BINARIZATION TECHNIQUE USED FOR RECOVERING DEGRADED DOCUMENT IMAGES

Miss. Nikita Mote  
SCSMCOE, Ahmednagar, India

Miss. Shital Avhad  
SCSMCOE, Ahmednagar, India

Miss. Sonali Jangale  
SCSMCOE, Ahmednagar, India

Abstract: Document Image binarization converts an acquired gray-scale document image to binary format, the objective of binarization is to automatically choose a threshold that separates the foreground and background information. Document image binarization is a process that is usually carried out in the pre-processing stage of document image processing. Primary aim of this document image binarization is to extract the foreground text from the document background. In the case of degraded document images this text extraction or segmentation is a difficult task. In this paper, we propose a simple and efficient document image binarization technique it makes use of the adaptive image contrast and some of the noise reduction methods. In the proposed technique, first input degraded document image is normalized to improve the quality of output binarized document image. Second, an adaptive image contrast map is constructed for the normalized image. Third, adaptive image contrast map is binarized and combined with Canny’s edge map to identify the text stroke edge pixels. Then the document text is segmented by a local threshold that is estimated based on the intensities of the detected text stroke edge pixels. Finally, the output document image is filtered to reduce noise. The proposed method requires only minimum number of parameters. This method shows superior performance over various datasets in terms of various performance measures.

Keywords: Degradation, Equations, Histograms, Image edge detection, Image segmentation.

1. Introduction

In recent years, the field of document image processing has increasingly widespread applicability and powerful growth. Document Image Binarization is usually performed in the preprocessing stage of Document image processing. Frequently, binarization is used as a pre-processor before Optical Character Recognition (OCR). Image binarization converts an image of up to 256 gray levels to a black and white image. Document Image Binarization converts a gray-scale document image into binary document image. The main of this document image binarization technique is to segment or extract foreground text from the document background. In the case of degraded document images this foreground text extraction is a challenging task due to variations in the document image properties. By degradations we mean every sort of less-than ideal properties of a real document image, example coarsening of document image, ink or toner drop-outs, smear, thinning and thickening, geometric deformations etc.
Handwritten text within document images also shows some variations in stroke width, stoke connection etc. In addition historical document images are often degraded by bleed-through. Many document image binarization techniques have been reported for binarization of degraded document images. In the year 1998, a recursive thresholding binarization technique for image segmentation has been proposed. This approach is only applicable to gray-scale images specifically for real-life bank checks. Performance analysis indicates that this method is more efficient to segment darkest object in a given image. An Iterative multimodel subimage binarization technique has been proposed for handwritten document images in the year 2004. This approach can be used for different types of handwritten document images where we do not do not have prior knowledge about noisiness of document image. In the year 2005, an image binarization technique has been proposed for degraded historical document images. This approach is mainly based on a decompose algorithm. Main drawback of this approach is that the algorithm does not work well on document images with big pattern or pictures. In order to give best results on heavily degraded document images a document image binarization technique using Markov field model has been proposed. This method is more effective to detect text than other local thresholding methods. An improved document image binarization technique has been proposed in the year 2008. This method is mainly based on the combination of different document image binarization technique and efficient edge information about gray scale images. A document image binarization using background estimation and stroke edges has been proposed in the year 2010. The proposed document binarization method still has several limitations. One of the drawbacks is that the proposed technique is worked for the binarization of scanned document images that have no or weak slanting. Another approach for document image binarization is using local maximum and minimum filter. We attempt to create vigorous and productive report picture binarization methods which have the capacity to handle great effects for severely debased archive pictures. Generally, they can be classified into three major types: global binarization, local binarization and hybrid binarization methods.

1.1 Global Binarization

The global thresholding technique computes an optimal threshold for the entire image; these techniques need few computations and can work well in simple cases. But fails in complex backgrounds, such as non-uniform color and poor illuminated backgrounds. These methods are usually not suitable for degraded document images, separates foreground text and background because they do not have a clear pattern.

1.2 Local Binarization

The local binarization techniques set different thresholds for different target pixels depending on their neighbourhood/local information. Generally, these techniques are sensitive to background noises due to large variance in case of a poor illuminated document or bleed-through degradation.

1.3 Hybrid Binarization

Hybrid binarization approach combines global and local thresholding. A first step consists in carrying out a global thresholding to classify a part of the background of the document image and keep only the part containing the foreground (graphics or text in our case). A second step aims to refine the image obtained by the previous step in order to obtain a sharper result by applying an adaptive thresholding technique.

1.4 Dynamic Threshold Binarization

Dynamic Threshold Binarization such as iteration method defines the threshold of a pixel with the grey-level values of its own and neighbouring pixels and the coordinate of the pixel. This binarization method is commonly used for the bad quality images, especially the images with single peak histogram. However, owing to the dynamic threshold calculation, the method has high computation complexity and slow speed.

2. Related Work
Many thresholding techniques have been reported for document image binarization. As many degraded documents do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding, which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images. For example, the early window-based adaptive thresholding techniques, estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window. The main drawback of these window-based thresholding techniques is that the thresholding performance depends heavily on the window size and hence the character stroke width. Other approaches have also been reported, including background subtraction, texture analysis, recursive method, decomposition method, contour completion, Markov Random Field, matched wavelet, cross section sequence graph analysis, self-learning, Laplacian energy user assistance, and combination of binarization techniques. These methods combine different types of image information and domain knowledge and are often complex. The local image contrast and the local image gradient are very useful features for segmenting the text from the document background because the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques. In previous paper, the local contrast is defined as follows: $C(x,y) = \frac{I_{\text{max}}(x,y) - I_{\text{min}}(x,y)}{I_{\text{max}}(x,y) + I_{\text{min}}(x,y)}$ where $C(x,y)$ denotes the contrast of an image pixel $(x,y), I_{\text{max}}(x,y)$ and $I_{\text{min}}(x,y)$ denote the maximum and minimum intensities within a local neighborhood windows of $(x,y)$ respectively. If the local contrast $C(x,y)$ is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of $I_{\text{max}}(x,y)$ and $I_{\text{min}}(x,y)$. Bernsen’s method is simple, but cannot work properly on degraded document images with a complex document background. We have earlier proposed a novel document image binarization method by using the local image contrast.

3. Literature Review

Abdenour Sehad et al. (2013) has present a capable scheme for binarization of ancient and degraded document images, grounded on texture qualities. The suggested technique is an adaptive threshold-based. It has been calculated by using a descriptor centred on a co-occurrence matrix and the scheme is verified objectively, on DIBCO dataset degraded documents furthermore subjectively, utilizing a set of ancient degraded documents offered by a national library. The outcomes are acceptable and assuring, present an improvement to classical approaches. Konstantinos Ntirogiannis et al. (2013), has analysed that document image binarization is of incredible value in the document image examination and recognition pipeline as it disturbs further phases of the recognition procedure. The assessment of a binarization technique helps in examining its algorithmic conduct, and also confirming its adequacy, by giving qualitative and quantitative sign of its execution. A pixel-based binarization assessment approach for recorded hand written/machine-printed document image has been proposed. In the proposed assessment procedure, the review and accuracy assessment measures are fittingly adjusted utilizing a weighting plan that decreases any potential assessment unfairness. Extra execution measurements of the proposed assessment plan comprise of the rate rates of broken and missed content, false alerts, foundation commotion, character amplification, and combining. Djamel GACEB et al. (2013), has studied a smart-binarization technique of the images. In this technique, considered different degradations document images. The nature of every pixel is approximate using a hierarchical local thresholding in order to classify it as foreground, background or ambiguous pixel. The ambiguous pixels that represent the corrupted zones cannot be binarized with the same local thresholding. The global quality of the image is estimated from the density of theses degraded pixels. If image is degraded then apply a second separation on the ambiguous pixels to split them into background or foreground. Second process uses our improved relaxation method. Marian Wagdy et al. (2013), has implemented a quick and proficient document image clean up and binarization technique depend on retinae hypo
thesis and global thresholding. This technique joins of local and global thresholding with concept of retinex theory which can efficiently improve the degraded and poor quality document image. Then, quick global threshold is utilized to change over the document image into binary form. The new method conquers the limitations of the related global threshold techniques. Vassilis Papavassiliou et al. (2012) has discussed an capable technique dependent upon mathematical morphology for extracting text regions from degraded document images. The fundamental stages of methodology area) top-hat-by-reconstruction to construct a filtered image with sensible background) region growing beginning from a set of seed points and attaching to each seed similar intensity neighbour pixels and conditional extension of the first detected text regions based on the values of the second derivative of the filtered image.

4. Proposed Methodology

This section describes the proposed document image binarization techniques. Given a degraded document image, an inversion contrast map is first constructed and the text stroke edges are then detected through the grayscale conversion of contrast image. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

4.1 Construction of contrast image

Primary aim of the contrast image construction is to detect text stroke edge pixels properly. In prior to the construction of adaptive image contrast map for the input degraded document image input image is normalized to improve the quality of output binarized image. Adaptive image contrast is a combination of local image contrast and local image gradient. The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many non-stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To ex-tract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background. For the image pixels within dark regions, it will produce a small denominator and ac-cordingly result in a relatively high image contrast. How-ever, the image contrast has one typical limitation that it may not handle document images with the bright text properly.

![Block diagram of Proposed System](image)

4.2 Detection of text stroke edge pixels

We can extract the foreground text from the document background once the high contrast edge pixels are detected properly. Text Stroke edge pixels can be detected easily by using, previously constructed contrast image. Adaptive image contrast computed at the text stroke is considerably higher than that computed within document background. Contrast map is then binarized using a global thresholding method which can extract the
stroke edge pixels properly. The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image has a clear bimodal pattern, where the inversion image contrast computed at text stroke edges is obviously larger than that computed within the document background.

4.3 Estimation of Local Threshold

After high contrast text stroke edge pixels are detected properly, we can segment the foreground text from the document background by a local threshold that is estimated based on the intensities of the detected text stroke edge pixels. If we analyze different kinds of document images we can observe that the text pixels are close to the detected text stroke edge pixels and there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels.

4.4 Post-Processing

Once the initial binarization result is derived from Equation as described in previous subsections, the binarization result can be further improved by incorporating certain domain knowledge, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes. One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators.

5. Acknowledgement

We would like to take this opportunity to express our profound gratitude and deep regard to my Project Guide Prof-Deepa Agale, for his exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project. His valuable suggestions were of immense help throughout my project work. His perceptive criticism kept me working to make this project in a much better way. Working under him was an extremely knowledgeable experience for me.

6. Conclusion

This paper presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique is simple and robust, only few parameters are involved. Moreover, it works for different kinds of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets.

7. Reference

