IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Efficient Binarization of Degraded Document Image

Prof. Kiran Patil[1], Ms. Surya Pushparajan[2], Ms. Joshi Amruta[3], Ms. Suryavanshi Purnima[4], Ms. Shaikh Afrin[5]

Assistant Professor, Department of Information Technology, PVG's College of Engineering, Nashik, India[1]

UG Student, Department of Information Technology, PVG's College of Engineering, Nashik, India[2,3,4,5]

Abstract: Image binarization is the method of separation of pixel values into dual collections, black as background and white as foreground. 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.

Keywords: Binarization; local thresholding; global thresholding; binary image.

I. INTRODUCTION

After years of studies in document image binarization, the thresholding of degraded document there are many challenging task in images because of high inter variation or intra variation within the text stroke and the document background across various document images. The stroke width, stroke brightness, stroke connection, and document background vary in the handwritten text within the degraded documents. Moreover, bleed through degradation is observed in historical documents by variety of imaging outputs. For most of the existing techniques many kinds of document degradations, it is still an unsolved problem of degraded document image binarization due to the document thresholding error.

A document image binarization technique presented in this paper is an extended version of an existing local maximum minimum method. The method can handle different degraded document images with least number of parameters, making it simple & robust. It uses the adaptive image contrast which is a combination of local image contrast & local image gradient. Thus it is capable of tolerating the text & background variation induced by different types of document degradation hence we proposed a binarization technique for degared document to analyze the document, its image is binarized before processing it. It is nothing but segmenting the document background & the foreground text. For the confirmation of document image processing task an accurate document image binarization technique is a must.

In document image binarization technique that gives these issues by using image contrast. The image contrast have two type local image contrast / local image gradient. Image contrast is use to remote the text and background variations which caused by degraded documents. In our the proposed method, an contrast map is first apply on an input ie on degraded document images contrast map is then binarized the image Next step is combination of contrast map with Canny's edge map. This help to recognize the text stroke edge pixels. The degraded document text is further divided by a local thresholding. Local thresholding is rough calculation of intensities of detected text stroke pixels within a local window. The proposed technique is simple, easy , robust, and it requires minimum parameter for calculation. testing is based on three different type of public datasets that are used in the current degraded document image binarization contest DIBCO dataset 2009 and DIBCO dataset 2011 and handwritten-DIBCO or HDIBCO dataset 2010. Accuracies achieves by system is of 93.5%, 87.8%, and 92.03%, respectively, that are significantly the best-performing methods reported in the given contests. Our proposed method, show the superior performance compared with other techniques.

Figure 1.Different type of Degraded document

Different type of degraded/historical document image. As shown in Fig. 1, the hand written text within the degraded

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

documents often shows a certain amount of different type variation in document .variation in terms of the edge stroke width, edge stroke brightness, edge stroke connection, and document foreground, background. degraded documents are often degraded by the bleed through as shown in Fig. 1(a) and (c) in both fig ink of the other side seeps through other side to the front. furthermore, historical documents or degraded document are often degraded by different types of imaging artifacts as shown in Fig. 1(e). These dissimilar types of historical document degradations tend to induce the document thresholding error and make degraded document image binarization big challenge to most state-of-the-art techniques.

II. RALATED WORK

The three feature vectors described below were used to test the local regions and classify them into three types: heavy strokes, faint strokes or background. Typical examples of these three types of regions are shown in Fig.2.The background of a document does not contain any useful content information. A background area typically has lower values of edge strength and variance. A background which is totally noise-free also has a small mean-gradient value. Faint stroke areas contain faint strokes, which are very difficult to detect from the background. This kind of area typically has a medium value of edge strength and mean gradient but less variance. Heavy stroke areas have strong edge strength, more variance and larger mean gradient value. The proposed weighted gradient thresholding method is applied to the different classes of sub block.

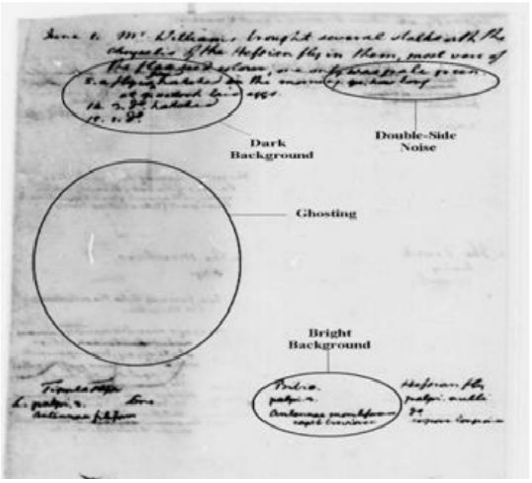

Fig.2. Example of typical historical document image

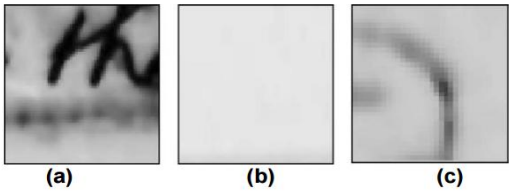

Fig.3.Examples of sub-regions containing (a) heavy strokes, (b) background and (c) faint strokes.

Enhancement of faint strokes is necessary for further processing. To avoid the enhancement of noise, a Wiener filter was first applied. The enhancement can be divided into two steps.

i. Use 3x3 windows to enhance the image by finding the maximum and minimum grey value in the window

ii. Mini =min (elements in the window)
Maxi = max (elements in the window)

Compare „pixel – mini" and „maxi – pixel", where „pixel" is the pixel-value. If the former is greater, the „pixel" is closer to the highest grey value than the lowest value in this window; hence the value of „pixel" is set to the highest grey value („pixel"=„maxi"). If the former is smaller, then the value of „pixel" is set to the lowest grey value („pixel"=„mini").A new weighted method based on mean gradient direction is proposed for thresholding [6]-[10] faint strokes. Handwritten English or Western-style scripts normally contain strokes written in several directions.

III. LITERATURE SURVEY

[1]Optical Character Recognition: Patvardhan et al. (2012) has studied that images may contain difficult background i.e. shading or a denoising. Binarization method of document images creates them suitable for OCR using discrete curvelet transform.

Curvelet transform is used for eliminate difficult image background, white Gaussian noise and gives improved binarized document image. The Curvelet transform also helps to enhanced in text shape still in the occurrence of noise. This method is capable to eliminate high frequency Gaussian noise and low frequency complex backgrounds and shows better performance.

[2] Markov Random Field model: Bolan Su et al. (2012) has studied a document image binarization structure that makes utilization of the Markov Random Field model. Structure isolates the document image pixels into three classes i.e. document background text, document foreground text, and uncertain pixels established binarization method. Uncertain pixels are belong to foreground and background categories by incorporating MRF model and boundary information.

[3] Retinex Hypothesis: Marian Wagdy et al. (2013) has implemented a quick and proficient document image clean up and binarization technique depend on retinex hypothesis 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.

[4] hierarchical local thresholding: Djamel GACEB et al. (2013) has studied a smart binarization technique of the images. In this technique, considered different

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

degradations document images. The nature of each and every pixel is approximatly using a hierarchical local thresholding method in order to classify it as foreground and background, ambiguous pixel. The ambiguous pixels that represent the corrupted pixel zones can not be binarized with the same local thresholding method. The global quality of the image is calculated from the density of theses degraded pixels image. If image is degraded then apply a second separation method on the ambiguous pixels to split them into background pixel or foreground pixel.

[5] Document shaving degradation: Vincent Rabeux et al.(2013) has an approach to expect the outcome of binarization algorithms on a known document image according to its situation of degradation. Document shaving degradation which result in binarization errors. To characterize the degradation of document by using different features. Which is based on the strength, amount and position of the degradation. These characteristics allow us to build calculation models of binarization algorithms that are very accurate according to **R2** values and p-values. The model of prediction are used to chose the excellent binarization algorithm for a given document image.

[6]Mathematical morphology for extracting text regions: Vassilis Papavassiliou (2012) has discussed an capable technique dependent up on mathematical morphology for extracting text regions from degraded document images. The fundamental stages of methodology are a) 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 Neighbor pixels and c) conditional extension of the First detected text regions based on the values of the second derivative of the filtered image.

[7]Retinex hypothesis and global thresholding: Marian Wagdy et al. (2013) has implemented a quick and proficient document image cleanup and binarization technique depend on retinex hypothesis 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.

[8]A pixel based: Konstantinos Ntirogiannis (2013) has analyzed 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 Handwritten / 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.

[9]Adaptive threshold based: Abdenour Sehad (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 centered 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.

IV. PROPOSED METHOD

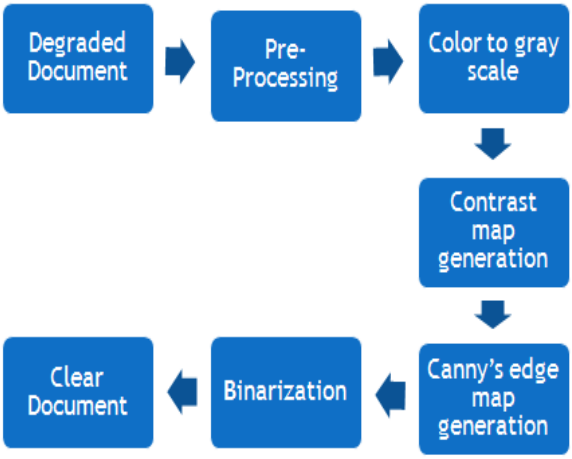

Fig 4.system architecture

A. Input or Degraded Document

Input to the system will be the number of degraded image or any historical paper that normal eye cannot read because of more degradation or could be old document and the data which is on paper or on image can't be readable due to some ink is spread on it.

B. Contrast Map Generation

The image gradient has been mostly used for text edge detection . It can be used to detect/extract the edges stroke text of the degraded document images effectively that have a constant document background.

Fig.5. Contrast Images constructed

DOI 10.17148/IJARCCE.2016.5411

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Other than that , it also detects many non- stroke edges from the backgrounds of degraded documents or historical documents that often contains some amount of image variations due to noises, un-even brightness, bleed-through, etc. To take out only the text stroke edge in images properly, the image gradients needs to be normalized to recoup the image variation within the document background image.

C. Text Stroke Edge Detection

The purpose of the contradistinction image construction is to detects only the stroke edge pixels of the image of degraded document text properly. The contrast image constructed has a clear bimodal pattern, text stroke of image contrast is obviously greater than computed within the document background. Hence we detects the text stroke edge pixel candidates by using Otsu's global thresholding technique. As the both local image contrast & gradient are assess by the difference between the maximum /minimum intensities in a local windows of pixels at both sides of the text stroke will be chose as the high contrast image pixels. The binary contrast map can be also improved through the combination with the edges stroke by Canny's edge detector, because the property Canny's edge detector has a good localization property that Cannys edge can spot the edges which is close to real edge locations in the detecting text stroke edge image. Furthermore, canny edge detector uses two adaptive thresholding ie local threshold and global threshold and is more ignores to different imaging artifacts. Artifact means shading. Canny's edge detector by itself also take out a large amount of non-stroke without tuning the parameter manually. For best result we combined it with contrast map, therefore we keep only those pixels that appears within both the high contrast image pixel map and canny's edge map.

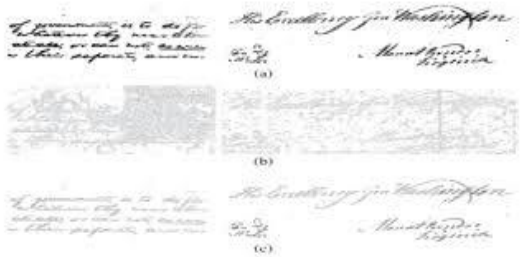

Fig.6. Text Stroke Edge Pixel Detection using canny edge maps

D. Local Threshold Estimation

The text can then be take out from the document background pixels once the high contrast text stroke edge pixels of images are detected properly. Two characteristics can be observed from different kinds of document images. First, the text pixels in the image are close to the detected text stroke edge pixels of output image. Second, there is a unique intensity variation between the high contrast stroke edge pixels and the surrounding background of image pixels.

Ensure: The rough calculation Text Stroke Edge Width WE

1) Get the widths and heights of J
2) for Each Row j = 1 to height in Edge do
3) Scan from left to right to top to bottom find edge pixels that meet the following criteria:

a) label 0 if background;

b) labeled as 1 for next pixel edge.

4) investigate the intensities in J of those pixels selected in Step 3, and accept those pixels that have a hight intensity than the following pixel next to it in the same row of J.
5) check for matching pixel with remaining adjacent pixels, and measure the distance between the two pixels.
6) end for
7) calculate distances is use for Constructing a histogram.
8) make use the most frequently occurring distance as the estimated text stroke edge widths WE.

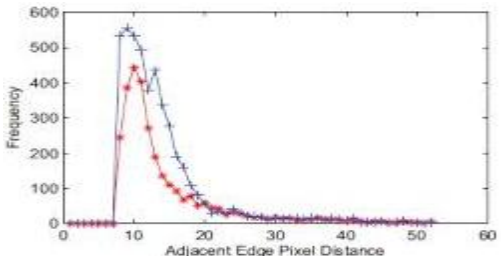

Fig 7.Local Threshold Estimation

E. Post-Processing

Once the initial binarization result is derived from equation. The binarization result can be further improved by incorporating certain domain knowledge. First, the differentiate foreground pixels that do not connect with other foreground pixels and then filtered out this unconnected pixel to make the edge of text pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text).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.

F. Output

Proposed technique combines bath method i.e local image contrast and local image gradient. This method helps to suppress the background variation and other variation to avoid the over normalization of degraded document images with less variation. Next, the combination of edge map with canny's edge helps to produce a precise stroke edge map. The proposed method makes use of the text stroke edges that help to detect the foreground text and background text from the degraded document.

IV. EXPERIMENTAL RESULTS

The evaluation measures are adapted from the DIBCO report including F-measure, peak signal-to-noise ratio (PSNR), negative rate metric (NRM), and

DOI 10.17148/IJARCCE.2016.5411

misclassification penalty metric (MPM). In particular, the F-measure is defined as follows:

FM= 2∗RC∗PRRC+PR
where RC and PR refer the binarization recall and the binarization precision, respectively. This metric measures how well an algorithm can retrieve the desire pixels. The PSNR is defined as follows:

PSNR=10log(C2MSE)
where MSE denotes the mean square error and C is a constant and can be set at 1. This metric measures how close the result image to the ground truth image. The NRM is defined as follows:
NRM= NfnNfn+Ntp+ NfpNfp+Ntn2

where Ntp, Nfp, Ntn, Nfn denote the number of true positives, false positives, true negatives, and false negatives respectively. This metric measures pixel mismatch rate between the ground truth image and result image. The MPM is defined as follows:

MPM= dfni+ dfpjNfpj=1Nfni=12D
Where di FN and d j fp denote the distance of the ith false negative and the j th false positive pixel from the contour of the ground truth segmentation. The normalization factor D is the sum over all the pixel-to-contour distances of the ground truth object. This metric measures how well the result image represents the contour of ground truth image.

V. CONCLUSIONS

We have studied in detail about our topic. We found and studied the previous techniques and papers and proposed our new technique that will overcome the drawbacks and hurdles in the previous papers. We searched all the related papers related to our topic and literature survey. We also studied algorithms that will best suite our project. And we also prepared all the UML diagrams. We have completed our planning till today.

REFERENCES

[1]. Patvardhan, C. A. K. Verma, C. Vasantha Lakshmi. "Document image denoising and binarization using Curvelet transform for OCR applications."Engineering Nirma University International Conference on. IEEE, 2012.
[2]. S u Bolan Shijian Lu Chew Lim Tan. "A learning framework for degraded document image binarization using Markov random field." Pattern Recognition 21st International Conference on. IEEE, 2012.
[3]. Wagdy M , Ibrahima Faye , Dayang Rohaya."Fast and efficient document image clean up and binarization based on retinex theory." Signal Processing and its Applications 9th International Colloquium on. IEEE, 2013.
[4]. Gaceb , Djamel , Frank Lebourgeois , and Jean Duong. "Adaptative Smart Binarization Method: For Images of Business Documents." Document Analysis and Recognition, 2013 12th International Conference on IEEE, 2013.A.B. Lewko and B. Waters, "Decentralizing Attribute-Based Encryption," Proc. Ann. Int'l Conf. Advances in Cryptology , pp. 568-588, 2011.
[5]. Rabeux, Vincent, etal."Quality evaluation of ancient digitized documents for binarization prediction." Document Analysis and Recognition, 2013 12th International Conference on. IEEE.
[6]. Papavassiliou, Vassilis, etal."A Morphology Based Approach for Binarization of Handwritten Documents." Frontiers in Handwriting Recognition, 2012
[7]. Wagdy, M., Ibrahima Faye, and Dayang Rohaya. "Fast and efficient document image clean up and binarization based on retinex theory." Signal Processing and its Applications 9th International Colloquium on .IEEE, 2013.
[8]. Ntirogiannis , Konstantinos , Basilios Gatos, and Ioannis Pratikakis. "Performance Evaluation Methodology for Historical Document Image Binarization." Image Processing, IEEE Transactions on Efficient Binarization of Degraded Document Image.
[9]. Sehad , Abdenour , et al. "Ancient degraded document image binarization based on texture features." Image and Signal Processing and Analysis International Symposium on.IEEE, 2013.