VOL. 10, NO. 2, FEBRUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

COMPOUND BINARIZATION FOR DEGRADED DOCUMENT IMAGES

Arwa Mahmoud AL-Khatatneh, Sakinah Ali Pitchay and Musab Kasim Al-qudah
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai, Nilai, Negeri Sembilan, Malaysia
E-Mail: arwa_khatatneh@yahoo.com

ABSTRACT

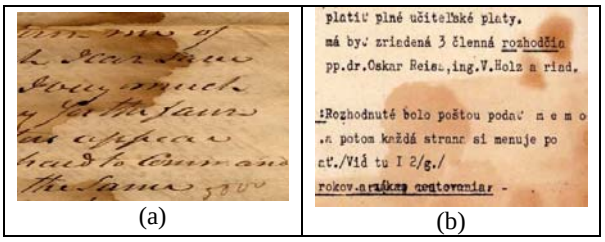
In this paper, we propose a new binarization method for degraded document images. Hence, the existing work is focus on finding a good global or local method in order to remove smear, strains, uneven illumination etc. We propose a new compound method that combines the advantages of both global and local thresholding methods. Our method is applicable for various types of degradation cases and the value of factors could be determined automatically. We compare our method with five state-of-the-art degraded document images. It also has been tested over the dataset that is obtained from the recent Document Image Binarization Contest (DIBCO) 2011 and 2013 for the experiments. Experimental results prove the effectiveness of the proposed technique compared to previous methods.

Keywords: degraded document, binarization, thresholding method.

INTRODUCTION

Text binarization is a crucial step in document image processing, it represent a preprocessing task which automatically converts the document images from a gray-scale or color image into a binary image in way that the background information is represented by white pixels and the foreground by black ones, the thresholding applies to allow the document to be recognized and retrieved more efficiently.

The binarization is a simple task when apply a simple thresholding techniques to good quality image. On the other hand, it has been proved to be a very difficult task, especially in the case of degraded document such as uneven illumination, shadows, low contrast, smears and heavy noise. An example is shown in Figure-1.



(a) (b)

Figuare-1. Example of degraded documents from DIBCO contest (a) handwritten document (b) printed document.

The binarization techniques divided into two categories (Jagroop Kaur and Mahajan, 2014). First, simple technique that used simple thresholding equation and it is easy to implement. There are two main approaches for simple a thresholding technique: (i) global thresholding and (ii) local thresholding. The global approach such as (Otsu, 1979) calculate single threshold for all the pixels in the image to classify them as text or background. They obtained a good result in some cases of document image where it consist a clear image. However, this approaches usually not a suitable approach for the degraded document binarization.

The local approach such in (Niblack, 1986) divides the main image into several window depends on the case study and finds different threshold values. Therefore, threshold varies according to the properties of an image area. Nevertheless, this approach is time-consuming and computationally expensive. Second category is the compound technique which is used hybrid algorithms based on existing techniques. Wherefore, its combine the advantages for both global and local thresholding.

Many works have been reported for the document binarization task in literature. (Kefali *et al*., 2010) work, asset 12 outstanding methods over the old Arabic document images. They used 120 images with different problems and find as a result that Nick and Sauvola method are the best methods. (Stathis *et al*., 2008) have written deep evaluation paper based on 30 well-known binarization techniques including local, global and hybrid algorithms. The methods were tested on 150 degraded document images. (Leedham *et al*., 2002) compare between five binarization algorithms by using the precision and recall analysis of the resultant words in the foreground.

Generally, the choice of a suitable method for each degradation case proved to be a very difficult procedure itself (Jagroop Kaur and Mahajan, 2014). Hence, it affects the following steps in the documents analysis systems. On other hand the evaluation of these algorithms is another difficult tasks where there is no objective way to compare the performance of these algorithms (Leedham *et al*., 2003).

The aim of this work is to propose a compound binarization method. This method find automatically two thresholding equation, the threshold value is selected depending on the comparison between global and local standard deviation to determine the suitable threshold equation. We compare our method with other current such as: Otsu, Niblack, Sauvola, Wolf and Nick methods. Our method produces more accurate binarization results and the findings will be discussed in the experimental result.

The rest of this paper is organized as follow: first we cite the five binarization methods. Afterwards we describe our proposed method. Then, we present our experimental results of proposed and these methods. Finally we present the future work, conclusion and the References.

STATE OF THE ART

In degraded document the quality of image is degraded and less readable. Therefore, there is a need for techniques to separate the foreground from background, for example (Lu and Tan, 2007) uses background subtraction, (Dawoud, 2007) depend on cross section sequence graph analysis, and many other methods. The methods which have higher accuracy are as follow:

Otsu method (1979)

Otsu is one of the famous global methods. It separates the gray-level histogram into two clusters foreground and background and choosing the threshold that minimizes the interclass variance of the threshold. This method gives good result in most cases, in other side it does not influence when the image contain dark or light region which switch the image into black and white. The threshold is given by this equation:

$$\mu_0(T+1) = \frac{\mu_0(T)\cdot\omega_0(T) - N_T}{\omega_0(T+1)} \quad (1)$$

μ: is a class mean
N: gray level

where $\omega_0(T) = \sum_{i=0}^{T-1} P(i)$ (2)

and $\omega_0(T) = \sum_{i=0}^{N-1} P(i)$

$\sigma_B^2(T)$ = The Variance of the Pixels in Background (Below Threshold)
$\sigma_O^2(T)$ = The Variance of the Pixels in Foreground (Above Threshold) (3)

Niblack method (1986)

Local method that calculates the local threshold based on local mean and local standard deviation. This method used for all kind of images, but it does not able to remove the unimportant details during the processing especially in a blank area. The threshold is given by:

$$T = m + (-0.2 * S) \quad (4)$$

where m is the mean and S is the standard deviation.

Sauvola method (2000):

This method is an improvement of Niblacks method by solving the problem of presence of a large amount of noise in the background areas. This method computes the threshold using the dynamic range of image gray-value standard deviation. On other hand this method faces some problems when the text pixel value close to foreground image. The threshold is given by:

$$T = m * \left(1 - K * \left(1 - \frac{S}{R}\right)\right) \quad (5)$$

where R is the dynamic of the standard deviation (R=128), and the parameter K takes positives values (K=0.5).

Wolf method (2002)

In this method, they normalize the contrast, the mean gray value of the image and then compute the threshold in order to address the (Sauvola, 2000) issues. The threshold is given by:

$$T = (1-k)*m + k*M + k*\frac{s}{R}(m-M) \quad (6)$$

where k is fixed to 0.5, M is the minimum gray-level of the image and R is the maximum standard deviation of gray-level obtained over all windows.

Nick method (2009)

This method is an improvement of Niblack method where it works very well for most degraded document (Khurshid et al., 2009), it solves the problem of noise in white pages and low contrast problem. Nevertheless, this method faces the problem of small noisy that cross in windows. The threshold T is calculated as follow:

$$T = m + k * \sqrt{\frac{\left(\sum_{i=1}^{NP} P_i^2\right) - (m^2)}{NP}} \quad (7)$$

where k fixed to -0.1 and -0.2 depend on the application, m: the average gray-level, pi: the gray-level of pixel i and NP is the total number of pixels for slide windows.

PROPOSED METHOD

Based on the previous works, the global binarization methods (Jagroop Kaur and Mahajan, 2014) give poor result when it deal with degraded document and unable to remove noise that it does not distributed in the entire image. On the other hand, the local binarization methods (Sauvola and Pietikäinen, 2000), (Gatos, Pratikakis and Perantonis, 2006) produce bad results with complex noise or a low quality. Hence, a simple binarization method produces a poor result for degraded document case. Therefore, there is a need to use the compound method based on global and local methods to overcome this problem.

A compound method is fast like global thresholding, while providing high quality binarization results like the local thresholding algorithms.

The diagram of the proposed compound binarization algorithm is illustrated in Figure-2.

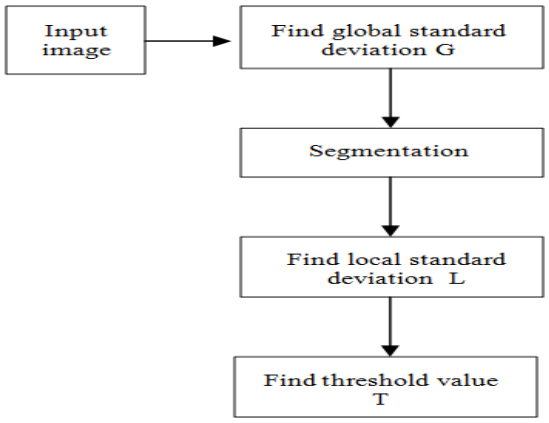


Figure-2. The flowchart of proposed binarization method.

In our proposed method, we combine the global and local thresholding techniques. After the algorithm reads the image, we compute the global standard deviation (G) using mean for all pixels in that image. For whiter images, the value of mean (m) will be set high (near to 255). A low standard deviation indicates that the pixel value tends to be very close to the mean.

The second step is segmenting the image into windows of size 40 x 40. Then, find the standard deviation for each window in order to obtain a sharper result. For that reason, we have a local standard deviation which depends on the value of pixels into each window.

The threshold value is depending on the comparison between global and local standard deviation to determine the suitable threshold equation. Our method aims to solve the problem of degraded image in term of each segment. Each window tells us if it's light or dark with respect to the rest of the image. When the standard deviation of image greater than the local standard deviation this is mean that the pixels value is near black, while when the greater less than local standard deviation this is mean that the pixels near white. Therefore, we propose two thresholding equations for each case.

The procedure of proposed method

The Pseudo code for our proposed method is given as follows:

```
Find global standard deviation (G)
Segment the image into windows
For each window
    Find local standard deviation (L)
        If (G) ≥ (L)   Then
            T = m * (1 - k * (1 - (Abs (L - G) / 128)))
        Else
            T = m – f * Sqrt ( L^ 3 +m ^ 2)
```

Such as k and f are a parameters used for determining the number of edge pixels considered as object pixels, and takes a fixed value 0.5 for k and 0.7 for f which produces best results. Hence, some method like Nick method did not have one value for k. Therefore, it depends on application that is used.

In this way, we get a binarization algorithm that can deal with document images suffering from uneven illumination, shadows, low contrast, smears and heavy noise.

EXPERIMENTAL RESULT

The evaluation of the proposed approach was based on visual criteria by estimate the quality of document images. The objective evaluation used evaluation measurement for the resulted image from the proposed method. On the other hand, the method will be tested in OCR (Optical Character Recognition) application to test the readability of proposed method in degraded documents in near future.

The experiment was performed using both printed and hand-written documents. The different between these two styles is in typing and structured, the hand-written had different behavior of typing. The used dataset include degraded document, we adopt DIBCO contest 2011 and 2013. A dataset contain 32 benchmark 16 from DIBCO 2011 (8 for each printed and hand written) (Gatos, Ntirogiannis and Pratikakis, 2011) and 16 from DIBCO 2013 (8 for each printed and hand written) (Pratikakis, Gatos and Ntirogiannis (2013).

The evaluation measurement applied to compare between the performances of document images. In order to evaluate our method and previous five binarization methods we use "F-measure" and Peak Signal-to-Noise-Ratio (PSNR), which is consider as traditional measures of image quality. The following equations are described F-Measure:

$$F-Measure = \frac{2 * recall * Precision}{recall + Precision} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

where true is a positive TP (pixels that were black in the ground truth image and in the binarize image), false is a positive FP (white in the ground truth image and black in the binarize image) and false - negative FN (black in the ground truth image, but white in the binarize image).

PSNR is a measurement of how close is an image to another higher value of PSNR meant higher similarity of the two images. We consider that the difference between foreground and background equals to MAX values of best ground truth image. The MAX is the maximum possible pixel value of the image. The following equations described PSNR and Mean Squared Error (MSE) where I is a ground truth Image and I' is image result acquired from binarization methods.

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \quad (11)$$

$$MSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[I(x,y) - I'(x,y)]^2}{MN} \quad (12)$$

RESULTS AND DISCUSSIONS

In our experiments, we found that the proposed method is simple and achieved the best performance, its can suppress the other five methods by produces a higher precision and demonstrated superior performance. Based on visual criteria as shown in Figure-3, we observe that our method extracts the text properly from document images that suffer from different types of document degradation.

Table-1 presents the evaluation measurements using F-measure and PSNR measure. A higher quality binarize image has higher F-Measure and PSNR, the table shows that our proposed method achieved the highest performance in general. The evaluation results are summarized as follows:

- Otsu's global thresholding method misses some text while classifies dark background pixels as the text pixels improperly.
- Niblack's method segments the characters very well, but also the area without text contain large amount of noise.
- Sauvola's algorithm overcomes Niblack method problem but have another problem with thinner characters and holes,
- Wolf solve Sauvola problem but the image still have a noise
- Nick method works very well for most degraded document but have a problem of low contrast.

Table-1. Evaluation result for Niblak, Wolf, Nick, Sauvola, Otsu and the proposed method.

Evaluation Metrics	Niblack	Wolf	Nick	Sauvola	Otsu	Proposed
Printed (F-measure)	0.4516	0.73615	0.81845	0.79345	0.85065	0.87605
Hand-written (F measure)	0.2898	0.6333	0.7821	0.72365	0.72925	0.7963
Printed (PSNR)	6.2459	12.6317	14.82783	14.8209	15.9036	15.93535
Hand-written (PSNR)	5.508	13.3015	16.6795	16.39	15.896	16.5565

We can conclude from Table-1 that for F-measure, our proposed method performs better in term of printed and hand written documents. This means that the proposed method preserves the text stroke contour better and produces a higher precision as shown in the Figure-3 the text is clear in document image while other methods produce certain amount of noise. The global method Otsu method is second method after our method in performance in term of printed documents for both F measure and PSNR measure. Nick method out form our method in PSNR measure for hand-written document and the second method after our method for printer F measure. In general the other method performance order is Sauvola, Nick then Niblack method.

Figure-3. Binarization results of the degraded document image from DIBCO contest for the state of art methods and our method. The state of art methods and our method.

FUTURE WORK

In order to make the proposed binarization algorithm more effective and increase the clarity of text we try to detect the problematic area in the document image based on certain detection condition as future work. In other hand we can apply post processing step to algorithm which will increases the performance and quality of overall document image.

CONCLUSIONS

In this paper, a compound binarization technique has been presented that combines the advantages of global and local binarization for degraded document images binarization. The proposed method depends on segment the image into windows and finds the standard deviation for each one and for the whole image. Then, the threshold is calculated according for each window with respect to the rest of the image. The proposed method has been tested on the dataset that is used in the DIBCO (2011, 2013) contest. Experiments show that the proposed method produces a good result in term of the F-measure and PSNR. The recovered image using different existing methods are represented and also to support the proposed method. The binarization results using our proposed method enable us to recover the degraded document images by removing the unwanted background than the existing methods would, and this can be applied, e.g. in typing and handwritten applications.

ACKNOWLEDGEMENTS

The authors wish to thank Universiti Sains Islam Malaysia (USIM) for the support financially and facilities provided.

REFERENCES

Dawoud A. 2007. Iterative cross section sequence graph for handwritten character segmentation. Image Processing, IEEE Transactions on. 16(8): 2150-2154.

Gatos B., Ntirogiannis K. and Pratikakis I. 2011. DIBCO 2011: Document Image Binarisation Contest. International Journal on Document Analysis and Recognition. 14(1): 35-44.

Gatos B., Pratikakis I. and Perantonis S. J. 2006. Adaptive degraded document image binarization. Pattern recognition. 39(3): 317-327.

Jagroop Kaur D. and Mahajan R. 2014. A Review of Degraded Document Image Binarization Techniques. Changes. 3(5).

Kefali A., Sari T. and Sellami M. 2010. Evaluation of several binarization techniques for old Arabic documents images. In The First International Symposium on Modeling and Implementing Complex Systems MISC. 1: 88-99.

Khurshid K., Siddiqi I., Faure C. and Vincent N. 2009. Comparison of Niblack inspired Binarization methods for ancient documents. In IS&T/SPIE Electronic Imaging (pp. 72470U-72470U). International Society for Optics and Photonics.

Leedham G., Varma S., Patankar A. and Govindaraju V. 2002. Separating text and background in degraded document images-a comparison of global thresholding techniques for multi-stage thresholding. In Frontiers in Handwriting Recognition, 2002. Proceedings. IEEE 8th International Workshop on. pp. 244-249.

Leedham G., Yan C., Takru K., Tan J. H. N. and Mian L. 2003, August. Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In: 12th International Conference on Document Analysis and Recognition. IEEE Computer Society. 2: 859-859.

Lu S. and Tan C. L. 2007, September. Binarization of badly illuminated document images through shading estimation and compensation. In: Document Analysis and Recognition. ICDAR 2007. IEEE 9th International Conference on. 1: 312-316.

Niblack W. 1986. An Introduction to Image Processing, Prentice-Hall, Englewood Cliffs, NJ. Strandberg Publishing Company.

Otsu N. 1979. A threshold selection method from gray-level histogram. IEEE Transactions on Systems, Man and Cybernetics. 9(1): 62-66.

Pratikakis I., Gatos B. and Ntirogiannis K. 2013, August. ICDAR 2013 document image binarization contest (DIBCO 2013). In: Document Analysis and Recognition (ICDAR). IEEE 12th International Conference on. pp. 1471-1476.

Sauvola J. and Pietikäinen M. 2000. Adaptive document image binarization. Pattern recognition. 33(2): 225-236.

Stathis P., Kavallieratou E. and Papamarkos N. 2008. An Evaluation Technique for Binarization Algorithms. J. UCS. 14(18): 3011-3030.

Wolf C. and Jolion J. M. 2003. Extraction and recognition of artificial text in multimedia documents. Formal Pattern Analysis and Applications. 6(4): 309-326.