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}(I(x,y) - I'(x,y))^2}{MN} \quad (8)$$

I is the original and I' is binarized image.

The fourth evaluation measure is NRM: based on pixel-wise mismatches between the ground truth and observations in a frame.

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \quad (9)$$

$$NR_{FN} = \frac{FN}{FN + TP} \quad (10)$$

Where

$$NR_{FP} = \frac{FP}{FP + TN} \quad (11)$$

and

where NR_{FN} and NR_{FP} denote the number of false negative and false positive pixels respectively. TN and TP are the number of true negatives and true positives.

4 Results and discussions

The proposed methodology of binarization was tested on two sets of degraded document images. First dataset is collected from old newspapers, old books, computer generated, internet [18–20] and camera captured document images. All document images of first dataset contain different types of degradations such as variable background and shadows, non-uniform illumination, ink bleed-through, blur caused by humidity, different background complexity, noise levels, and very low illumination. In second dataset, we have used standard benchmark datasets: DIBCO (Document Image Binarization Contest: DIBCO 2009,

Fig. 5 Starting from left to right: **a-1**, **a-2**, and **a-3** are the input images, **b-1**, **b-2**, and **b-3** are the binarized output images of Otsu's method, **c-1**, **c-2**, and **c-3** are the binarized output images of Gatos's method, **d-1**, **d-2**, and **d-3** are the binarized output images of Niblack's method, **e-1**, **e-2**, and **e-3** are the output images of Bernsen's method, **f-1**, **f-2**, **f-3** are output images of Souvola's, method, **g-1**, **g-2**, and **g-3** are the output images of proposed binarization method

Fig. 6 DIBCO datasets results

H-DIBCO 2010 and DIBCO 2011) series [21–23], which include a variety of degraded document images. Sample images of both datasets are shown in Fig. 4. The experimental results of proposed approaches are shown in Figs. 5 and 6.

The observations are based on the two categories of experiments, first is based on visual perception and second is based on evaluation using four well known measures: F-measure, PSNR and NRM. We compared the performance of our methodology with five well-known binarization techniques: one global and four local. We evaluated and compared the following: Otsu's global thresholding method [4], Niblack's [7], Gatos et al. [6], Sauvola and Pietikainen [8], Bernsen's Method [9], and our proposed adaptive methods using above stated four well known evaluation measures.

On the basis of visual observations from the results, proposed binarization methodology over Otsu's, Gatos et al.'s, Niblack's, Souvola et al.'s, Bernson et al.'s methods are promising. Comparative experimental results of binarization are shown in Fig. 5. In Fig. 5, starting from left to right, a-1, a-2, and a-3 are the input images, b-1, b-2, and b-3 are binarized output images of Otsu's method, c-1, c-2, and c-3 are binarized output images of Gatos's method, d-1, d-2, and d-3 are binarized output images of Niblack's method, e-1, e-2, and e-3 are output images of Bernsen's method, f-1, f-2, and f-3 are output images of Souvola's, method, g-1, g-2, and g-3 are output images of proposed binarization method.

In the first set of experiment, which is based on the visual perception, the results are shown that the Otsu's method not gives satisfactory results in any test input document image as it is a global thresholding algorithm and degradations generally have local variance noise. We have compared Otsu's method just to show the comparison with at least one global method. Gatos method is a contrast independent binarization algorithm. It efficiently extracts the text and eliminates the background and works only for degradations which occur due to shadow, non-uniform illumination, low contrast, large signal-dependent noise, smear and strain but fails in case of low resolution documents, produces broken characters and variable stroke width and slow in case of high resolution documents. Niblack's method completely recovers text from degraded images but produces large noise, fails in the case of poor contrast images, the text regions are eliminated and cannot separate background and foreground. Bernsen's method produces large noise in all degraded images, fails in the case of poor contrast images, text regions are eliminated, and cannot separate background and foreground. Sauvola's method handles ink bleed-through and humidity exposed images but produces broken text in other cases, which is not suitable for OCR. Proposed method is superior to all four well known adaptive methods. It does not fail in any of the degradation conditions although it does not produce good results in case of very low resolution and very poor contrast.

In the second set of experiment, we evaluated the following: Otsu's global thresholding method [2], Niblack's adaptive thresholding method [8], Gatos et al. [6], Sauvola and Pietikainen adaptive method [11], and our proposed adaptive method using four well known evaluation measures: F-measure, PSNR and NRM. The evaluation results are shown in Table 1. All the experiments are carried out

Table 1 Evaluation results

Method	Time (s)	F-measure (%)	PSNR (dB)	NRM (10^{-2})
Otsu	0.017	36.80	07.68	27.21
Gatos	9.53	72.63	39.78	13.28
Niblack	02.83	47.92	17.26	14.35
Sauvola	02.92	42.38	29.34	32.29
Bernsen	02.87	54.32	21.09	11.21
Proposed	12.34	90.42	40.42	7. 59

using the hardware specifications of GPU: GeForce 9500 GT, 1 MB DDR2, No of processors = 4, No of core = 32, RAM 1 GB, frequency 1.35 GHz, DDR2 and CPU: Intel Core 2 Duo, 2.66 GHZ, No of cores available = 2, No of thread = 1, No of thread/core = 1, Physical Memory = 2 GB, DDR2.

The measures are calculated only for 15 degraded document images (1800 × 1000 pixels) containing all four types of degradations. The average values achieved from all document images on all evaluation measures are shown in Table 1.

Table 1 shows that the execution time of Otsu's, Gatos, Niblack, Sauvola, Bernsen, and proposed method are 0.017, 9.53, 02.83, 02.92, 02.87, 12.34 s respectively. Otsu's method is the fastest method among others and proposed method is slower than other methods. However, it is obvious that Otsu's method is global method which always is fastest than local methods. Niblack's method is the fastest method among all four adaptive methods: Gatos, Niblack, Bernsen and proposed. Proposed methodology is computationally high but produces better results than others.

F-measure is 36.80 % for Otsu's method, 72.63 % for Gatos's method, 47.92 % for Niblack's method, 42.38 % for Sauvola's method, 54.32 % for Bernsen's method, and 90.42 % for the proposed method. Proposed method's F-measure is the highest and Otsu's contains the lowest F-measure but in comparison with local methods, Sauvola's method produces the lowest F-measure.

PSNR is 07.68 db for Otsu's method, 39.78 db for Gatos's method, 17.26 db for Niblack's method, 29.34 db for Sauvola's method, 21.09 db for Bernsen's method, and 40.42 db for the proposed method. Proposed method's PSNR, 40.42 db is the highest and Otsu's contains the lowest 07.68 PSNR, but in comparison with local methods, Niblack's method produces the lowest PSNR.

NRM is 27.21for Otsu's method, 13.28 for Gatos's method, 14.35 for Niblack's method, 32.29 for Sauvola's method, 11.21 for Bernsen's method, and 7. 59 for the proposed method. Proposed method's NRM which is 7. 59, the lowest and Souvola's NRM is the highest 32.29. But in comparison with local methods, Sauvola's method produces the highest NRM.

5 Conclusions and future scope

In this paper, we proposed a robust approach for removing degradations and recovering text from Latin and Devanagari machine printed and handwritten scanned document images. Comparative experimental results of binarization on numerous degraded document images, demonstrate the efficiency of our proposed methodology. Our proposed methodology of binarization fails in case of very low resolution images because it is a contrast independent binarization algorithm.

Most of the document analysis and recognition works reported are on good-quality documents. But still it remains a highly challenging task to implement an OCR that works under all possible conditions and gives highly accurate results. Elaborate studies on poor-quality documents are not much undertaken by the scientists in the development of script independent OCR. Experiments should be made to observe the effect of poor quality paper as well as noise of various types, and take corrective measures.

References

1. Lasmar AG, Kricha A, Essoukri N and Amara B (2006) A segmentation text/background method for degraded ancient Arabic manuscript, pp 1327–1331, IEEE
2. He J, Do QDM, Downton AC and Kim JH (2005) A comparison of binarization methods for historical archive documents. In: Eighth international conference on document analysis and recognition (ICDAR'05), pp 538–542
3. Jindal M.K, Sharma RK and Lehal GS(2007) A study of different kinds of degradation in printed Gurmukhi script. In: Proceedings of the international conference on computing: theory and applications (ICCTA'07), pp 538–544, IEEE
4. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 9(1):62–66
5. Trier OD, Jain AK (1995) Goal-directed evaluation of binarization methods. Pattern Anal Mach Intell IEEE Trans 17(12):1191–1201
6. Gatos B, Pratikakis I, Perantonis SJ (2006) Adaptive degraded document image binarization. Pattern Recognit 39:317–327
7. Niblack W (1986) An introduction to digital image processing. Prentice-Hall, Englewood Cliffs, pp 115–116
8. Sauvola J, Pietikainen M (2000) Adaptive document image thresholding. Pattern Recognit 33:225–236
9. Bernsen J (1986) Dynamic thresholding of grey-level images. In: Proceedings of the eighth ICPR, pp 1251–1255
10. Kittler J, Illingworth J (1985) On threshold selection using clustering criteria. IEEE Trans Syst Man Cybernet 15:652–655
11. Brink AD (1992) Thresholding of digital images using two-dimensional entropies. Pattern Recognit 25(8):803–808
12. Yan H (1996) Unified formulation of a class of image thresholding techniques. Pattern Recognit 29(12):2025–2032
13. Sahoo PK, Soltani S, Wong AKC (1988) A survey of thresholding techniques. Comput Vis Graph Image Process 41(2):233–260
14. Kim IK, Jung DW, Park RH (2002) Document image thresholding based on topographic analysis using a water flow model. Pattern Recognit 35:265–277

15. Yang J, Chen Y, Hsu W (1994) Adaptive thresholding algorithm and its hardware implementation. Pattern Recognit Lett 15(2):141–150
16. Parker JR, Jennings C and Salkauskas AG (1993) Thresholding using an illumination model. In: International conference on document analysis and recognition (ICDAR'93), pp. 270–273, IEEE
17. Gonzalez RC, Woods RE, Eddins SL (2010) Digital image processing using MATLAB, 2nd edn. Mc Graw Hill Education, New Delhi, India
18. The Library of Congress (http://memory.loc.gov/). Accessed Nov 2010
19. Holy Monastery of St. Catherine at Mount Sinai (http://www.sinaimonastery.com/). Accessed Nov 2010
20. Bodleian Library—University of Oxford, (http://www.bodley.ox.ac.uk/). Accessed Nov 2010
21. Users.iit.demokritos.gr/∼ bgat/DIBCO2009/benchmark/. Accessed Nov 2010
22. Users.iit.demokritos.gr/∼ bgat/H-DIBCO2010/benchmark/. Accessed Nov 2010
23. Utopia.duth.gr/∼ ipratika/DIBCO2011/benchmark/. Accessed Nov 2010