 ISBN: 978-93-86005-02-1

An Improved Method for Document Image Binarization

Nilima Paul* and Harinandan Tunga

Department of Computer Science and Engineering, RCC Institute of Information Technology, Kolkata - 700015, West Bengal, India; mnpaul45@gmail.com

Abstract

Handwriting analysis of document image has four parts- preprocessing, segmentation, feature extraction and classification. Image pre-processing technique is used to improve the quality of the image for easily and efficiently processing in future steps. Principal stage of image pre-processing is binarization, according to which the pixels are classified into text and background. It is a crucial stage that can affect further stages including the final character recognition stage. This paper proposed a binarization technique which is based on Otsu which has been already used for handwriting document binarization. But in order to tolerate badly degraded document images, present work proposed a binarization technique with the help of Otsu algorithm, which can segment the foreground from the background if text document is badly degraded, such as uneven illumination, image contrast variation, bleeding-through, and smear. The proposed method was tested on text image of H-DIBCO2012 and DIBCO2009. Experimental results show that proposed technique achieved a high precision that gives better result than the Otsu algorithm.

Keywords: Binarization, Gray Scale Image, Line Segment, Otsu, Threshold

1. Introduction

In the area of image processing, Document Image Binarization is performed in the preprocessing stage for document analysis and it aims to segment the foreground text from the document background. A fast and accurate document image binarization technique is important for the ensuing document image processing tasks such as Optical Character Recognition (OCR). Many techniques have been proposed so far for document binarization as shown in literature survey. All these methods are having their own advantages and disadvantages. It has been concluded from the existing research is that no technique is perfect for every case. Therefore still some research is required in this field of image binarization. The proposed work tried to implement a new method for Document Image Binarization. Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem due to the high inter/intra-variation between the text stroke and the document background across different document images.

**Author for correspondence*

The main goal of binarization is to convert a Gray scale input image into a binary image, because many vision algorithms and operators only handle the binary image rather Gray scale image. Selection of a global threshold value is a general technique to convert Gray scale image into binary image. Such general methods binarize the entire image using a single threshold value. One simple way is to automatically select the value at the valley of intensity histogram of the image, assuming that there are two peaks in the histogram, one corresponding to the foreground, the other to the background. Poor contrast and strong noise in the input image is also a challenge for selection of proper threshold value because due to the poor contrast, many images do not have such two peaks in the histogram.

Many document image binarization methods have been proposed which are usually classified in two main categories, namely global and local. Global methods binarize the entire image using a single threshold value, where as in local thresholding algorithms compute a separate threshold for each pixel based on a neighborhood of the pixel. Reference points in binarization are considered the global thresholding method of[2] and the local adaptive methods of[14,15] which are widely incorporated in binarization methods. After that various binarization approaches was introduced in document image binarization. In most of the systems researchers generally use the[2] method for binarization.

Document binarization is an important application of vision processing. The main objective is to evaluating the short comings of algorithms for degraded image binarization. It has been found that each technique has its own benefits and limitations; no technique is best for every case. The main limitations of existing workers are found to be noisy and low intensity images. So, in this work we have proposed a new algorithm which will use more reliable methodology to enhance the work.

Historical printed documents, such as old books and old newspapers, are being digitized and made available through software interfaces such as web-based libraries. Scanned images of the original documents are usually displayed in grayscale or color for the benefit of human readers, but Optical Character Recognition (OCR) is used to enable keyword searches, document categorization, and other referencing tasks.

As illustrated in Figure 1, the handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. In addition, historical documents are often degraded by the bleed through as illustrated in Figure 1(d) and (e) where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts as illustrated in Figure 1(e). These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge to most state-of-the-art techniques.

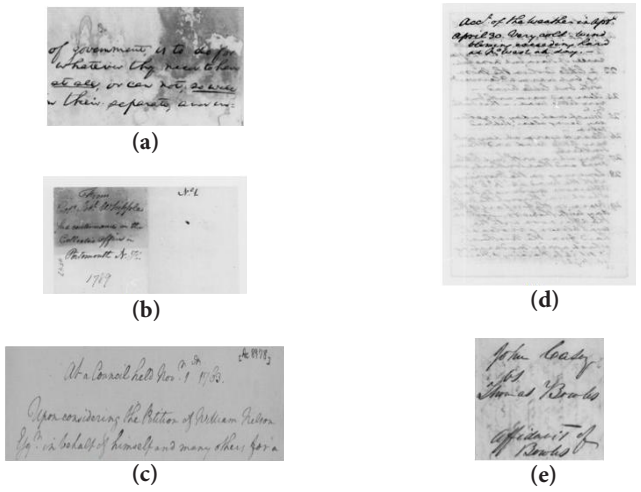

Figure 1. Example of five degraded document images taken from DIBC02009 test images handwritten database.

Five global thresholding techniques were compared which is referred in[4], with its performance. It was shown that Otsu's[2] and Kittler and Illingworth's[3] thresholding techniques generate relatively good results, but, performance of their techniques varies depending on data sets. In this paper, an efficient approach is proposed which is particularly for degraded historical handwriting document images. Proposed method based on Otsu which was already one of the best thresholding techniques for image binarization. But some time Otsu does not produce best result for heavily degraded image. In order to tolerate the degradation quality of the document images, present work proposed an efficient method for degraded document image binarization with some technical help of[2] algorithm. Proposed method can handle printed text as well.

2. Litarature Review

In[2] proposed a nonparametric and unsupervised method of automatic threshold selection for image segmentation. The proposed optimal threshold is selected by the discriminant criterion, namely, so as to maximize the separability, utilizing only zoroth and first-order cumulative moments of the gray level histogram. Several experimental results have presented to support the validity of the proposed method.

In[6] have proposed a novel document image binarization technique that concentrates on these issues by using adaptive image contrast. An adaptive contrast map is first assembling for an input degraded document image. The contrast map is then Binarized and collective with Canny's edge map to identify the text stroke edge pixels. The document text is additional segmented by a local threshold that is approximate based on the intensities of detected text stroke edge pixels within a local window. The Beckley diary dataset that consists of numerous sticky bad quality document images also show the superior performance of proposed method, compared with other techniques.

In[10] have proposed a new method for image binarization. This is a modified and improved version of the iterative partition based algorithm. This method has been compared with other five representative binarization methods. The USC-SIPI image database has been used for experimental verification purposes. The results of implementation of the algorithms unearth the superiority of the proposed method compared to the other five methods in terms of two quantitative measures, namely, misclassification error and the relative foreground area error.

In[5] have proposed that document image binarization is a significant pre-processing technique for document image analysis that segments the text from the document image backgrounds. They have propose a learning framework that makes use of the Markov Random Field to advance the performance of the existing document image binarization methods for those degraded document images. Extensive experiments on the recent Document Image Binarization Contest datasets express that significant enhancement of the existing binarization methods when applying framework.

In[1] have proposed Document Image Binarization is a method to segment text out from the background region of a document image, which is a challenging task due to high intensity variations of the document foreground and background. They have proposed a self-learning classification framework that combines binary outputs of different binarization methods. The proposed framework makes used of the sparse representation to re-classify the document pixels and generate a better binary results.

3. Proposed work

3.1 Proposed Binarization Method

At first proposed work collect color and gray scale handwriting document from the H-DIBCO2012 and DIBCO2009 data sets. Then it

checks the image quality to measures degradation level. If the image is heavily degraded then the intensity of maximum occurrence of pixel's is closer to black, the intensity of minimum occurrence of pixel's is also closer to black and intensity difference between maximum and minimum occurrence pixels are small. In this case proposed method measures the maximum and minimum threshold values of the image.

To measure the degradation quality, first calculate the mean of entire image. Then divide the image into n parts and calculate the mean of local n images. Now find the mean difference between the mean of entire image and mean of local image and set a threshold. If all the difference is less than the threshold then it is called that the input document is non-degraded in nature means uniform background picture, otherwise it is non-uniform image.

If image is uniform then OTSU method is applied directly to produce binary image. But for some images, OTSU does not give better result. So the proposed method is better approach to give better binary image where text's edge is better than OTSU. First calculate the histogram of an image, and find the pixel value which occurs more, and find the mean value of the image. Now it finds the difference between mean value and pixel value which occurs maximum time. If the difference value is less than 20, then threshold is chosen by subtracting 30 from then minimum value between mean and pixel value.

Now the problem arises when image is non-uniform. In this case, proposed method first uses OTSU method for finding threshold. Depending on this threshold value document image is converted to binary image. But using OTSU we do not get the true binary image. So the intensity value of input gray image is taken where we get black pixel of binary image. Again OTSU is applied. Now binarization method is called using threshold and gives the binary image. This final binary image is more than better than previous binary image. Yet this binarization technique does not give best output. By this some text is omitted due to intra variation of image quality and noise between two lines came. So if we use line segmentation and apply the proposed method, then noise can be removed. And also it gives better result than final binary image that is taken from binarization method without line segmentation.

To use this method, first normal OTSU method is applied, then depending on the binary image, lines are segmented[12,13] from the original image. After that Otsu thresholding technique is applied to each segmented line individually to create the segmented binary lines. After getting all segmented binary lines, lines are concatenated in similar order as original image to construct original non-degraded binary image.

3.2 Algorithm

3.2.1 Algorithm 1

```
Algorithm START_binarization()
{
  Take an image I
  GR= gray image (I)
    Check uniform or non-uniform background image
  If uniform image
    BW=MAX_MEAN_method(I)
  else
    Binary_OTSU=Binarization (I)
    Segment each same background line image from original image
    IMG
    for each same background line LN do
      if segmented line is non-uniform image then
        Binary_Line= degradedBinarization (LN)
        BW=concatLine(BW, Binary_Line)
      else
      Binary_Line=MAX_MEAN_binarization(LN)
```

```
            BW = concatLine(Bw, Binary_Line)
        end if
    end for
  end if
}
```

3.2.2 Algorithm 2

```
Algorithm MAX_MEAN_Binarization (I)
{
        [r c]=size(gr)
        mn=mean(gr)
        h=histogram(gr);
        [mx,ind]=max(h)
        if(abs(mn-ind)<20)
                m=min(mn,ind)-30
        else
                m=min(mn,ind)-10
        end
          B=Binarization(gr, m);
}
```

3.2.4 Algorithm 3

```
Algorithm Binarization(GR,th)
{
        [r c]=size(GR)
        for x=1 to r do
          for y=1 to c do
            if GR(x,y)<th
              B(x,y)=0
            else
              B(x,y)=1
}
```

3.2.5 Algorithm 4

```
Algorithm degradedBinarization (LN)
{
th_line= OTSU (LN)
AR=Binarization(LN, th_line); % send gray image +histogram
for i=1: r
  for j=1:c
    if(AR(i,j)==0)
      ITR(i,j)=A(i,j);
    else
      ITR(i,j)=A(i,j)+50;
th= OTSU (ITR)
B=Binarization(ITR,th-40)
}
```

3.2.6 Algorithm 5

Algorithm OTSU(I)

{

Compute thistogram and probabilities of each intensity level

Step through t=1 to maximum intensity level

Compute w_b, σ^2_b and w_f, σ^2_f for the two class sepetarted bt t

Compute $\sigma^2_{wcv}(t) = W_b(t) * \sigma^2_b(t) + W_f(t) * \sigma^2_f(t)$

Loop until find minimum variance $\sigma^2_{wcv}(t)$

Desired threshold= t;

}

3.2.7 Algorithm 6

```
Algorithm LineSegmeantion()
{
Take an Binary Image I
BW=convert to Binary Image
```

Repeat

R1=start_line(BW)// return row no which has maximum white pixel or minimum black pixel and next row has reverse condition

R2= end_line(BW) // return row no which has maximum black pixel or minimum no of white pixel and next row has reverse condition

LN= I(R1:R2,1:c)

Until(end the end of the last row)

}

3.3 Experimental Results

The proposed work implemented in MATLAB on H-DIBCO2012 and DIBCO2009 over 100 text images containing 3800 words.

As expected, Otsu's algorithm did not perform well in extracting characters from complicated backgrounds shown in Figure 4 and Figure 6 for the image shown in Figure 2 and Figure 3. The proposed algorithm performed better than Otsu's method for noisy images as well as the images with degraded contrast. The experiment results of proposed uniform background image binarization method over Otsu method are shown in Figure 5 and Figure 7.

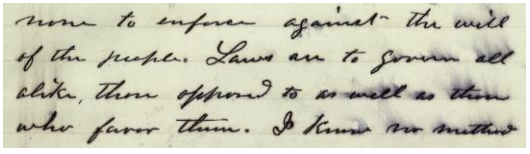

Figure 2. Example of document image from H-DIBC02012 test images handwritten database.

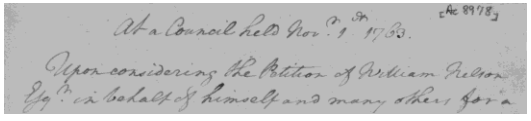

Figure 3. Example of document image from DIBC02009 test images handwritten database

Now the following experimental result shows the binarization image for badly degraded image. Figure 8 shows a image and if the algorithm degraded binarization() is applied on this image directly then some text is deleted from the image. So same background line segmentaion is applied in the image, then accoring to the image degradation level it applies max_mean_method or degradation Binarization method described in proposed work.

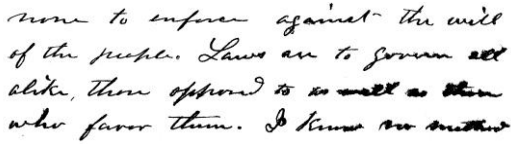

Figure 4. Binary image after appying OTSU method on the figure 2.

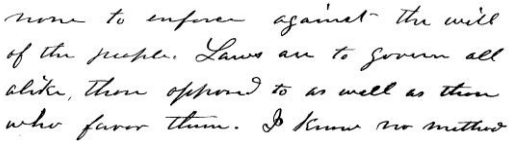

Figure 5. Binary image after appying proposed method on the Figure 2.

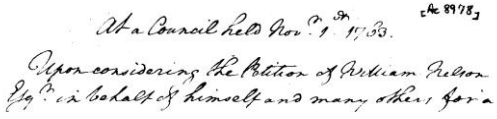

Figure 6. Binary image after appying OTSU method on the Figure 3.

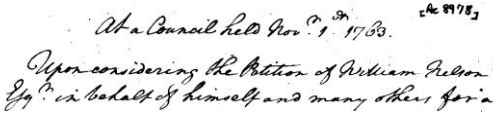

Figure 7. Binary image after appying proposed method on the figure 3.

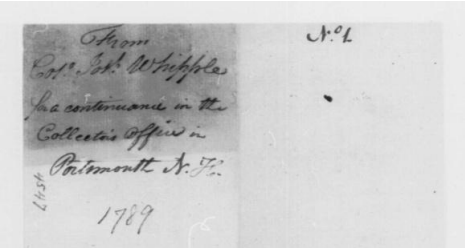

Figure 8. Example of document image from DIBC02009 test images handwritten database.

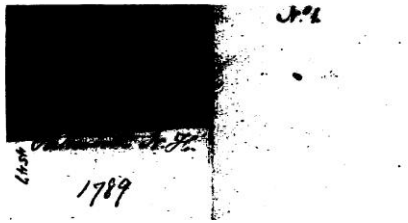

Figure 9. Binary image after appying OTSU method on the figure 8.

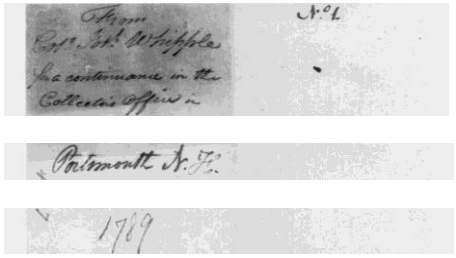

Figure 10. Same background image line segmentaion on figure 8.

From the experiement result of proposed method shown in Figure 5, Figure 7 and Figure 11, it is clear that proposed method extract more prominent charecters from the uniform backgorund document image rather that the Otsu method.

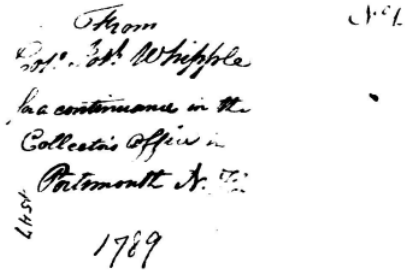

Figure 11. Binary image after appying proposed method on the figure 8.

4. Conclusion and Future Work

Document binarization is an important application of vision processing. The main objective is to evaluating the short comings of algorithms for degraded image binarization. It has been found that each technique has its own benefits and limitations; no technique is best for every case. The present work proposed a binarization technique in current scenario. The proposed method was tested on more than 550 text image of H-DIBCO2012and DIBCO2009. It has been seen that Otsu algorithm is not optimal for all kind of degraded images. In our application, Otsu's algorithm does not always segment characters well out of troublesome images. The superior performance was demonstrated in the experimental part with the proposed algorithm as compared with other algorithms. Using the proposed method most of the degraded images are binarized well with compare to Otsu algorithm which are shown in experimental part.

As a future work, we want to override the weakness of proposed approaches and obtaining a more robust system. Also want to include handwriting features with the proposed method like line segmentation, word segmentation, character segmentation, size of the letter pen pressure and some others personality.

5. Acknowledgment

The authors are very much grateful to the Department of Computer Science and engineering for giving the opportunity to work on An Efficient Methods for Skew Normalization of Handwriting Image. Both the authors sincerely express his gratitude to Dr. Arup Kumar Bhaumik, Principal of RCC Institute of Information Technology College for giving constant encouragement in doing research in the field of image processing.

6. References

1. Bolan S, Lu S, Tan CL. Robust document image binarization technique for degraded document images. IEEE Transactions on Image Processing. 2013; 22(4):1408–17.
2. Otsu N. A threshold selection method from gray-scale histogram. IEEE Trans Systems, Man, and Cybernetics. 1978; 8:62–6.
3. Kittler J, Illingworth J. On threshold selection using clustering criteria. IEEE Trans Systems, Man, and Cybernetics. 1985; 15:652–5.
4. Lee SU, Chung SY. A comparative performance study of several global thresholding techniques for segmentation. Computer Vision, Graphics, and Image Processing. 1990; 52:171–90.
5. Bolan S, Lu S, Tan CL. Combination of document image binarization techniques. 2011 IEEE International Conference on Document Analysis and Recognition (ICDAR); Beijing. 2011 Sep 18-21. p. 22–6
6. Lu S, Su B, Tan C. Document image binarization using background estimation and stroke edges. International Journal on Document Analysis and Recognition. 2010 Dec; 13:303–14.
7. DIBCO 2009 (Document Image Binarization Contest) image dataset.
8. H-DIBCO 2012 (Handwritten Document Image Binarization Contest) image dataset.
9. Gill TK. Document image binarization techniques- a review. International Journal of Computer Applications. 2014 Jul; 98(12).
10. Shaikh SH, Maiti A, Chaki N. Image binarization using iterative partitioning: A global thresholding approach. International Conference on IEEE Recent Trends in Information Systems (ReTIS); Kolkata. 2011 Dec 21-23. p. 281–6.
11. Gupta MR, Jacobson NP, Garcia EK. OCR binarization and image pre-processing for searching historical documents. The Journal of the Pattern Recognition Society. 2007; 40(2):389–97.
12. Bal A, Saha R. An efficient method for skew normalization of handwriting image. 6th IEEE International Conference on Communication Systems and Network Technologies; Chandigarh. 2016. p. 222–8. ISBN: 978-1-4673-9950-0.
13. Bal A, Saha R. An improved method for text segmentation and skew normalization of handwriting image. 4th Springer International Conference on Advanced Computing, Networking, and Informatics (ICACNI-2016); India: National Institute of Technology Rourkela. 2016 Sep 22-24. ISSN: 1876-1100.
14. Niblack W. An introduction to digital image processing. Englewood Cliffs: Prentice Hall; 1986.
15. Sauvola J, Seppanen T, Haapakoski S, Pietikainen M. Adaptive document binarization. 4th Int Conf on Document Analysis and Recognition; Ulm, Germany. 1997. p. 147–52.