International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A Review on image binarization in degraded documents images

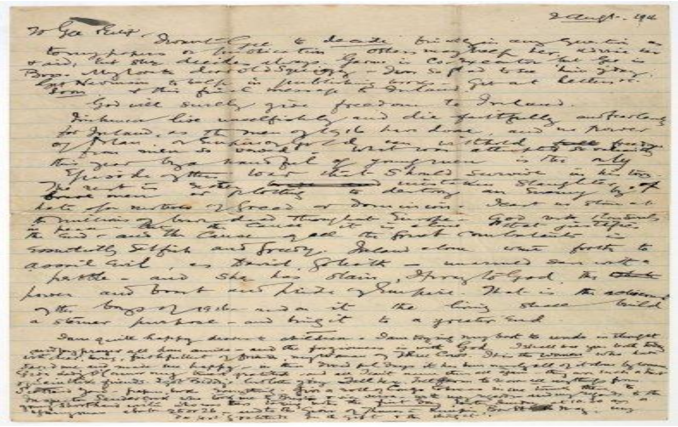
[1]DivyaJyoti, [2]Arun Sharma, [3]Bodh Raj, [4]KapilKapoor
[1,2,3,4]Department of Computer Science Engg.
Himachal Pradesh Technical University, Hamirpur

Abstract: *At present time image processing is a hot immerging research area, and in this paper we have reviewed one of the hot immerging area of image processing that is image binarization, with the help of image binarization we would able to retrieve text from degraded documents, we have also present a solution for improving the existing research in the later section of the paper.*

I. INTRODUCTION

Digital image processing is a vital field in image processing. Images are processed using algorithms. It has numerous plus points over analog image processing. In digital image processing large number of algorithms can be used with input. Problems like noise, signal distortion can be avoided. Image segmentation is a process of dividing a digital image into different segments so as to analyze it properly and represent it in a more meaningful way. The very initial steps of image segmentation are pattern recognition and image analysis. Image segmentation techniques are based upon two approaches: 1) Detection of discontinuities and 2) Detection of similarities. Detection of discontinuities includes algorithms like edge detection. Partitioning of an image depends upon abrupt changes in intensities. Detection of similarities include algorithms like region based segmentation methods and thresholding. In this approach image is partitioned according to the similarity between the regions and some predefined criterion.

Document images, as a substitute of paper documents, mainly consist of common symbols such as handwritten or machine-printed characters, symbols and graphics. In many practical applications, we only need to keep the content of the document, so it is sufficient to represent text and diagrams in binary format which will be more efficiently transmit and processed instead of the original gray-scale image. It is essential to threshold the document image reliably in order to extract useful information and make further processing such as character recognition and feature extraction, especially for those poor quality document images with shadows, non-uniform illumination, low contrast, large signal-dependent noise, smear and smudge. Therefore, thresholding a scanned gray-scale image into two levels is the must step and also a critical part in most document image analysis systems since any error in this stage will propagate to all later phases. Document Image Binarization aims to segment the foreground text from the document background and is performed in the preprocessing stage for document analysis. For the ensuing document image processing tasks such as optical character recognition (OCR), a fast and accurate document image binarization technique is essential. Though document image binarization has been developed for many years, the thresholding of degraded document images is still an unsolved problem due to the high inter/intra variation between the document background and the text stroke across different document images.



[illegible]
God will surely give freedom to Ireland.
[illegible]

Fig.1(a)

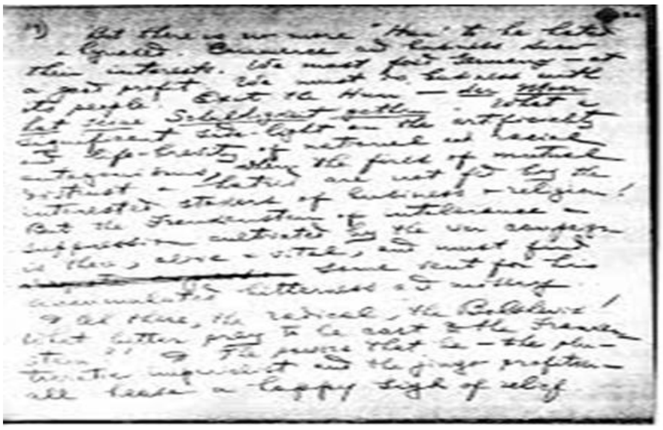


Fig.1(b)

Fig.1: (a) degraded documents having a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, (b) Degraded document image with bleed-through.

As illustrated in Fig. 1(a), 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. 1(b), where the ink of the otherside seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts. These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge.Document image binarization plays a key role in document processing since its performance affects the degree ofsuccess in subsequent character segmentation and recognition. In general, image binarization is categorized in two main classes: (i) global and (ii) local. Binarization is a preprocessing stage for document analysis and it is used to segment the foreground text from the document background. This technique ensures faster and accurate document image processing tasks. Most document analysis algorithms are built based on underlying binarized image data. The use of bilevel information decreases the computational load and helps in using simplified analysis methods compared to 256 levels of grey-scale or colour image information. Document image understanding methods require logical and semantic content preservation for thresholding. Though document image binarization has been studied for many years, the thresholding of images is still a challenging task due to the high variation between the text stroke and the document background. For an input image, some processing stages should be used before the text extraction. One of the steps includes binarization.

II. LITERATURE SURVEY

In 2009, [5] V. Papavassiliou, et al., presented two novel approaches for segmentation of handwritten documents into words and text lines. For line extraction, an image is divided into perpendicular or vertical zones, and then each zone is over-segmented into “text” and “gap” regions by using extreme points of piece-wise projection profile. For locating the optimal succession of gap and text areas in each zone the statistics of an HMM are estimated. These estimations are used in Viterbi algorithm for line segmentation. By concerning the closeness and local foreground density of adjacent zones their line separators are combined. After applying text line separator drawing technique, the connected components are assigned to text lines. A novel metric “gap” is used for measuring the separability of consecutive connected components and its underlying pdf is estimated for entire page of document. The rightmost minimum of pdf is used as a threshold and by using it, the candidate word separators are classified as “between” or “within”. At last, two benchmarking datasets of IC- DAR07 handwriting segmentation contest are used for testing algorithms and it is concluded that the proposed techniques achieved better and improved results than other techniques. Hence, these approaches are believed to be appropriate for handwritten document retrieval systems.

In 2010, [HYPERLINK \l "Min10" 1]M. Li et al., proposed a new method for the segmentation of text from the images with complex background. This new method is based upon conditional random field approach.. The proposed method fully considers the contextual information and is a supervised learning method. Contextual label information and local visual information are incorporated into a conditional random field by some components. Some of these components focus on visual image information while others focus on contextual label information. Visual image information is used for the prediction of the category within the image sites. To find out the patterns within the label field, contextual label information is used. The proposed method is compared with two methods:1) Clustering method and 2) Conventional threshold method. The experimental demonstrations clearly verify that the proposed method segments the text visibly from the complex background and resolves local ambiguities. The proposed scheme also achieves enhanced performance and outperforms Markov Random Field method and Conditional random field methods in complex background.

In 2012, [2]A. Fernandez-Caballero et al., proposed best binarization values for commercial optical character recognition system. The main objective of this work is evaluate the textual information such that the work that earlier were performed with human intervention can be now performed automatically. Optical character recognition system is described in brief. The proposed system for generic displays will recognize the text that appears on the display. It will also recognize the color of the text characters and its background. The threshold values are learnt for different display cell and then the best threshold is taken for each cell. In the end, the proposed work has been tested by using a multi display simulator and a commercial Optical Character Recognition system.

In 2011, 3]**}**H. Ben Ameur et al., proposed an algorithm that provides a finite series of segmented images. It also includes color regions that increases with a count of one with every iteration. This algorithm is called optimal adaptive segmentation algorithm. This algorithm starts extracting the significant regions from extremely beginning due to its adaptive character. Error is minimized and it results in a segmented image that contains regions equal to the number of distinct colors in original image due to its optimal character. This algorithm can manage the preferred number of regions because it works on significant criterion to end refinement process. The author also includes multiscalar algorithms. According to it, distinct segmentation is provided for each colour component. Properties like flexibility, soundness and robustness are inherited by the resulting segmentation algorithm from its optimal control origin.

In 2010, [5]P.RMarpuet al., proposed an enhanced and better concept for the quality estimation and assessment of segmentation results. The author introduced inclusive and comprehensive criteria for detecting under-segmentation and over-segmentation of all reference areas.These criteria help in understanding the segmentation results. It also provides essential support to find optimal parameter settings. The evaluation criteria are designed to handle the results of multi-level segmentation algorithms. These algorithms are commonly used in GEOBIA. GEOBIA is Geographic object-based image analysis which is a methodology for image analysis, in which images are first segmented into segments andthen analyzed based upon context, shape, texture and spectral features.

III. PROBLEM DEFINITION

In the existing system, the author have used edge detection technique for detecting the edges of the old documents manuscripts, though the technique was new and also the outputs were improved from the existing technique in literature but not that much accurate.

Motivation

Libraries usually provide accessto ancient and historical document image collections. **S**pecialized processing is common for such document images in order to remove backgroundnoise and make them more readable. These historical documents are assets for every nation and need to be protected. Specialized approaches and techniques are required for information acquisition from such document images.

Limitation of existing system

Canny edge detection has a limit to constrain only inside the edge that means a part of text remains unexplored.

Objective

The objectives of current thesis are database arrangement; to design an algorithm for region based segmentation, evaluation of the performance in terms of PSNR, F-Measure, MPM and NRM and to compare the existing system with proposed system.

IV. PROPOSED WORK

In current thesis, the proposed work is that I will try to implement existing system using morphological operators and will improve the values of parameters like PSNR, F-Measure, NRM and MPM. I will do region based segmentation instead of edge based segmentation.

Classification of Parameters

For evaluation, following parameters are used:

F-Measure:

It is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of correct results that should have been returned. The measure can be interpreted as a weighted average of the precision and recall, where an Fmeasure reaches its best value at 1 and worst score at 0.

The traditional F-measure is the harmonic mean of precision and recall:

$$\text{F Measure} = 2.\text{precision}.\text{recall} / \text{precision} + \text{recall}$$

Peak signal-to-noise ratio (PSNR):

PSNR is a term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale. PSNR is most commonly used to measure the quality of reconstruction of lossy compression codec such as for image compression. The signal in this case is the original data, and the noise is the error introduced by compression. When comparing compression codec, PSNR is an approximation to human perception of reconstruction quality. Although a higher PSNR generally indicates that the reconstruction is of higher quality, in some cases it may not. One has to be extremely careful with the range of validity of this metric. It is only valid when it is used to compare results from the same codec and same content. PSNR is most easily defined via the mean squared error (MSE).

To compute the PSNR, the block first calculates the mean-squared error using the following equation:

$$\mathrm{MSE} = \sum_{MN} \frac{[I_1(m,n) - I_2(m,n)]^2}{M * N}$$

In the previous equation, M and N are the number of rows and columns in the input images, respectively. Then the block computes the PSNR using the following equation:

$$\mathrm{PSNR} = 10 \log_{10}\left(\frac{R^2}{\mathrm{MSE}}\right)$$

R is the maximum fluctuation in the input image data type.

Negative Rate Metric (NRM):

The negative rate metric NRM is based on the pixel-wise mismatches between the GT and prediction. It combines the false negative rate $NR_{FN}$and the false positive rate NR_{FP}. It is denoted as follows:

$$\mathrm{NRM} = \frac{NR_{FN +} NR_{FP}}{2}$$

Where $NR_{FN} = \frac{NR_{FN}}{NR_{FN+NR_{TP}}}$, $NR_{FP} = \frac{NR_{FP}}{NR_{FP+NR_{TN}}}$

N_{TP}denotes the number of true positives, N_{FP}denotes the number of false positives, N_{TN} denotes the number of true negatives, N_{FN}denotes the number of false negatives. In contrast to F-Measure and PSNR, the binarization quality is better for lower NRM.

Misclassification penalty metric (MPM)

The Misclassification penalty metric MPM evaluates the prediction against the Ground Truth (GT) on an object-by object basis. Misclassification pixels are penalized by their distance from the ground truth object's border.

$$\mathrm{MPM} = \frac{MP_{FN +} MP_{FP}}{2}$$

Where $MP_{FN} = \frac{\sum_{i=1}^{N_{FN}} d_{FN}^{i}}{D}$, $MP_{FP} = \frac{\sum_{j=1}^{N_{FP}} d_{FP}^{j}}{D}$

d^i_{FN} and d^j_{FP} denote the distance of the i^{th} false negative and the j^{th} false positive pixel from the contour of the text in the GT image. The normalization factor D is the sum over all the pixel-tocontour distances of the GT object. A low MPM score denotes that the algorithm is good at identifying an object's boundary.

CONCLUSION

A detailed survey about the principles of image binarization techniques is introduced in this paper. A comprehensive review is given. A number of classical methodologies together with the recent works are considered for comparison and study of the concept of binarization for both document and graphic images.

References

[1] Meng Bai, Chunheng Wang, Baihua Xiao Minhua Li a, "Conditional random field for text segmentation from images with complex background," Pattern Recognition Letters, pp. 2295-2308, july 2010.

[2] Maria T. Lopez, Jose Carlos Castillo Antonio Fernandez-Caballero, "Display text segmentation after learning best-fitted OCR binarization parameters," Expert Systems with Applications, pp. 4032-4043, 2012.

[3] G. Chavent, F. Clement & P. Weis H. Ben Ameur, "Image segmentation with multidimensional refinement indicators," Inverse Problems in Science and Engineering, vol. 19, pp. 577-597, july 2011.

[4] Themos Stafylakisa, Vassilis Katsourosa, George Carayannis Vassilis Papavassilioua, "Handwritten document image segmentation into text lines and words," Pattern Recognition, pp. 369-377, may 2009.

[5] M. Neubert , H. Herold & I. Niemeyer P. R. Marpu, "Enhanced evaluation of image segmentation results," Journal of Spatial Science, vol. 55, pp. 55–68, july 2010.