IMPROVED ADAPTIVE BINARIZATION TECHNIQUE FOR DOCUMENT IMAGE ANALYSIS

Lal Chandra, Puja Lal, Raju Gupta, Arun Tayal
Newgen Software Technologies Limited, A-6, Satsang Vihar Marg
Qutab Institutional Area, New Delhi-110067, India

Dinesh Ganotra
GGSIP University, Kashmere Gate, Delhi-110006, India

Keywords: Binarization, Thresholding, Gamma Correction, Reverse Video, Contrast Stretch, ICR, OCR.

Abstract: Technology of image capturing devices has graduated from Black & White (B&W) to Color, still majority of document image analysis and extraction functionalities work on B&W documents only. Quality of document images directly scanned as B&W is not good enough for further analysis. Moreover, nowadays documents are getting more and more complex with use of variety of background schemes, color combinations and light text on dark background (reverse video) etc. Hence an efficient binarization algorithm becomes an integral step of preprocessing stage. In proposed algorithm we have modified Adaptive Niblack's Method (Rais et al., 2004) of thresholding to make it more efficient and handle reverse video cases also. The proposed algorithm is fast and invariant of factors involved in thresholding of document images like ambient illumination, contrast stretch and shading effects. We have also used gamma correction before applying the proposed binarization algorithm. This gamma correction is adaptive to brightness of document image and is found from predetermined equation of brightness versus gamma. Based upon result of experiments, an optimal size of window for local binarization scheme is also proposed.

1 INTRODUCTION

Binarization of image is the central part to many image processing applications like form processing, check processing etc. The binarization process computes the threshold value that differentiates objects and background pixels.

A number of methods have already been proposed for image binarization, most of them being specific to their respective applications. The accuracy of these image binarization or thresholding techniques depends on the quality of the scanned image and the nature (such as background, prints, data cell sizes, contrast stretch) of the document. If the input gray image is of good quality then the output of most binarization technique can be used for good analysis of image. However in real time scenario, the quality and nature of input image may not always provide desirable accuracy with most binarization methods.

There are many factors that complicate the thresholding scheme like ambient illumination, variation of gray levels within the object and the background, inadequate contrast etc. In case of noisy background the binarization can become a challenging job.

To improve binarization, (Gonzalez and Richard, 2005) use equation (6) for setting gamma of the image and Adaptive Thresholding Technique (Rais et al., 2004; Niblack, 1990). The basis of binarization method is the calculation of appropriate threshold. A thresholding selection method (Otsu, 1979) suggests minimizing the weighted sum of variances of the objects and background pixels to establish an optimum threshold. A fixed value of threshold cannot give satisfactory results in case of images with illumination variance. In document image analysis, logical and semantic content conservation is needed. Global threshold cannot conserve these kinds of contents. Contrary to global method, local method is an adaptive approach in

Chandra L., Lal P., Gupta R., Tayal A. and Ganotra D. (2007).
IMPROVED ADAPTIVE BINARIZATION TECHNIQUE FOR DOCUMENT IMAGE ANALYSIS.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IFP/IA*, pages 317-321

which a threshold value is determined over a small region. The Niblack's method (Niblack, 1990) uses mean and standard deviation of image to compute threshold over a small region (75 × 75 window). Dynamic Thresholding (Bemsen, 1986) uses mean and standard deviation of the image along with the contrast to compute the threshold. Sauvola et al. (1997) presented a modified Niblack's method (Niblack, 1990), which uses adaptive contribution of standard deviation in determining local threshold. Gray-level Thresholding (Parker, 1991) is done by computing local threshold value by classifying object and background pixels and then using region growing technique to produce the binarized image.

In our proposed algorithm we have modified Adaptive Niblack's Method (Rais et al., 2004) of thresholding to make it more efficient and handle reverse video cases also. The proposed algorithm is fast and invariant of factors involved in thresholding of document images like ambient illumination, contrast stretch and shading effects. We have also used gamma correction before applying the proposed binarization algorithm. Gamma correction is adaptive to brightness of document image and is found from predetermined equation of brightness versus gamma. Based upon result of experiments, an optimal size of window for local binarization scheme is also proposed.

2 NIBLACK'S METHOD

Niblack's method (Niblack, 1990) is a local thresholding method that adapts the threshold according to the local mean and local standard deviation over a specific window size around each pixel location. The local threshold at any pixel (i, j) is calculated by equation (1)

$$T_{i,j} = M_{i,j} + k\sigma_{i,j}^2 \qquad (1)$$

Where $M_{i,j}$ and $\sigma_{i,j}^2$ are the mean and variance of a window in the image respectively. The size of the window depends upon the application. The value of the weight 'k' is used to control and adjust the effect of standard deviation due to changes in object's features. Niblack's algorithm suggests the value of 'k' to be −0.2.

Niblack's algorithm suffers from the problem of local thresholding by providing details in the binarized images that may not be required in processing. Niblack's method uses fixed value of the weight 'k' which is not the optimum value.

3 ADAPTIVE NIBLACK'S METHOD

Adaptive Niblack Method (Rais et al., 2004) offers improvement over original Niblack's method (Niblack, 1990). It not only depends upon image's local statistics characteristics but also considers the global statistics. This algorithm calculates 'k' dynamically for each pixel and thresholding is done using Niblack's method. The normalized difference between global and local mean provides information about the illumination difference for each pixel window with respect to global illumination.

Equation (2) provides a reasonable value for factor 'k', but it fails to adapt to changes in images with different contrast.

$$K_{i,j} = \frac{M - M_{i,j}}{\max(M, M_{i,j})} \qquad (2)$$

Here M is the global mean of the image and $M_{i,j}$ is the local mean computed on each window.

The use of standard deviation of image and local window improves the result. This algorithm uses the interrelation of global and local characteristics and sets the threshold based on the relative change of local and global mean and standard deviation. The effect of standard deviation remains same on different images having different local illumination and contrast.

Adaptive Niblack method uses eq (2) for images which do not have large variation in contrast. For images with large variation in contrast eq (3) is used.

$$K_{i,j} = -0.03\frac{(M\sigma - M_{i,j}\sigma_{i,j})}{\max(M\sigma, M_{i,j}\sigma_{i,j})} \qquad (3)$$

4 PROPOSED ALGORITHM

This section describes the proposed algorithm along with improvements over adaptive Niblack's algorithm. We have used images with 256 gray levels, scanned at 300dpi.

4.1 Image Quality Improvement

Equation (4 and 5) gives an appropriate gamma (Gonzalez and Richard, 2005) for each pixel in the image. We devised these curves by manually adjusting gamma for 150 document images and optimising the Newgen Software Tech. (2005) Find Data Cell (FDC) engine's results. These equations depend on the global mean of the image. In this algorithm we propose two equations for calculating gamma. Though quadratic equation is proposed, however our experimental results demonstrate that linear equation is sufficient for proper binarization. Quadratic equation may be used in cases where linear equation is giving unsatisfactory results. However it may be time consuming. Figure 1 represents the plot of linear and quadratic equation with respect to gamma and mean of the image.

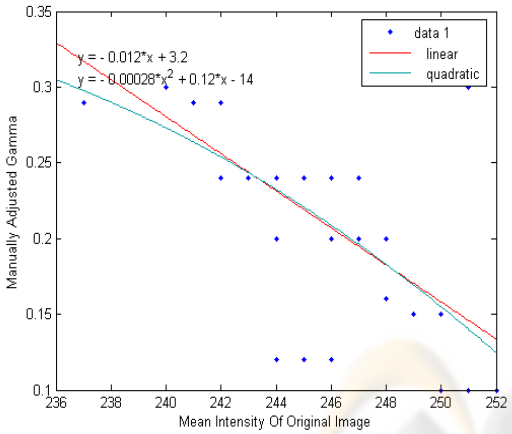


Figure 1: Manually adjusted gamma vs. mean intensity of original images.

$$\gamma = -0.012M + 3.2 \quad (4)$$

$$\gamma = -0.00028M^2 + 0.12M - 14 \quad (5)$$

Here M is mean of original image. We set gamma for all gray pixels of the image. We use

$$I = I^{(1/\gamma)} \quad (6)$$

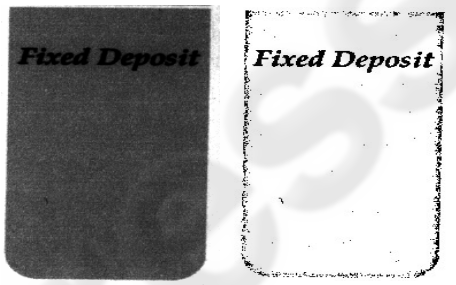
convention for gamma change, where I is the image's intensity.

4.2 Binarization

The proposed binarization algorithm is based on Adaptive Thresholding Technique (Rias et al., 2004). The algorithm is executed in three steps:

1. If the selected pixel's gray value is greater than 240 then saturate the pixel as white (gray value 255) or if the pixel's gray value is less than 30 then saturate the pixel as black (gray value 0).
2. If the condition to execute first step is not satisfied then set a window of size 15×15 and calculate the mean. If the difference between selected pixel's gray value and mean is less than 10 then set the pixel as white. This operation is specific for removing the gray background and preserving the text and other components present over it.
3. For remaining pixels apply adaptive thresholding technique. With the use of 15×15 window, the algorithm requires less computation time without deteriorating performance.

Figure 2(a) shows input image with dark background and figure 2(b) is the output image after processing with proposed algorithm.



(a) (b)

Figure 2: Execution of step 2 for dark gray background image. (a) Dark background image (b) Image after background removal.

4.3 Performance of the Proposed Algorithm

Figure 3 shows the performance of the proposed algorithm over adaptive thresholding technique.

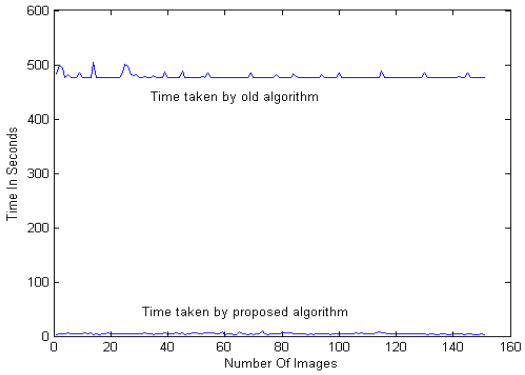


Figure 3: Performance of proposed algorithm.

We executed both algorithms on 150 different images. The images were off-the-shelf available

bank account opening forms. They had large variation in terms of background, prints, data cell sizes, illumination, contrast stretch etc. About 30,000 different cells were selected on them and Newgen Software Tech. (2005) FDC engine was used to detect intelligent character recognition (ICR) cells and optical character recognition (OCR) cells. The comparison of FDC results is given in table 1.

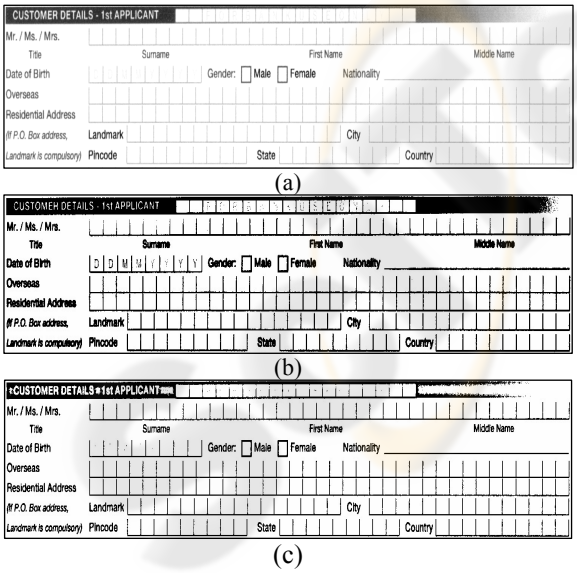
Table 1: Comparison of FDC engine.

Forms with 29089 ICR cells	Detected ICR cells	% Accuracy
Original forms	17742	61.0
Forms with gamma correction	27119	93.0
Forms binarized with proposed algorithm	28592	98.0

4.4 Results and Analysis

Figure 4 and 5 show the comparison of document images with white background and gray background respectively. Experiments were conducted on images with varying background scanned with scanner AVISION830C.

Figure 4(a) is a scanned document image. There is not much difference between print and background gray levels.



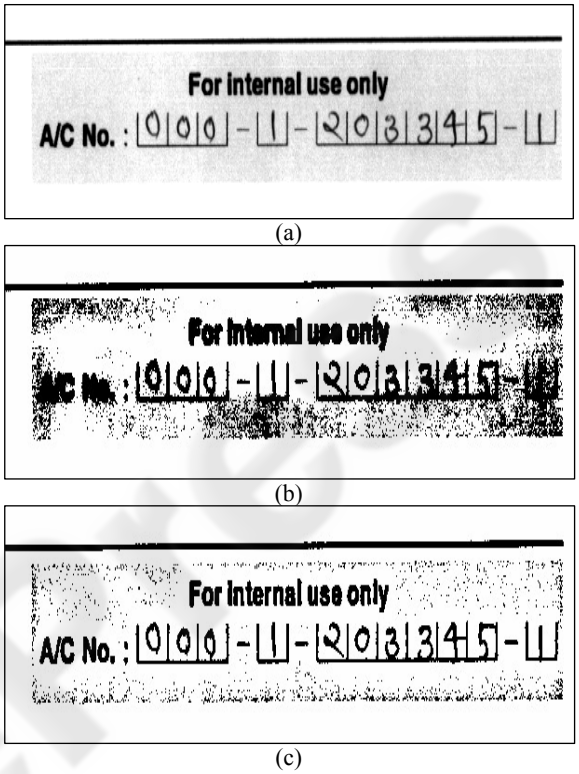
(a)

(b)

(c)

Figure 4: Comparison of output images with light lines and text gray value with respect to background. (a) Original Gray level image (b) Binarized through adaptive thresholding (c) Binarized through proposed Algorithm.

Figure 4(b) is the binarized image through adaptive thresholding technique. The ICR and OCR cells are clear but the text quality is not so good and also the background is still noisy. While as shown in figure 4(c), the output of proposed algorithm, all ICR and OCR cells are clear and the text quality is good as compared to figure 4(b) and the background noise is also removed.



(a)

(b)

(c)

Figure 5: Comparison of output images with non-continuous gray background (a) Gray Background Image (b) Binarized through adaptive thresholding technique (c) Binarized through proposed algorithm.

A document image over gray background is shown in figure 5(a). In this case the gray background is non-uniform. The figure 5(b) is the binarized image through adaptive thresholding technique. Figure 5(c) is the binarized image through proposed algorithm. In this case the image quality is improved over existing algorithm.

4.5 Flowchart of Proposed Algorithm

Flowchart of the algorithm is shown in fig 6.

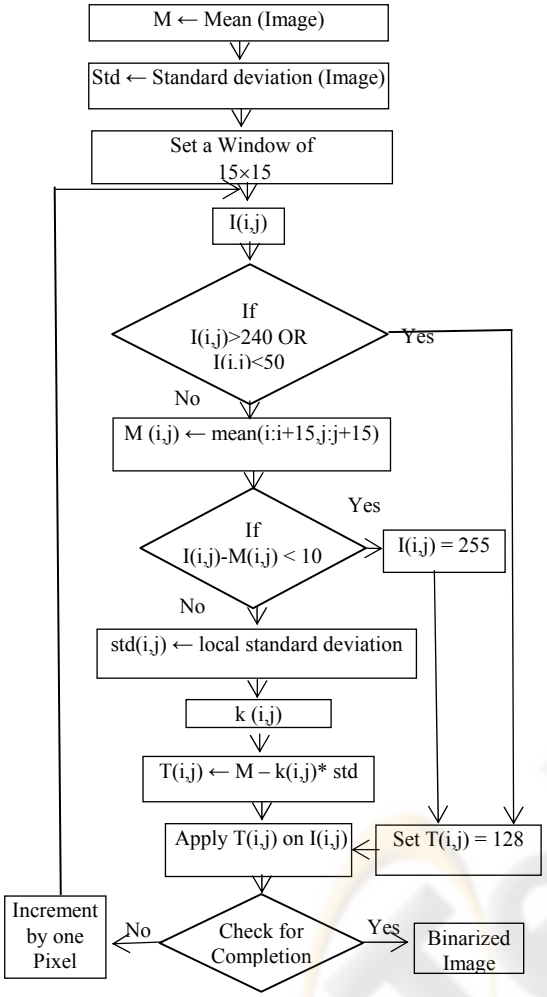


Figure 6: Flow chart of the proposed algorithm.

5 CONCLUSION

In this paper, we have proposed an algorithm of binarization for document image analysis. This algorithm was tested over images with large variation in terms of background, prints, data cell sizes, illumination, contrast stretch etc and it gave satisfactory results. The result shows that it had improved document image binarization significantly over adaptive Niblack's algorithm (Raise et al., 2004) .We have also shown that a local window size of 15 × 15 is the appropriate option for the thresholding scheme as it provides significant time optimization without deteriorating performance. The use of local and global statistics has made our algorithm strong and robust. Though targeted here for document image analysis, it will be a good candidate for other kind of applications as well scene processing and image segmentation.

REFERENCES

Rais, N. B., Hanif, M. S., Taj, I. A., 2004. "Adaptive Thresholding Technique for Document Image Analysis.", *Proc. INMIC 2004, 8th Int. Multitopic conf*, 61-66 *IEEE*.

Gonzalez, R. C., Richard, E. W., 2005. "*Digital Image Processing*", Pearson Prentice-Hall Inc.

Niblack, W., 1990. "*An Introduction to Digital Image Processing*", Prentice-Hall Inc.

Otsu, N, 1979. "A threshold selection method from gray level histograms." *IEEE Tran. on Sys. Man. Cyber.*

Bemsen, J., 1986. "Dynamic thresholding of gray-level images". *Proc. 8th ICPR, Paris*. 1251-1255.

Sauvola, J., Seppanen, T., Haapakoski, R. Pietikainen, M., 1997. "Adaptive Document Binarization." *Int. Conf. Doc. Ana. Rec.*, 147-152.

Parker, J.K., 1991. "Gray level thresholding in badly illuminated images". *IEEE Trans. Patt. Ana. Mac. Intell.* Volume 13, 813-819.

Newgen Software Technologies Ltd. www.newgensoft.com/2005/products/omniextract.htm