Binarization and its Evaluation for Urdu Nastalique Document Images

Mamoona Naz Qurat ul Ain Akram Sarmad Hussain
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore, Pakistan
mamoona.naz@kics.edu.pk, ainie.akram@kics.edu.pk, sarmad.hussain@kics.edu.pk

Abstract - **Binarization converts a colored or gray scale image into a black and white image and is normally a preliminary step in optical character recognition. Binarization of images of Urdu language documents written in Nastalique writing style requires particular attention because Nastalique is not written with a uniform stroke but as a sequence of thin and thick strokes with a variety of marks. In the current work, three binarization methods are compared to determine an accurate and efficient technique for Urdu. This technique is further tuned for binarizing Urdu document images written in Nastalique writing style, to avoid disconnecting thin character connections but also to simultaneously prevent joining of diacritics with main bodies due to thickened strokes.**

Keywords – Urdu Optical Character Recognition, binarization, Urdu image corpus

I. Introduction

Urdu is spoken by more than 100 million speakers (as first or second language). It uses Arabic script, with enhanced character-set compared to Arabic and Persian languages [1]. Urdu is written in Nastalique writing style, which is cursive and written from right to left. Based on whether a word contains joining or non-joining characters, a word is normally divided into sub-parts, each called a ligature. For example, the word پاکستان ("Pakistan") contains three joined portions or ligatures. Each ligature is composed of a main body and zero or more diacritics (e.g. dots). A main body or a diacritic can also be generally referred to as a connected component (CC). Nastalique writing is done with a flat nib whose width is referred to a *Qat*. Thus, as the direction of the stroke changes it results in change in thickness of the stroke.

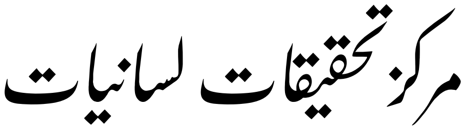

مرکز تحقیقات لسانیات

Fig. 1. Sequence of Thick and Thin Strokes in Nastalique Writing

Nastalique is written in a way that characters join within a ligature are always on a thin stroke and thick in other places see Figure 1. As the writing is very thin at the joins, and binarization technique has to be robust to avoid disconnecting the stroke of the main body (of ligature) at these joins. However, if we make the binarization technique liberal to avoid such disconnectivity at very thin connections, the multiple marks (including dots, etc.) start joining with the strokes of the main bodies. The current work explores how various methods compare in this context and further enhances the most promising technique for optimally balancing the constraints imposed by Urdu Nastalique writing style.

Section 2 gives an overview of some relevant binarization methods. Section 3 gives the methodology, and Section 4 discusses experimental results.

II. Overview of Binarization Methods

Binarization is normally done by setting threshold to identify whether a pixel in an image should be converted to black or white value. Based on how this threshold is set, binarization methods are categorized into the following three classes: (i) Global, (ii) Local and (iii) Hybrid thresholding methods. Global methods compute a threshold value for whole image. These are computationally inexpensive and better for scanned documents having uniform illumination but produce noise artifacts if gray scale document contains non uniform illumination [2]. The local methods divide whole image into smaller windows and compute a different threshold for each window. Local methods overcome the drawbacks of global binarization methods but less efficient [2]. The hybrid binarization methods combine information of global and local thresholds for better accuracy but are complex in nature [3].

Among global binarization methods Otsu global binarization is widely used for binarizing images [4].

This method sets threshold to minimize intra class variance, represented by (1), where $w_1(t)$ is sum of all intensity values below the threshold *t* for foreground and $w_2(t)$ is sum of all intensity values from *t* to maximum intensity value in an image (which is normally 255 for gray scale) for background. $\sigma_1^2(t)$ and $\sigma_2^2(t)$ are variances of foreground and background classes respectively. For different values of *t* which range from 0-255, $\sigma_w^2(t)$ is computed and the best value of *t* is selected which has minimum value of σ_w^2 .

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \qquad (1)$$

Niblack [5] partitions a page by sliding a fixed size window and calculates the threshold for each pixel based on local context, using the formula in (2), where *m* is mean and *s* is standard deviation of all pixel values in a window, and *k* is constant value that has value in the range of 0 and 1 which determines how effectively an edge of an object is retained. Niblack method is an appropriate choice for detecting region of images in a field of low quality [3].

$$T = m + k * s \qquad (2)$$

Sauvola [6] extends Niblack method, using the equation in (3), where *R* is dynamic range of standard deviation calculated for the document and has value 125. This method performs better for document images having background with light texture and larger variation in illumination [3].

$$t(x,y) = m(x,y) * (1 + k(\tfrac{s(x,y)}{R} - 1)) \qquad (3)$$

Shafait [3] presents local adaptive thresholding using Sauvola method with modification in calculating local mean and variance. This approach computes local mean and variance by using integral image. The binarization is same as Sauvola method but this technique is as efficient as global methods.

Nick method [9] is an extension of Niblack algorithm, in which an image is partitioned by moving a rectangular sliding window across the gray level image. Window threshold is computed using (4), where P_i is pixel gray-scale value, *NP* is total number of pixels in the window and *k* has value in the range between -0.1 and -0.2. Nick's method can be used to solve the low contrast problem .

$$T = m + k\sqrt{\frac{\sum(P_i{}^2 - m^2)}{NP}} \qquad (4)$$

Bukhari [7] proposes a modification to Sauvola method by introducing varying k value depending upon presence of ridges using (5), where *m(x,y)* is mean, $\sigma(x,y)$ is standard deviation and *k(x,y)* = 0.05 if any ridge is present in the local neighborhood window centered around the pixel *(x,y),* otherwise = 0.2. This method is preliminary designed for degraded hand-held camera-captured document images and solves the problems like non-uniform illumination, bad shading, blurring, smearing and low resolution.

$$t(x,y) = m(x,y)[1 + k(x,y)(\tfrac{\sigma(x,y)}{R} - 1)] \qquad (5)$$

Bataineh [2] uses local adaptive thresholding method to overcome problems of low contrast images and thin pen stroke, using global mean and adaptive standard deviation of window in computing threshold, using (6), where m_w and σ_w are mean and standard deviation of selected window and m_g is mean of whole image. $\sigma_{adaptive}$ is adaptive standard deviation of window and uses standard deviation of window, maximum and minimum standard deviation value among all windows in whole image.

$$T = m_w - (m_w^2 - \sigma_w)/((m_g + \sigma_w) * (\sigma_{adaptive} + \sigma_w)) \qquad (6)$$

III.METHODOLOGY

Three algorithms are initially selected, including Otsu global method, Bataineh adaptive method and Sauvola local method. The Otsu global method is culled because of its time efficiency [3] while Sauvola method is selected because it performs better among local binarization techniques [3]. Bataineh method is selected due to its adaptive nature [2]. In Otsu global method and Bataineh adaptive methods there are no free parameters. Sauvola method produces different results for as the free parameter *k* is varied. To capture a range, k values of 0.03, 0.08, 0.09, 0.1, 0.13, 0.14 and 0.2 are selected after initial experimentation. Impact of value of *R* is not found to be significant and is set to 128. Further, non-overlapping windows are considered for efficient processing.

The binarization techniques are evaluated using a gray scale Urdu document image corpus. The corpus is designed to capture variation in publishers, publication dates (since 1995), paper quality, print quality and paper transparency from a variety of published books. A subset of the corpus (300 pages from 100 books) in 14 point size is chosen for evaluation of binarization, as this size is used for publishing Urdu books. The same pages are scanned using HP Scanjet G3110 and directly binarized by the scanner to create a reference corpus. This scanner is selected after manual screening of a variety of scanners in common use.

Both efficiency and accuracy of the binarization techniques are computed for evaluation. Efficiency is calculated for 300 Urdu document images. Average time per page is used to compare the selected methods. Manual verification for accuracy is very time consuming for a large set of document images. Therefore, initially we generated a set of reference documents and developed automated methods to evaluate the accuracy, as per the details given below.

We measure accuracy of binarization using the criterion that black and white pixels of binarized image should maximally match the corresponding pixels in the reference image for the CC. This is computed in a two phases. First, the reference and binarized document images are auto-correlated to find maximal alignment using (7). This is computed for 100 placements of the binarized image over the reference image to get the best match (starting by mapping the center of the binarized image at the center for the reference page, and moving the former in a *10x10* window). Placement against the maximum score is selected for each page, where *BlackonBlack* is true if a black pixel in the binarized image matches with the corresponding black pixel in the reference. *WhiteonBlack* and *BlackonWhite* indicate pixels of binarized image that do not match with the pixels in the reference image, and therefore add a penalty score.

$$Alignment\ Score = \sum BlackonBlack - \sum WhiteonBlack - \sum BlackonWhite \qquad (7)$$

Once the reference and binarized images from the binarization methods are aligned, all the corresponding CCs are evaluated separately for the second phase of alignment. Four cases are identified:

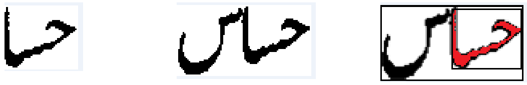
a) CC in binarized image aligns with the corresponding one in reference image (Figure 2)
b) CC disconnected/broken into smaller portions and does not align (Figure 3)
c) CC joined with other so does not align (Figure 4)
d) CC is noise in binarized document

a) Reference Image b) Binarized Image

Fig. 2. Aligned CC of Reference and Binarized Image

a) Reference Image b) Binarized Image

Fig. 3. Broken CC in Binarized Image



a) Reference Image b) Binarized Image c) Overlayed Image (a) on (b)

Fig. 4. Joined CC in Binarized Image

All alignments discussed in case (a) above are considered accurate and tabulated. Cases (b) - (d) are erroneous and are discarded. As reference and binarized images may not exactly agree for all pixels, match for each connected component (as discussed in case (a) above) is considered accurate if the bounding boxes are within 8 pixels[1] in *x* and *y* dimensions.

The CCs which are matched through all the three binarization methods are short listed to compare these techniques. For each pair of reference and binarized CC alignment score is computed by (7), using auto-correlation as discussed. The binarization technique yielding the maximum overall alignment score for all the matched CCs across all 300 document images is selected. The results are discussed in Section 4.

IV. Experimental Results

The time of three binarization methods is computed for 300 document images. The results show least time for Otsu method (33ms) whereas Sauvola(67ms) and Bataineh (71ms) methods are comparable.

The alignment score of the three binarization techniques over all connected components of 300 pages for various *k* values of Sauvola method are given in Figure 5.

[1] The threshold of eight pixels is experimentally determined to match the mean size of a *Nuqta* (single dot diacritic) at 14 point size in Nastalique

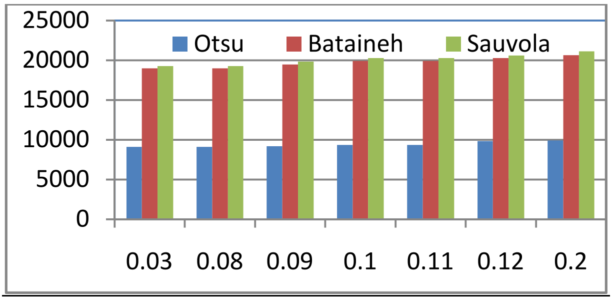
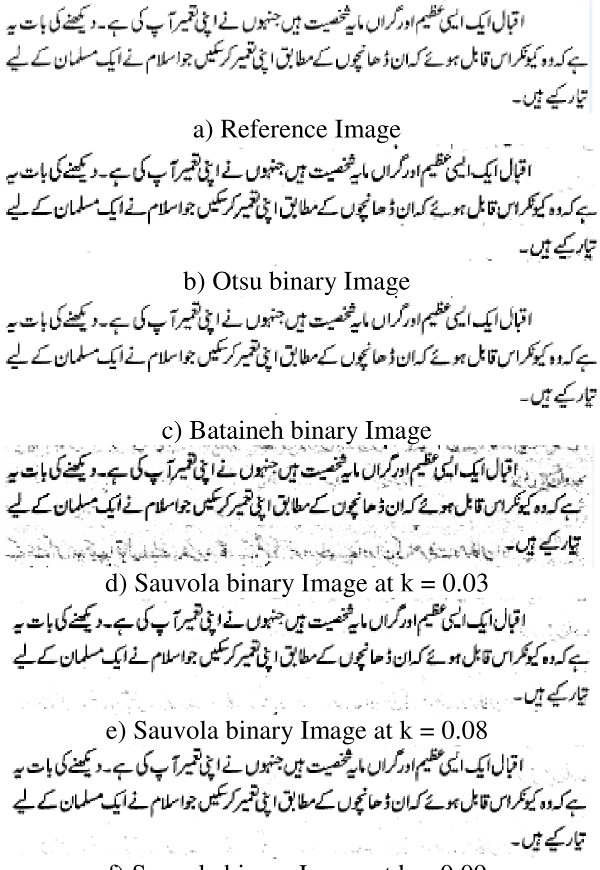


Fig. 5. Matched connected component counts for the three binarization methods for different *k* values

Although Otsu method has significantly less computational time, the counts in Figure 5 show that it is not accurate. Again, both Sauvola and Bataineh perform equally well, though Sauvola method is slightly better in accuracy as well as efficiency for the Urdu document images. Therefore, this method is shortlisted for further processing. Sample images are given in Figure 6 below.



اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

a) Reference Image

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

b) Otsu binary Image

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

c) Bataineh binary Image

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

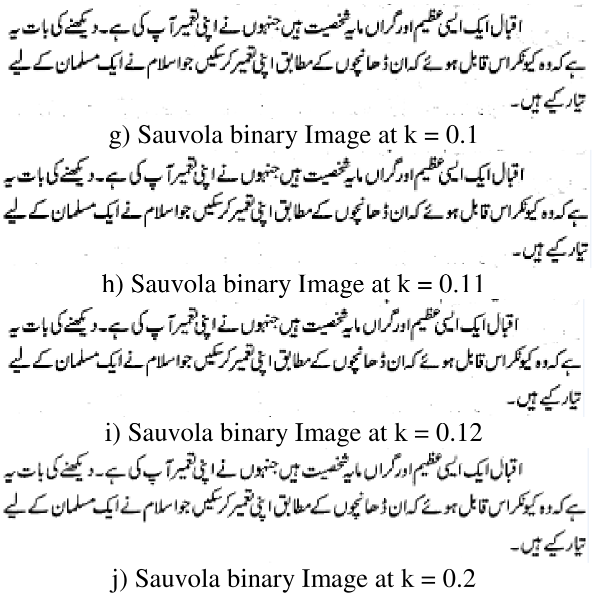
d) Sauvola binary Image at k = 0.03

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

e) Sauvola binary Image at k = 0.08

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

f) Sauvola binary Image at k = 0.09



اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

g) Sauvola binary Image at k = 0.1

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

h) Sauvola binary Image at k = 0.11

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

i) Sauvola binary Image at k = 0.12

اقبال ایک ایسی عظیم اور گراں مایہ شخصیت ہیں جنہوں نے اپنی تعمیرآپ کی ہے۔ دیکھنے کی بات یہ
ہے کہ وہ کیونکر اس قابل ہوئے کہ ان ڈھانچوں کے مطابق اپنی تعمیر کرسکیں جو اسلام نے ایک مسلمان کے لیے
تیار کیے ہیں۔

j) Sauvola binary Image at k = 0.2

Fig. 6. Sample Binarized Images of Various Methods

At *k*=0.03 a clear foreground image is produced but with noisy background. At *k*= 0.08, 0.09, 0.1, 0.11 and 0.12 similar binary images are produced giving best results at 0.1, with clear image and background. At 0.2 the image has clear background but faded foreground.

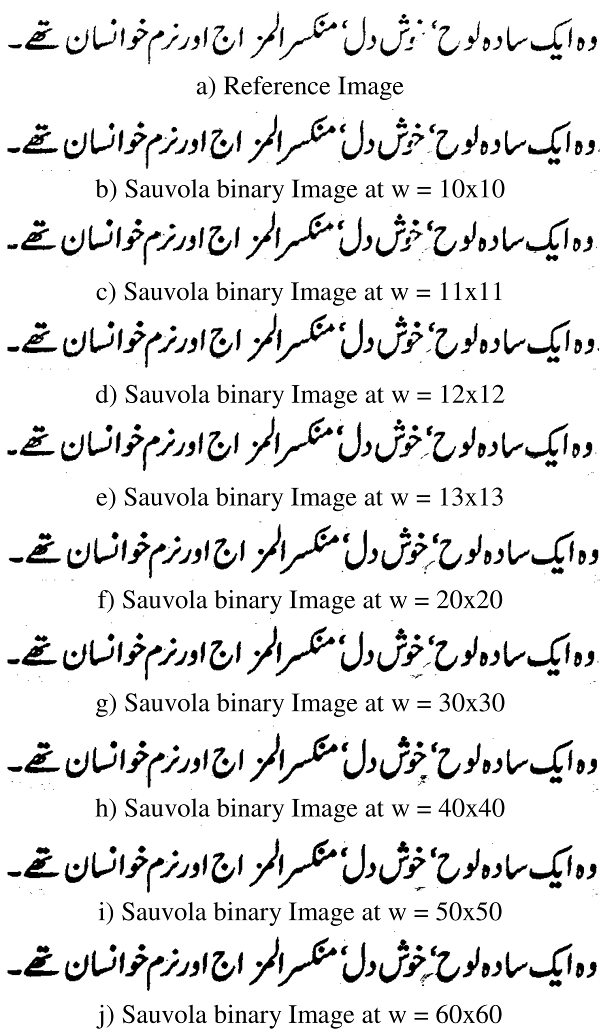
Results in Figure 5 are based on a window of size of 40x40. However, a more detailed analysis shows that Sauvola method causes some undesired diacritic and main body joining. Therefore, the method is further tuned by altering the window sizes. Binarized versions of 10 document images of varying print and paper quality are created using *w*x*w* window size for each of the following *w* values: 10, 20, 30, 40, 50, and 60. Manual inspection of resulting images shows best results around *w*=10. Therefore, further tuning at *w*=11, 12 and 13 is also conducted. Overall results are computed for 251 lines containing 5121 diacritics and 7666 main bodies (a total of 12787 CCs, computed from the parallel text of these pages). Table 1 gives the CCs obtained through binarization at different window sizes and their difference from the reference counts. Minimum difference from the reference is desired. A larger window size creates a thicker image which is better for eventual recognition phase for an optical character recognition system as it reduces disconnections and performs repair of broken images[2]; however it also results in joining diacritics and main bodies. To balance these two constraints, the value of

[2] Repair is visible in Figure 7 for the word خوش

w= 12 is optimal and is selected. Sample output or Urdu at these window sizes is given in Figure 7 below.

TABLE 1. IMPACT ON WINDOW SIZE ON JOINING OF CCs (ACTUAL COUNT = 12787 CCS)

Size	Dia.	MB	Total CCs	Diff.
10	5160	6996	12156	631
11	5150	7002	12152	635
12	*5127*	*7024*	*12151*	*636*
13	5063	6982	12045	742
20	4914	6880	11794	993
30	4739	6840	11579	1208
40	4642	6761	11403	1384
50	4591	6706	11297	1490
60	4544	6630	11174	1613



a) Reference Image

b) Sauvola binary Image at w = 10x10

c) Sauvola binary Image at w = 11x11

d) Sauvola binary Image at w = 12x12

e) Sauvola binary Image at w = 13x13

f) Sauvola binary Image at w = 20x20

g) Sauvola binary Image at w = 30x30

h) Sauvola binary Image at w = 40x40

i) Sauvola binary Image at w = 50x50

j) Sauvola binary Image at w = 60x60

Fig. 7. Binary Image by Sauvola Method at Different Window Sizes

V. CONCLUSION

Sauvola method performs better for Urdu document images, giving optimal results at *k*= 0.1 and a window size of 12x12. The binary image is efficiently produced and the output is appropriate for an optical character recognition system. Bataineh method is comparable; however, Otsu global method does not perform as well over a variety of documents. These parameter settings work well for text areas, but create noise in figures areas. This aspect will be further explored and addressed if needed, e.g. by setting different threshold values for text and figure areas as in [8].

ACKNOWLEDGEMENT

This work has been supported by Urdu Nastalique OCR research project grant by ICTR&D Fund, Ministry of IT, Govt. of Pakistan.

REFERENCES

[1] A. Wali and S. Hussain, “Context Sensitive Shape-Substitution in Nastalique Writing system: Analysis and Formulation,” Proceedings of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE2006), 2006.

[2] B. Bataineh, S. N. H. S. Abdullah, K. Omar, and M. Faidzul, “Adaptive thresholding methods for documents image binarization,” Proceedings of the Third Mexican Conference on Pattern Recognition (MCPR'11), 2011.

[3] F. Shafait, D. Keysers and T. M. Breuel, “Efficient implementation of local adaptive thresholding techniques using integral images,” DRR 2008: 681510

[4] N. Otsu, “A threshold selection method from graylevel histograms,” IEEE Trans. Systems, Man, and Cybernetics, 9(1), 1979, pp.62-66.

[5] W. Niblack, An introduction to Digital Image Processing, Prentice Hall, Englewood Cliffs, 1986.

[6] J. Sauvola, T. Seppanen, S. Haapakoski and M. Pietikainen, “Adaptive Document Binarization,” 4th Int.Conf. on Document Analysis and Recognition (ICDAR’97), Germany, 1997, pp. 147-152.

[7] S. S. Bukhari, F. Shafait, and T. M. Breuel, “Foreground-Background Regions Guided Binarization of Camera-Captured Document Images,” 3rd Int. Workshop on Camera-Based Document Analysis and Recognition (CBDAR’09). Barcelona, Spain, July 2009.

[8] J. Sauvola and M. Pietikainen, “Adaptive document Image binarization,” Pattern Recognition, 33(2), 2000, pp. 225–236.

[9] K. Khurshid, I. Siddiqi, C. Faure, N. Vincent, " Comparison of Niblack inspired Binarization methods for ancient documents," 16th International conference on Document Recognition and Retrieval. SPIE, USA, 2010