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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

Sub Pixel Mapping in Degraded Document for Text Retrieval

A. Anandhi [1], J. Renuka jothy[2], A. Vijayalakshmi[3], D. Sathiyavani[4]

Assistant Professor, Department of CSE, Christ college of Engineering and Technology, Puducherry, India [1]

Student, Department of CSE, Christ college of Engineering and Technology, Puducherry, India [2,3,4]

Abstract: Segmentation of text from poorly despoiled document image is a very demanding task due to the high inters/intra deviation between the document background and the foreground text of different document scanned images. Text withdrawal from normal picture images is an emerging field in computer graphics. Extracted text contains essential information that can be used for a variety of purposes like vehicle number plate to identify the vehicle, to provide information of neighbouring to visually impair persons, preservation of information for chronological documents etc. In our project we propose new method called sub pixel mapping that addresses the issues by using adaptive image contrast and Binarization Technique. The adaptive image contrast is a mixture of the local image contrast and the local image gradient that is forbearing to text and background variation caused by dissimilar types of document degradations, binarization algorithm for each case proved to be a very difficult procedure itself. We use the case of degraded historical documents, then we apply the proposed technique called sub pixel mapping to convert the document images to the text format.

Keywords: Adaptive image contrast, document analysis, degraded document image binarization, pixel classification; sub pixel mapping**.**

I. INTRODUCTION

In imaging science, image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or a set of character or parameters related to the image. Document Image Binarization is performed in the pre-processing stage for document analysis and it aims to segment the foreground text from the document background. This simple procedure has been proved to be a very difficult task, especially in the case of historical documents that very specialized problems have to be dealt with, such as variation in contrast and illumination, smearing and smudging of text, seeping of ink to the other side of the page and general degradation of the paper and ink due to aging. On the other hand, such a task is necessary for the further stages of document analysis either we are interested in performing OCR, or document segmentation, or just presentation of the document after some restoration stages.

Sub pixel mapping is a pre-processing task, very useful to document analysis systems. It automatically converts the document images in a text form in such way that the foreground information is represented by black pixels and the background by white ones.

The Sub pixel mapping is easy to evaluate the results by comparing the resulted image, pixel by pixel, with the original document. A document image binarization technique is for the ensuing document image processing tasks such as optical character recognition (OCR).

Though document image binarization has been considered for many years, the thresholding of corrupted document images is still an peculiar trouble due to the dissimilarity between the text stroke and the document surroundings across different document images.

In case of historical documents where their quality in many cases obstructs the recognition, and sometimes even the word segmentation, this way of evaluation can be proved problematic. On the other hand, we need a different evaluation technique, more direct, able to evaluate just the binarization stage. The ideal way of evaluation should be able to decide, for each pixel, if it has finally succeeded the right colour (black or white) after the binarization. This is an easy task for a human observer but very difficult for a computer to perform it automatically for all the pixels of several images. 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 where the ink of the other side seeps through to the front.

II. RELATED WORK

Many thresholding method [1] [2] have been records for degraded documents of sub pixel mapping. A lot of degraded documents do not clarity of background image for bimodal pattern. Sub pixel plotting [3] is a pre-processing task, very useful to analyze the degraded document.

In such a way, foreground information is represented by black pixel and background image information is represented by white pixels. This is easy to evaluate the result by comparing the resulted image, pixel by pixel in

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

original document. Adaptive thresholding [4], segmenting an image by setting the all pixels are intensity values from a threshold value to a foreground value and all the remaining pixels to a background image values. Adaptive thresholding typically takes a gray scale value and colour image as input values of simplest implementation of thresholding.

The mean values is represented by T, and the maximum and minimum values is represented by ,

$$T = \frac{Max + Min}{2}$$

Sub pixel mapping [5] is display the combination of pure pixel and mixed pixels. There is presented by the information about the spatial location of the sub pixels. Every pixels is divide into a predefined lots of sub pixels, allow a more spatially detailed on sub pixel mapping of lower resolution pixels.

OCR [6], optical character recognition is a increasing important information is available in digital format for increased efficiency in data storage and retrieval of optical images. Optical reorganization is invented by a technology in early 1800. OTSU [7] is a calculating the threshold value by accepting the existence of foreground and background the maximizing the variation of threshold black and white pixels. SAUVOLA [8] its calculates the local threshold and variation of intensity values using mean values are different variation are displayed on this method.

Binarization techniques [9], is a pre-processing methods for degraded document images are scanned documents. For execution of binarization technique is only to retrieve the scanned document images. This techniques are combination of Canny edge map [10] algorithm used on the binarization techniques to compensate the variation between documents.

III. PROPOSED METHOD

Segmentation of document images leftovers a demanding vision problem. Although document images have a planned layout, capturing enough of it for segmentation can be difficult. Most current methods merge text pulling out and heuristics for extraction, but text extraction is flat to failure and measure accuracy leftovers a complicated brave. Furthermore, when accessible with important degradation technique.

This section describes the proposed document image sub pixel plotting techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the Image adjust, data histogram and Adaptive thresholding techniques. Then the sub pixels of the text are plotted then it is 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 quality.

The proposed document image sub pixel mapping techniques for a given degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the adaptive contrast map and the canny edge map.

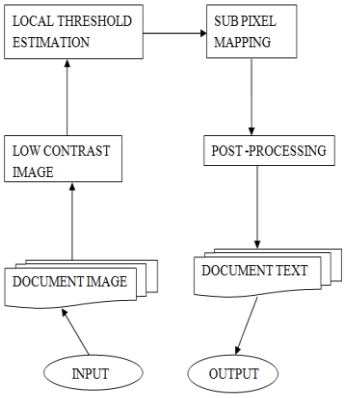

Fig. 1. Architecture diagram.

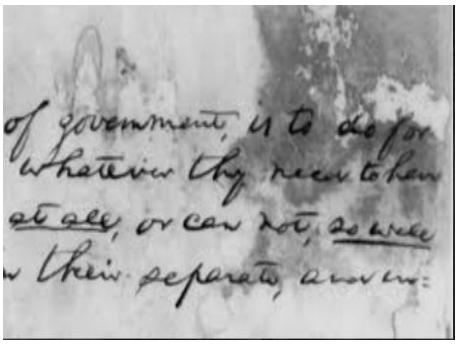

Fig. 2. Historical documents are often degraded by different types of imaging artifacts

The Fig. 1. Shows the architecture and flow of the proposed system. As illustrated in Fig. 2, 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 Fig. 2 shows historical documents are often degraded by different types of imaging artifacts. In addition, as illustrated in Fig.3. shows where the ink of the other side seeps through to the front. These are 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.

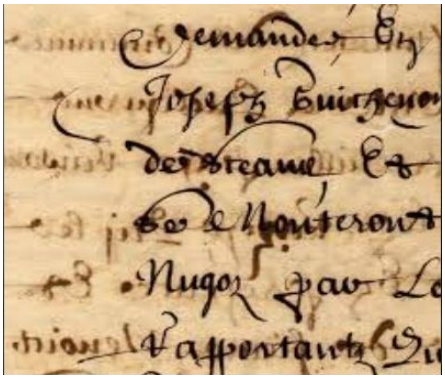

Fig. 3. The ink of the other side seeps through to the front.

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Then Sub-pixel Plotting technique is designed to obtain the spatial distribution of different classes in mixed pixels at the sub-pixel scale by transforming fraction images to classification map. However, sub-pixel mapping is an ill-posed as information in single low resolution image is not enough to obtain a high resolution. Then some of the post-processing is further applied to improve the document quality and convert to the text format.

A. *Contrast Enhancement*
By enhance the gray scale image and enhance colour images:

Imadjust - increase the contrast of the image by map the values of the input strength image to new values such that, by default, 1% of the data is soaked at low and high intensities of the input. data.Histeq - performs histogram equalization. It enhance the contrast of images by transform the values in an intensity image so that the histogram of the output image just about matches a specified histogram (uniform distribution by default).

Adapthisteq - performs contrast-limited adaptive histogram equalization. Unlike histeq, it operate on small data region (tiles) rather than the entire image. Each tile's contrast is enhanced so that the histogram of each output region roughly matches the specified histogram (uniform distribution by default). The contrast augmentation can be limited in order to avoid amplify the clatter which might be present in the image.

B. *Sub pixel plotting*
Sub-pixel plotting is a technique designed to obtain the spatial sharing of different classes in mixed pixels at the sub-pixel scale by transform division images to categorization map. However, sub-pixel plotting is an ill-posed problem as information in single low resolution image is not enough to obtain a high resolution. Exactness can be enhanced by incorporating auxiliary datasets to provide more information. Sub pixel plotting is a pre-processing task, very useful to document analysis systems. It automatically converts the document images in a text form in such way that the foreground information is represented by black pixels and the background by white ones. It is easy to evaluate the results by comparing the resulted image, pixel by pixel, with the original document.

Good detection – this method should mark as many actual edges in the image as probable.

Good localization – edges noticeable should be as close as probable to the edge in the actual image.

Minimal response – a given edge in the image should only be noticeable once, and where probable, image clamour should not create false edges.

C. *Local Threshold evaluation*
The text can then be extracted from the document background pixels once the contrast stroke edge pixels are detected properly. First, the text pixels are close to the detected text stroke edge pixels. Second, there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels.
The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

$$R(x, y) = \{1 \;\; I(x, y) \leq E\text{mean} + E\text{std}/2$$
$$0 \;\; \text{otherwise}\}$$

Where Emean and Estd are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighbourhood window W, respectively. First the edge image is scanned horizontally row by row and the edge pixel candidates are selected. If the edge pixels, which are labelled 0 (background) and the pixels next to them are labelled to 1 (edge) in the edge map (*Edg*), are correctly detected, they should have higher intensities than the following few pixels(should be the text stroke pixels).

So those improperly detected edge pixels are removed the remaining edge pixels in the same row, the two adjacent edge pixels are likely the two sides of a stroke, so these two adjacent edge pixels are matched to pairs and the distance between them are calculated. After that a histogram is constructed that records the frequency of the distance between two adjacent candidate pixels. The stroke edge width *EW* can then be approximately estimated by using the most frequently occurring distances of the adjacent edge pixels.

D. *Post-Processing*
Image Processing Toolbox provides reference-standards for post-processing tasks that solve frequent system problems, such as interfering noise, low dynamic range, out-of-focus optics, and the difference in colour representation between input and output. Post-Processing is a technique used in graphics that allows you to take a current input texture, and manipulate its pixels to produce a transformed image.

This can be used to apply shiny effects like volumetric lighting, or any other filter type effect you've seen in applications like Photoshop or histogram. Once the initial sub pixel mapping result is derived then it can be further improved by incorporating certain domain Knowledge. First, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely.

Second, the neighbourhood 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 labelled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artefacts along the text stroke boundaries are filtered out by using several logical operators.

IV. EXPERIMENTAL RESULTS

A small number of experiments are calculated to express the helpfulness and fitness of our proposed method. We first analyze the presentation of the proposed technique on

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 2, February 2014

some of the degraded documents collected from the previous techniques and methods.
We test the computation time of our proposed method and other up to date techniques implemented in Matlab.

The average execution time of the proposed method is minimum than the other existing methods and techniques like OTSU, BERN, NIBL, SAUV, GATO methods The proposed technique is comparable to the state-of-art adaptive document thresholding methods. The proposed technique is then tested and compared with methods and techniques of the collected documents. Due to need of view reality data in some datasets, No all of the metrics are apply on all images. compared to the collected or viewed methods and techniques the best results will produce the proposed method.

V. CONCLUSION

This project 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.

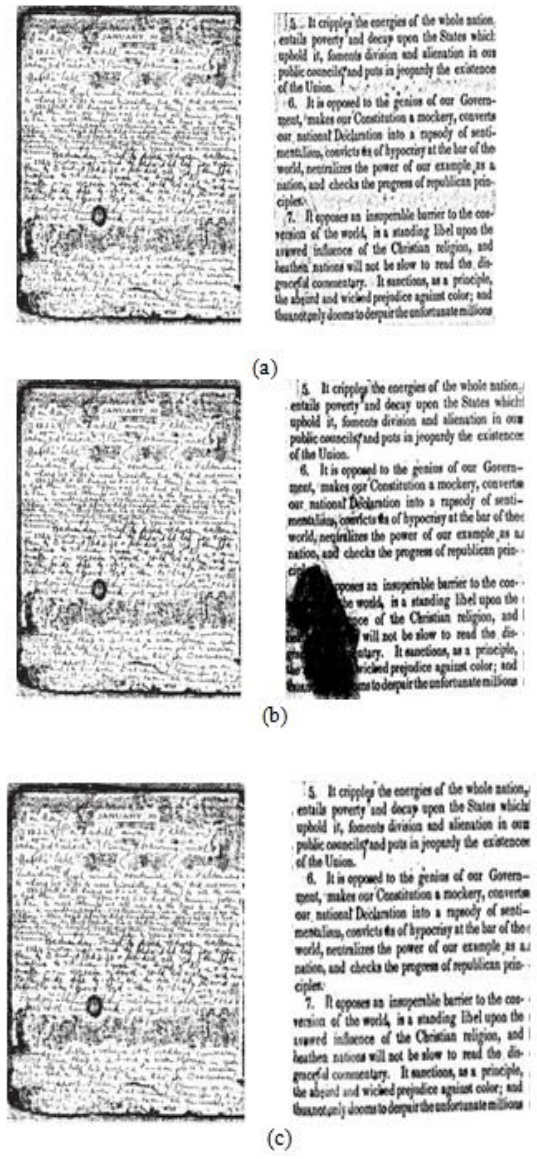

Fig. 4. (a) Wellner's Technique. (b) Adaptive Thersholding Technique. (c) Proposed Method.

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 max and min. The proposed method has been tested on the various datasets. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measures and it provides the result in the form of text retrieval.

ACKNOWLEDGMENT

The authors would like to thank Prof. Ms. A. Anandhi and her team for making the ground truth images of the Bickley diary dataset.

REFERENCES

[1] A. Brink, "Thresholding of digital images using two-dimensional entropies," Pattern Recognit., vol. 25, no. 8, pp. 803–808, 1992.
[2] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria," IEEE Trans. Syst., Man, Cybern., vol. 15, no. 5, pp. 652–655, Sep.–Oct. 1985.
[3] N. Otsu, "A threshold selection method from gray level histogram," IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 62–66, Jan. 1979.
[4] N. Papamarkos and B. Gatos, "A new approach for multithreshold selection," Comput. Vis. Graph. Image Process., vol. 56, no. 5, pp. 357–370, 1994.
[5] J. Bernsen, "Dynamic thresholding of gray-level images," in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 1251–1255.
[6] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images," in Proc. Int. Conf. Document Anal. Recognit., Sep. 1991, pp. 435–443.
[7] I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model," Pattern Recognit., vol. 35, no. 1, pp. 265–277, 2002.
[8] J. Parker, C. Jennings, and A. Salkauskas, "Thresholding using an illumination model," in Proc. Int. Conf. Doc. Anal. Recognit., Oct. 1993, pp. 270–273.
[9] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.
[10] S. Zhixin, S. Setlur, V. Govindaraju, "Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map", Proceedings of the Eighth International Conference on Document Analysis and Recognition, 2005, pp. 794-798.