Document Image Binarization Techniques- A Review

Tarnjot Kaur Gill
Department of CSE
Sai Institute of Engineering and Technology
Amritsar. Punjab, India

ABSTRACT
Image binarization is the procedure of parting of pixel values into dual collections, black as foreground and white as background. Thresholding has found to be a well-known technique used for binarization of document images. Thresholding is further divide into the global and local thresholding technique. In document with uniform contrast delivery of background and foreground, global thresholding is has found to be best technique. In degraded documents, where extensive background noise or difference in contrast and brightness exists i.e. there exists many pixels that cannot be effortlessly categorized as foreground or background. In such cases, local thresholding has significant over available techniques. The main objective of this paper is to evaluate the different image binarization techniques to find the gaps in existing techniques.

General Terms
Documents, Binarization, thresholding, Binary image

Keywords
Document Image binarization, Thresholding

1. INTRODUCTION
Document binarization is typically execute in the pre-processing phase of several document image processing associated fields such as optical character recognition(OCR) and document vision retrieval. Image binarization exchanges a picture up to 256 gray levels to a black and white picture. The simplest approach to apply image binarization is to prefer a threshold value and organize all pixels with values above this threshold as white and all other pixels as black.

Adaptive image binarization is required where an optimal thresholding is selected for each picture area. Thresholding is the simplest technique of image segmentation, from gray scale image. Thresholding can be used to generate binary images. Document images frequently experience from dissimilar types of degradation that renders the document binarization a challenging task.

Figure 1 Image (a)Before (b) After

1.1 Importance
Binarization is a significant phase in all systems of images processing and analysis. It has objective to reduce the amount of information present in the image and to lookout only applicable data which allows us to apply straightforward analysis technique performance of document analysis depends on binarization algorithm. It should on the one hand: conserve the most of information and details present in the entered picture and then it should remove the noise superposed to the original picture.

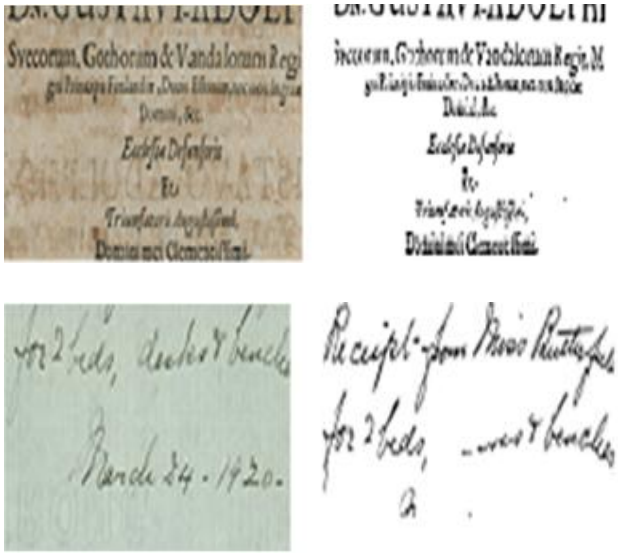

Figure 2 .(a) Original image (b) Binarized machine printed image (c) Original handwritten image (d) Binarized handwritten image

1.2 Techniques
Binarization method gray level images can be confidential in two categories: Global thresholding and Local thresholding. Global thresholding is that where one threshold is used in the entire picture to split it in two classes and local thresholding is that where the threshold values are resolute nearby pixel by pixel or region by region. Binarization methods were confidential according to the information that utilize in six categories methods based on histogram shape, entropy, clustering, and object attributes, spatial methods and local method.

1.2.1 Otsu Method:
Otsu method is solitary of the well-known global methods. This technique discovered the threshold T which split the gray level histogram into two segments. The computation of inter-classes or intra classes variances is based on the normalized histogram of the image H=[h0.....h255]where {hi=1}.Otsu method is apply to routinely execute clustering-based image thresholding. In this we thoroughly look for the threshold that minimizes the intra-class variance distinct as a weighted sum of variances of the two classes:

$$\sigma^2 \mathrm{prb}(t) = \mathrm{prb}_1(t)\sigma_1^2 + \mathrm{prb}2(t)\sigma^2 2(t)$$

Here prb are the probabilities of the two classes divided by threshold and variances of the classes. The class probability and class means can be computed iteratively.

1.2.2 ISODATA Method

Thresholding use ISODATA consists to discover a threshold by sorting out iteratively the gray-level histogram into two classes, with the appropriate information of the values connections to each class. This process starts by isolating the period of non-null values into two central parts.

1.2.3 Brensen Method

It is an adaptive local technique of which the threshold is designed for each pixel of the image. For each pixel of coordinates(x, y),the threshold is given by:

$$T(x,y) = \frac{Z_{low} + \mathrm{Z_{high}}}{2}$$

Zlow and Zhigh are the minimum and the maximum gray level in a squared window r*r centered more than the pixel (x, y).If the distinction quantity is lesser than a threshold 1, then the neighborhood consists of a single class: background or text.

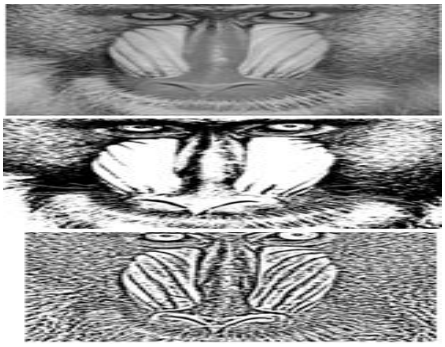

Figure 3.Image a) Original b) Reference c) Applying Brensen

1.2.4 Niblack method

Niblack algorithm analyze a confined threshold for every pixel by descending a rectangular window above entire picture .the calculation of the threshold is based on confined mean m and the standard deviations of all pixels in the window and is given by:

$$Tniblack = m + k * s$$

$$Tniblack = \mathrm{m} + k\sqrt{\frac{1}{NP}\sum}(\mathrm{pi} - \mathrm{m})^2$$

$$m + k\sqrt{\sum \frac{pi^2}{NP} - m} = \mathrm{m} + \mathrm{k}\sqrt{\mathrm{B}}$$

The threshold T is deliberate by using mean m and standard deviation σ of all pixels in the window. Thus the threshold T is given by: T= m+ k * σ. Such as k is a parameter used for find out the number of edge pixels measured as object pixels and takes a negative values. Advantage of niblack is that it always recognize the text regions properly as foreground but it tend to generate a huge quantity of binarization noise in non – text region.

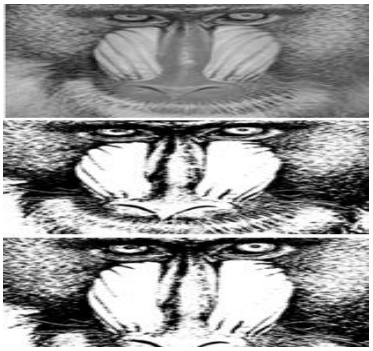

Figure 4.Image a) Original b) Reference c) Applying Niblack

1.2.5 Sauvola method

The Sauvola algorithm is alteration of niblack algorithm. It asserts to advance niblack's technique by calculating the threshold using the forceful variety of picture gray value standard deviation R:

$$\mathrm{T_{sauvola}} = \mathrm{m} * \left(1 - \mathrm{k} * \left(1 - \frac{\mathrm{S}}{\mathrm{R}}\right)\right)$$

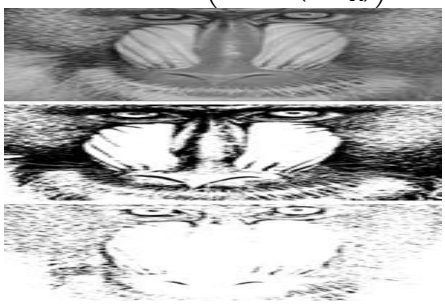

Figure 5 Image a) Original b) Reference c) Applying Sauvola

2. LITERATURE SURVEY

Bolan Su et al. (2013) [1] 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.

Bolan Su et al. (2013) [2] 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.

Wagdy et al. (2013) [3] have proposed that alteration from gray scale or color document image into binary image is the main and significant step in most of optical character recognition (OCR) systems and document analysis. Most of the preceding binarization methods that depend on local thresholding consume more time. They present fast and

efficient document image clean up and binarization method based on retinex theory and global threshold. The proposed method is fast and generates high quality results compared to the previous works.

Rabeux et al. (2013) [4] have proposed an approach to expect the result of binarization algorithms on a given document image according to its state of degradation. Historical documents experience from different types of degradation which result in binarization errors. They have proposed to characterize the degradation of a document image by using different features based on the intensity, quantity and location of the degradation. These features allow us to build prediction models of binarization algorithms that are very accurate according to R2 values and p-values. The forecast models are used to pick the best binarization algorithm for a given document image. This image-by-image strategy improves the binarization of the whole dataset.

GACEB et al. (2013) [5] have proposed the automatic reading systems of business documents requires fast and precise reading of interest zones using the OCR technology. The quality of each pixel is predictable using a hierarchical local thresholding in order to classify it as foreground, background or ambiguous pixel. The global quality of the image is thus predictable from the density of theses degraded pixels. If it is considered as degraded, we apply a second separation on the ambiguous pixels to separate them into background or foreground. This second process uses our improved relaxation method that we have accelerate for the first time to integrate it into a system of automatic reading document. Compared to existing binarization approaches (local or global), offers a better reading of characters by the OCR.

Seki et al. (2013) [6] have proposed a novel method using "color drop-out" for document images with "color shift" is proposed. Color shift phenomena occasionally arise in document images captured by a camera device or stand type scanner. It unfavourably affects the binarization and character recognition processes, because it generates pseudo color pixels on scanned image, which do not exist on the original document. To solve the "pseudo color problem," a binarization method based on the following three calculation steps is proposed. First, line and character areas are estimated coarsely by using form structure analysis and subtracting background from images, second, the color shift is detached by using morphological processing, third, each pixel of the background subtracted images is discriminated into character strings and lines accurately by dynamic color classification.

Smith et al. (2012) [7] have discussed that Image binarization has a great outcome on the rest of the document image analysis processes in character recognition. The preference of binarization ground truth affects the binarization algorithm design, either directly if design is by automated algorithm trying to contest the provided ground truth, or indirectly if human designers adjust their designs to execute better on the provided data. Three variations in pixel accurate ground truth were used to train a binarization classifier. The performance can differ considerably depending on choice of ground truth, which can manipulate binarization design choices.

Papavassiliou et al. (2012) [8] have proposed that Document image binarization is an original though critical stage towards the recognition of the text components of a document. The method is based on mathematical morphology for extracting text regions from degraded handwritten document images. The basic stages of our approach are: (a) top-hat-by-reconstruction to produce a filtered image with reasonable even background, (b) region growing starting from a set of seed points and attaching to each seed similar intensity neighbouring pixels and (c) conditional extension of the initially detected text regions based on the values of the second derivative of the filtered image.

Pratikakis et al. (2012) [9] have proposed that H-DIBCO 2012 is the International Document Image Binarization Competition which is devoted to handwritten document images structured in conjunction with ICFHR 2012 conference. The objective of the contest is to recognize existing advances in handwritten document image binarization using significant estimation performance measures.

Bolan Su et al. (2012) [10] 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 our proposed framework.

Le, T.H.N et al. (2011) [11] have proposed an enormous number of historical and badly degraded document images can be found in libraries, public, and national archives. A novel adaptive binarization algorithm using ternary entropy-based approach has been proposed. The pixels in the second region are relabelled by the local mean and the standard deviation. This method classifies noise into two categories which are processed by binary morphological operators, shrink and swell filters, and graph searching strategy. The estimation is based upon nine distinct measures. The proposed algorithm outperforms other state-of-the-art methods.

Bolan Su et al. (2011) [12] have proposed Document image binarization has been studied for decades, and several convenient binarization techniques have been proposed for different kinds of document images. They have proposed a categorization framework to merge different thresholding methods and generate improved performance for document image binarization. Given the binarization results of some reported methods, the proposed framework separates the document image pixels into three sets, namely, foreground pixels, background pixels and uncertain pixels. A classifier is then applied to iteratively classify those doubtful pixels into foreground and background, based on the pre-selected foreground and background sets. The proposed framework out performs most state-of-the-art methods significantly.

Shaikh, S.H. et al. (2011) [13] 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.

3. GAPS IN LITERATURE

Many techniques have been proposed so far for document binarization as shown in literature survey. 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. Following are the main limitations of this research work:-

1. Many researchers have used image filters to reduce the noise from the image but the use of the Decision based switching filter (best edge preserving filter) is not found. It may increase the accuracy of the available binarization methods
2. In the most of techniques the contrast enhancement is either done by tradition methods or not done. So adaptive contrast enhancement is required.
3. Most of the methods have neglected the use of edge map which has the ability to map the exact character in efficient manner.

4. CONCLUSION

This paper has focused on the degraded document binarization technique. Document binarization is an important application of vision processing. The main objective of this paper 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. In near future we will propose a new algorithm which will use more reliable methodology to enhance the work. We will propose a new algorithm which will use nonlinear enhancement as a pre-processing technique to improve the results further.

5. REFRENCES

[1] Su, Bolan, Shijian Lu, and Chew Lim Tan. "Robust document image binarization technique for degraded document images." *Image Processing, IEEE Transactions on* 22.4 (2013): 1408-1417.

[2] Su, Bolan, et al. "Self Learning Classification for Degraded Document Images by Sparse Representation." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.

[3] Wagdy, M., Ibrahima Faye, and DayangRohaya. "Fast and efficient document image clean up and binarization based on retinex theory." Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on. IEEE, 2013

[4] Rabeux, Vincent, et al. "Quality evaluation of ancient digitized documents for binarization prediction." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.

[5] Gaceb, Djamel, Frank Lebourgeois, and Jean Duong. "Adaptative Smart-Binarization Method: For Images of Business Documents." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.

[6] Seki, Minenobu, et al. "Color Drop-Out Binarization Method for Document Images with Color Shift." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013

[7] Smith, Elisa H. Barney, and Chang An. "Effect of" ground truth" on image binarization." Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

[8] Papavassiliou, Vassilis, et al. "A Morphology Based Approach for Binarization of Handwritten Documents." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[9] Pratikakis, Ioannis, Basilis Gatos, and KonstantinosNtirogiannis. "ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012)."Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[10] Su, Bolan, Shijian Lu, and Chew Lim Tan. "A learning framework for degraded document image binarization using Markov random field." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[11] Le, T. Hoang Ngan, Tien D. Bui, and Ching Y. Suen. "Ternary entropy-based binarization of degraded document images using morphological operators."Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

[12] Su, Bolan, Shijian Lu, and Chew Lim Tan. "Combination of document image binarization techniques." Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

[13] Shaikh, SoharabHossain, AsisMaiti, and NabenduChaki. "Image binarization using iterative partitioning: A global thresholding approach." Recent Trends in Information Systems (ReTIS), 2011 International Conference on. IEEE