International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072

Binarization of Degraded Text documents and Palm Leaf Manuscripts

A Venkata Srinivasa Rao[1], S Sanjay Pratap[2], V S Subrahmanyam[3], M Harshini[4]

[1] *Department of ECE, Sasi Institute of Technology & Engineering, Tadepalligudem, W.G.Dist, India.*
[2,3,4]*UG Student, Department of ECE, Sasi Institute of Technology & Engineering, Tadepalligudem, W.G.Dist, India*

Abstract - Document deterioration Image processing is a specialized field that has recently drawn an influx of researchers. The first step in preparing a document for further processing is binarization. Depending on the degree of degradation of the original document, global or local thresholding methods are preferred. The threshold phenomenon is a useful technique for identifying the cluster of pixels that are most likely associated with background information while simultaneously separating the object information. In this paper, we propose a technique based on the complement of the original image. This technique is used to evaluate ancient text documents and palm leaf manuscripts, and the quality of the resulting images is assessed using various qualitative metrics.

***Key Words*: Binarization, Complementary of the image, Image documentation, Threshold, local method**

1. INTRODUCTION

Researchers faced numerous challenges as a result of degraded document image analysis. Degraded conditions of historical documents (e.g., bleed-through, ink stains, torn pages, etc.) prompted researchers to develop binarization and enhancement algorithms that are appropriate for these challenges. Binarization is the first stage of pre-processing in all image processing and analysis systems. The majority of the cultural heritage document collection consists of digitised images. These priceless documents are available for manual annotation on the Internet in order to make their content more accessible. Light, particularly ultraviolet light, which is present in daylight, can damage documents. Having original documents at home or visiting a local archive or history centre allows us to handle old material with historical evidence, which comes at a cost.Frequent handling of these original documents causes physical wear and tear, eventually leading to document loss. Furthermore, changes in environment and light can harm the documents. In this context, image analysis for removing background noise and improving document readability requires distinguishing between ancient and modern documents, as well as processing.

Binarization techniques based on a global or local threshold are straightforward and practical. The Global threshold specifies a global value for all pixel intensities in an image in order to distinguish them as text object or background [1]. This method fails to remove image noise that is not distributed uniformly. In contrast, local threshold provides an adaptive solution for images with varying background intensities, with the threshold varying according to the properties of the local region [6,8]. There are a plethora of general-purpose Binarization methods available that can handle any document image with a complex background. These methods are categorized as either local or adaptive thresholding. Bernson proposed[2] a neighborhood-based local threshold, Niblack evaluated the threshold at each pixel using local mean and standard deviation, and Sauvola used two algorithms[4] to calculate a different threshold for each pixel. The authors [9] all proposed a fast entropy-based segmentation method for producing high-quality binarized images of documents with back-to-back interference. Xiao et al. proposed an entropic thresholding algorithm based on gray-level spatial correlation (GLSC) histograms [12]. They revised and expanded on Kapur et al algorithm. Syed Saqib Bukhari et al. proposed [11] a local binarization method adaptation that uses two distinct sets of free parameter values for the foreground and background regions. They show how to estimate foreground regions in a document image using ridge detection. Using a different set of free parameter values, this information is then used to calculate an appropriate threshold for the foreground and background regions. Chien-Hsing Chou an et al. [13] proposed a method for segmenting an image and determining how to binarize each segment. The decision rules are the result of a learning process that starts with training images. Rachid Hedjam.A et al. [14] proposed an adaptive method based on maximum-likelihood(ML) classification that uses apriori information and spatial relationships on the image domain in addition to main data to recover weak text and strokes. Mehmet Sezgin et al. proposed[5] a comprehensive assessment of existing local infrastructure as well as global thresholding methods. Mitianoudis and Papamarkos [18] proposed a three-stage approach to document image binarization. The background was removed in the first stage using an iterative median filter, then misclassified pixels were separated in the second stage using local co-occurrence mapping (LCM) and Gaussian mixture clustering, and finally, morphological operators were used to identify and suppress misclassified pixels for better classification of text and background image pixels. Khan and Mollah [19] proposed a binarization technique that involved first removing background noise and improving document quality, then using a variant of Sauvola's Binarization method, and finally performing post-processing to find small areas of connected components in an image and

removing unnecessary components. A more recent study [20] used a hierarchical deep supervised network (DSN) to binarize document images. The network architecture is split into two sections. The first section of the network distinguishes between foreground and noisy background using high-level features, while the second section preserves foreground (text) information while dealing with noisy background.To accomplish this, the network is designed in such a way that different level features are used to preserve high details of the foreground. The proposed network is made up of three distinct DSNs, the first of which has a small number of convolution layers to generate the low-level feature maps. The second DSN is a slightly deeper structure for producing mid-level feature maps, while the final DSN is a deep structure for producing high-level feature maps. The experimental results on three public datasets show that the proposed model completely outperforms the state-of-the-art binarization algorithms. Westphal et al. [21] proposed document image binarization using a recurrent neural network. To incorporate contextual information into each step of the binarization process, they used grid LSM cells to handle multidimensional input in this method. They were able to accomplish this by dividing the input image into 64 64 pixel non-overlapping blocks. These blocks serve as the input sequence for the RNN model. The output of four distinct grid LSTM layers is combined and aligned to produce the mid-level feature sequence (L1). The mid-level features are then fed into two different grid LSTM layers (L2), which are combined with a bi-directional LSTM to produce the high-level feature map. After that, a full connection layer (L3) was applied to the high-level feature map to produce the binarization result. The experimental results show that the quality of binarization has significantly improved.

The average value of the image is used as a threshold in the algorithm in this paper to propose a general technique for cleaning degraded documents. One method for distinguishing object information from background noise is to compute a global threshold of intensity value that can distinguish two clusters. We used an algorithm based on the average value of the image. It works best with documents with non-uniform noise distribution.

There are four sections to the paper. The first section provided a brief overview of the introduction, literature review, and problem definition. The second section discussed the algorithm for cleaning the noisy documents. The third section discussed the experimental results. Section four discusses the findings and the scope of future work.

2. METHODOLOGY

We present (Fig-1) a technique for binarizing ancient degraded documents and palm leaf manuscripts that falls under the clustering of pixels from an image's background and foreground. Grey level data is subjected to a clustering analysis in this class of technique, with the number of clusters always set to two. These two clusters correspond to the two peaks of the histogram. This method computes the average pixel value that can be used as a threshold. The flowchart of the proposed model, shown in Fig-1, will explain all of the algorithm's details.

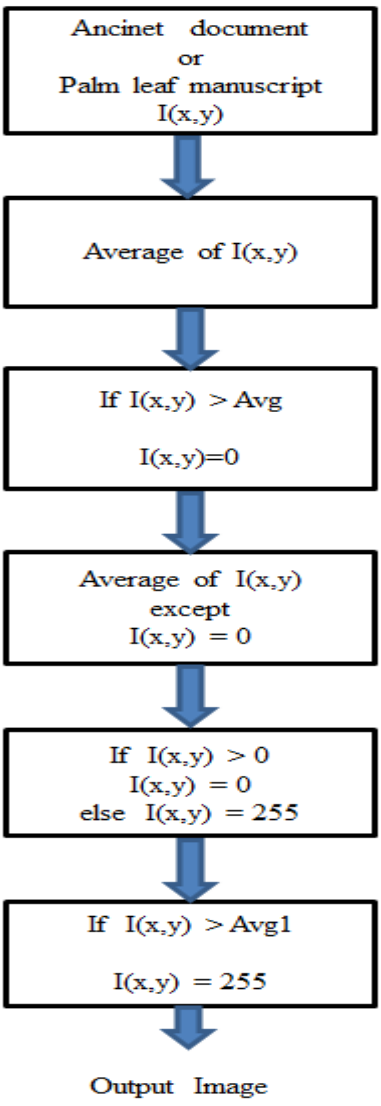


Fig.1 flowchart of the proposed model

The traditional approach in a Global thresholding technique[15] is to find a unique threshold to eliminate all pixels representing image background while preserving others as image foreground. Many real-world images have complicated backgrounds or poor image foregrounds (some foreground pixels have grey values very close to some background pixels). It is difficult to find a single threshold that can completely separate the object information from the background in such cases. The same is true for local thresholding, which determines threshold values locally, such as pixel by pixel or region by region. In the proposed algorithm, image equalisation is performed after each thresholding operation while evaluating the relative

importance of a respective pixel intensity toward background. The new threshold is computed to perform background elimination. This procedure will be carried out until the sensitive threshold is reached.

2.1. BINARIZATION ALGORITHM

The Binarization algorithm suitable for noisy documents requires a series of sequential steps, which are as follows:

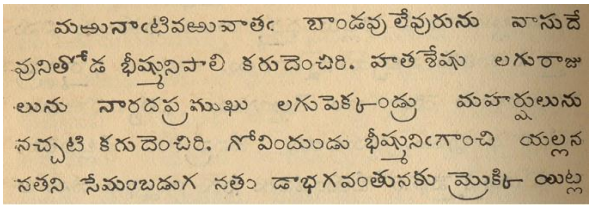
1. Document extraction from a degraded (noisy) state I (x,y)
2. Determine the average intensity Avg of the degraded document (x,y)
3. Using the average value, set the pixels to black;
4. Calculate the degraded document's average value Avg1, excluding the black pixels.
5. Use the image's second average to set the pixels to white (Avg1).

I stand in for the given image (x,y). Where x and y are the horizontal and vertical coordinates of the image, respectively, and I(x,y) can take any value between 0 and 1, with I=1 representing white and I=0 representing black. The proposed algorithm necessitates shifting the image's intermediate tones to the background. In general, any document image contains few useful pixels (foreground) in comparison to the image size (foreground+ background). Object information is typically less than 10% of the total pixels in the document. Taking advantage of this advantage, it was assumed that the background would determine the average value of the pixels, even if the document was quite clear.

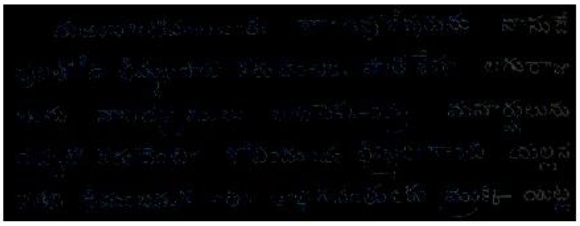
There is no proper differentiation between foreground and background because the pixel values in the degraded documents are not uniform. Some pixels from the foreground and background will occupy the same region in the image histogram. As a result, the proposed algorithm will distinguish pixels between the foreground and background regions. The threshold value of the image is critical in this segregation. In this case, the average value serves as the threshold for distinguishing the object information from the background information in the document. So, to begin, we'll compute the average value of the degraded document. Values that exceed the threshold value are set to black, which equals zero. Recreate the image using the complementary threshold value to the original image. Again, we must compute the average of the image while disregarding the black pixels. This will allow for proper separation of the image's foreground and background. Apply this second average to the image as a threshold value; values above the threshold value are set to white, indicating one. As a result of this action, the noise in the image's background is removed..

3. RESULTS AND DISCUSSIONS

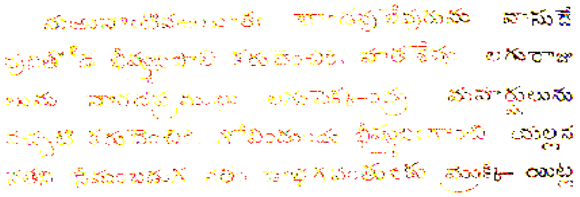
This algorithm is tested on a set of 30 document images. They are 50 to 60 years old and gathered from the Internet (a Telugu old book titled "Thiagarajaswami Krithis" was published in 1933 at Kesari Printing Press in Chennapuri) and scanned copies of old story books (a Telugu old book titled "vydula kathalu" was published in 1942 at Madras Printing Press). Figure 2(a) shows an example of a noisy document with non-uniformly distributed noise. The background of the noisy image is golden brown in colour. As shown in Fig.2, it is first converted into a complementary image, which means the background is black and the foreground is white (b). As shown in Fig. 2, the complemented image is then transformed into a noise-free image.



(a)



(b)

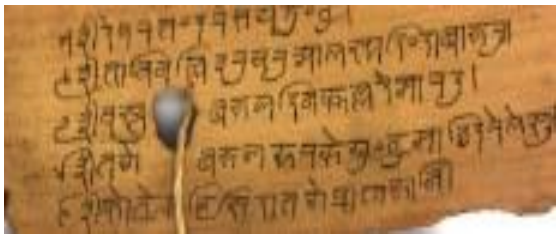


(c)

Fig 2. (a) Degraded Telugu text sample (b) Complementary of image (a) (c) Resultant image

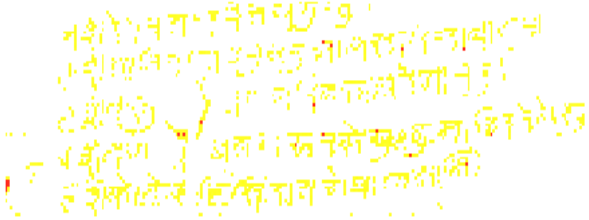
The algorithm is then applied to over 100-year-old palm leaf manuscripts collected from the Internet in the second phase. Nearly 30 palm leaf samples are tested using this algorithm. Figure 3 shows an example of a palm leaf manuscript (a). This is the typical case for testing palm leaf manuscripts with this algorithm because the noise concentration varies from

one area to another within the sample. As a result, it is difficult to distinguish the background from the foreground. In this case, the image's background is also golden brown. Figure 3 shows the complement to the original image (b). After applying the defined algorithm to the complemented image, the resultant image is displayed in Fig 3.(c).



(a)

(b)



(c)

Fig 3. (a) Palm leaf manuscript sample (b) Complementary of image (a),(c) Resultant image

4. PERFORMANCE EVALUATION

Three metrics are used to assess the algorithm's performance: Peak Signal to Noise Ratio, Mean Square Error, and Average Difference. These parameters are evaluated in order to compare the proposed algorithm's performance to that of the Otus method. Palm leaf manuscripts and antique text documents. When compared to the Otus method, the proposed algorithm's metrics show a significant improvement. The PSNR quantity increases in both graphs (Figs 4&5). This algorithm works best with older text documents. It needs to be improved if palm leaf manuscripts are to produce better results.

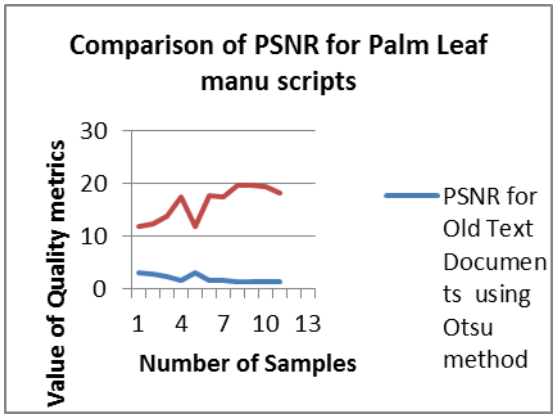


Fig 4: Comparison of PSNR for palm leaf manuscripts

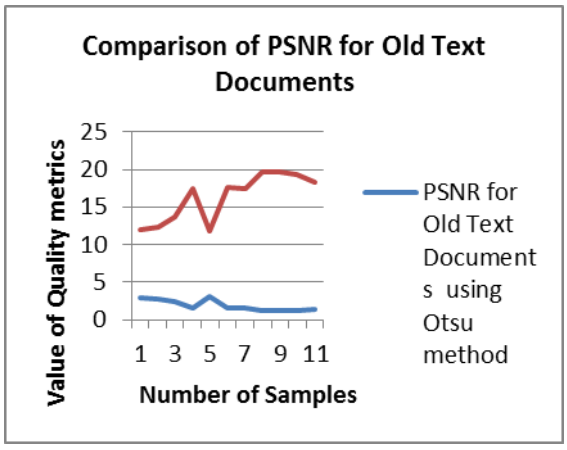


Fig 5: Comparison of PSNR for old text documents

5. CONCLUSION

The current work proposes using the average value of the image as a threshold. The image of the document under test is binarized using the average value of the image. The first average value of the image is used in this algorithm to complement the image and avoid unnecessary background information. The second average removes any residual background noise from the first average value. During the cleaning process, some text information and background noise are removed. This algorithm has been shown to be effective on low-noise historical document images as well as palm leaf manuscripts. However, it has been discovered that palm leaf manuscripts require further improvement.

6. FUTURE SCOPE OF THE WORK

The procedure described above could be extended to noise-free samples that are manually contaminated with various noises at various levels, such as pepper, Gaussian noise, and so on. The noisy documents are then cleaned using a defined algorithm and other methods, and quantitative metrics for identifying information loss during the cleaning process are established.

REFERENCES

[1] N.Otsu, "A threshold selection method based on grey level histograms," IEEE Transactions on Systems, Man, and Cybernet., 9(1), 1979, pp.62-66.

[2] W. Niblack " An Introduction to Digital Image Processing", Prentice Hall, 1986, pp. 115-116

[3] J.Sauvola, M.Pietikainen, " Adaptive Document Image Binarization ," Pattern Recognition, 33, 2000, pp.225-236

[4] Mehmet Sezgin, Bulent Sankur, " Survey over image thresholding techniques and quantitative performance evaluation," 146 / *Journal of electronic Imaging/* January 2004/ vol 13(1)

[5] E.Kavallieratou, "ABinarization Algorithm Specialized on document images and photos," 8th Int. Conf. on Document Analysis and Recognition, 2005, pp. 463-467.

[6] George D.C. Cavalcanti, Edducardo F. A. Silva "A Heuristic Binarization Algorithm for Documents with Complex Background",1-4244-0481, 2006 IEEE

[7] B.Gatos, I. Pratikakis, and S.J.Perantoni, " Adaptive degraded document image binarization," Pattern recognition, vol.39, pp.317- 327, 2006

[8] João Marcelo Monte da Silva, Rafael Dueire Lins, Fernando Mário Junqueira Martins, Rosita Wachenchauzer, "A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference," Journal of Universal Computer Science, vol. 14, no. 2 (2008), 299-313

[9] Xiao. Y, Cao. Z.G, and Zhang, "Entropic thresholding based on gray-level spatial correlation histogram," Proc. 19th Int. Conf. On Pattern Recognition (ICPR 2008), Tampa, FL, USA, 8-11 December 2008, pp. 1-4

[10] Syed Saqib Bukhari, Faisal Shafait, Thomas M. Breuel, " Adaptive Binarization of Unconstrained Hand-Held Camera-Captured Document Images," Journal of Universal Computer Science, vol. 15, no. 18 (2009), 3343-3363

[11] A.V.S.Rao, Tinnati Sreenivasu, N.V.Rao, T.S.K.Prabhu, A.S.C.S.Sastry, L.P.Reddy,"Binarization of Dcuments with complex Backgrounds," Procedings of International Conference on Machine Vision, 978-0-7695-3944-7/10, 2010.

[12] Chien-Hsing Chou. a, Wen-Hsiung Lin .b, Fu Chang .b.Ã, " A binarization method with learning- built rules for document images produced by cameras," Pattern Recognition 43 (2010) 1518–1530

[13] Rachid Hedjam Ã, Reza Farrahi Moghaddam, Mohamed Cheriet, " A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," Pattern Recognition(Elsevier), 2011.

[14] Maythapolnun Athimethphat , "A Review on Global Binarization Algorithms for Degraded Document Images " AU J.T. 14(3): 188-195 (Jan. 2011)

[15] Ntirogiannis .K, Gatos .B, Pratikakis .I, "Performance Evaluation Methodology for Historical Document Image Binarization", IEEE Transactions on Image Processing, Vol22, No.2, PP595-609, 2013.

[16] Sehad, Abdenour, "Ancient degraded document image binarization based on texture features." 8th International Symposium on Image and Signal Processing and Analysis (ISPA), PP189-193, 4-6 September 2013.

[17] Mitianoudis N., Papamarkos N. Document image binarization using local features and Gaussian mixture modeling. *Image Vis. Comput.* 2015;38:33–51.

[18] Pratikakis I., Zagoris K., Barlas G., Gatos B. ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016); Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR); Shenzhen, China. 23–26 October 2016; pp. 619–623.

[19] Vo Q.N., Kim S.H., Yang H.J., Lee G. Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition.* 2018;74:568–586.

[20] Westphal F., Lavesson N., Grahn H. Document image binarization using recurrent neural networks; Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS); Vienna, Austria. 24–27 April 2018; pp. 263–268.