ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Document Image Binarization Using Independent Component Analysis For OCR

Varada Sreeja, G.Guru Prasad
M.Tech(CMS), Department of E.C.E, sree vidyaniketan Engineering College, India
Assistant Professor, Department of E.C.E, sree vidyaniketan Engineering College, India

Abstract

The Image binarization plays a vital role in text segmentation which is used in OCR application. Binarization of text in degraded images is a challenging task due to the variations in size, color and font of the text and the results be often affected by complex backgrounds, dissimilar lighting conditions, reflections and shadow. A robust solution to this problem can significantly enhance the precision of scene text recognition algorithms leading to a variety of applications such as scene understanding, navigation, automatic localization and image retrieval. In this paper, we propose a novel method to extract and binarize text as of images that contains complex background. We apply an Independent Component Analysis (ICA) based technique to map out the text region, which is uniform in nature, while removing specularity, shadows and reflections, which are included in the background. This algorithm works better on images with different degradations. We implement our method on various DIBCO datasets.

Keywords: Adaptive image contrast, ICA, pixel classification, pixel intensity, thresholding..

Introduction

In the recent years, content-based image analysis techniques have received more attention with the advent of various digital image capture devices. The images captured. By these devices can vary significantly depending on lighting conditions, reflections, shadows and specularities. These images contain numerous degradations such as uneven lighting, complex background, multiple colors, blur etc. We propose a method for removing reflections, shadows and specularities in natural scene text images and extracting out the text from a single image. There are many algorithms that aim at extracting foreground text as of background in images but thresholding remains one of the oldest form that is used in many image processing applications.

Many sophisticated approaches often have thresholding as a pre-processing step. It is often used to segment images consisting of bright objects against dark backgrounds or vice versa. It typically works well for images where the foreground and background are clearly defined. For color thresholding images, most algorithms convert the RGB image into grayscale but here we will make use of the RGB channel as three different sources.
Traditional thresholding based binarization can be grouped into two categories: the one which uses global threshold for the given images like Otsu, Kittler and the one with local thresholds like Sauvola, Niblack. In global thresholding methods, global thresholds are used for all pixels in image. These methods are fast and robust, but not appropriate for degraded images with complex and bright back ground. On the other hand, local or adaptive binarization methods change the threshold over the image according to local region properties. Adaptive thresholding addresses variations in local intensities throughout the image. These methods are proposed to overcome global binarization drawbacks but they can be sensitive to image artifacts found in natural scene text images like shadows, specularities and reflections. Mishra et al has recently formulated the problem of binarization as an MRF optimization problem. The method shows superior performance over traditional binarization methods on many images, and we use it as the basis for our comparisons. However, their method is sensitive to the initial auto seeding process. Zhou et al also addresses the segmentation problem in text images which contains specular highlights.

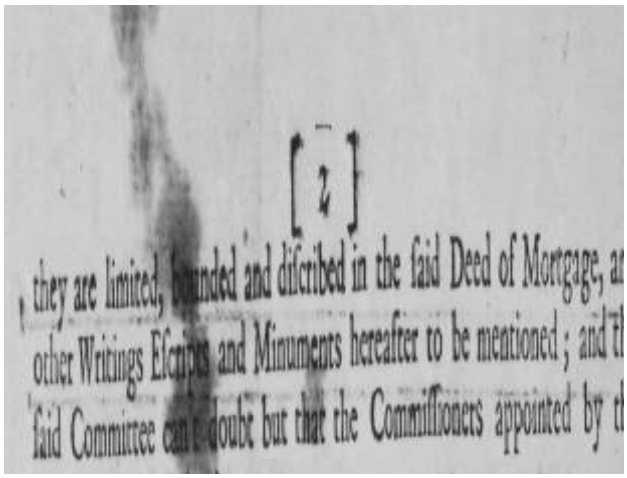
On the other hand, we propose a method that removes shadows, specularity and reflections and thus produces clean binary images even for the images with complex background. The primary issue

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

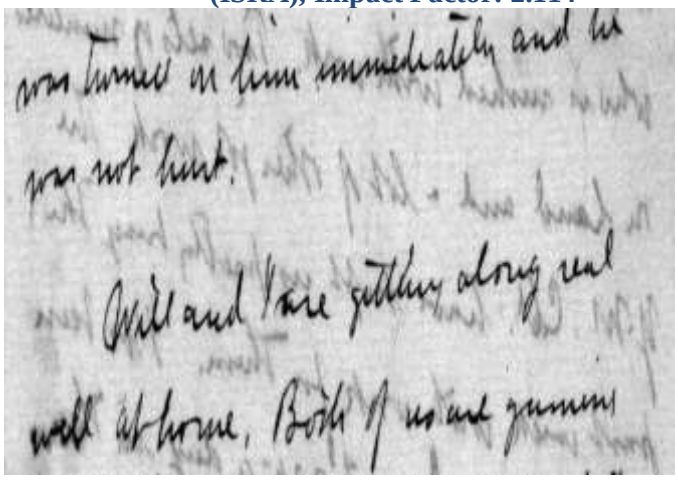
related to binarizing text from scene images is the presence of complex/textured background. When the background is uneven as a result of poor or non-uniform lighting conditions, the image will not be segmented correctly by a fixed gray-level threshold. These complex backgrounds vary dramatically depending on lighting, specularities, reflections and shadows. The above methods applied directly to such images give poor results and cannot be used in OCR systems.

In this paper, we do an ICA based decomposition which enables us to separate text from complex backgrounds containing, reflections, shadows and specularities. For binarization, we apply a global thresholding method on the independent components of the image and that with maximum textual properties is used for extracting the foreground text. Binarization results show significant improvement in the extraction of text over other methods. Some of the word images that are used for experiments are shown in Fig 1.

Figure.1:



(a)



(b)

(c)

Degraded Document Images from DIBCO Datasets.

Related work

Many thresholding techniques are used for document image binarization. Global thresholding is usually not a suitable approach for the degraded document which does not have a clear bimodal pattern. A better approach to deal with different variations within degraded document images is the adaptive threshold method. The early window-based adaptive thresholding techniques estimate the local threshold through the use of the mean and the standard variation of image pixels within a local neighborhood window. The main limitation of it is it depends on size of window and so the stroke width. The previous methods are back ground subtraction, recursive method, texture analysis, decomposition method, contour completion.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Proposed method

The proposed method first constructs the adaptive contrast map and then the text stroke edges are detected through the combination of binary map obtained through the independent component analysis and canny's edge map. Then by local thresholding the text is segmented based on the mean and standard deviation of the intensity of the pixels. To improve the quality of document binarization Post-processing is applied.

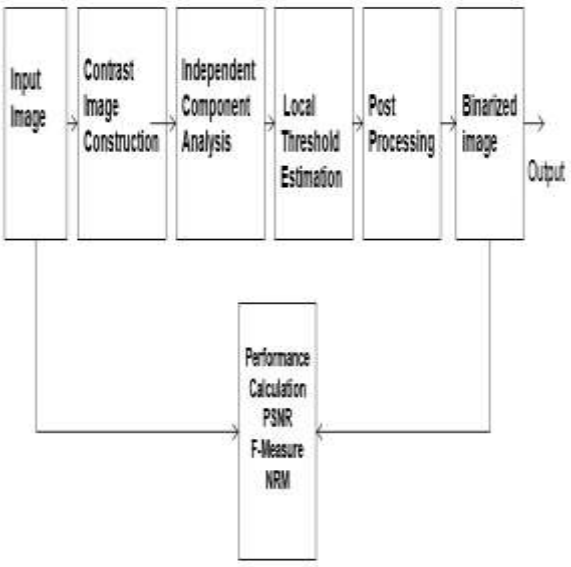
A. Contrast Image Construction.

B. Independent Component Analysis.

C. Local Threshold Estimator.

D. Post Processing Procedure.

Figure.2 :



Block diagram of the proposed method

Contrast Image Construction

Image gradient detects many non stroke edges from the background of degraded document that contains certain image variations due to noise, bleed-through, uneven lighting, etc.

$$C(i,j) = I_{\max}(i,j) - I_{\min}(i,j) \quad (1)$$

The image gradient is to be normalized to extract only the stroke edges of the image by compensating the image variation within the document background. The normalization factor that is used to suppress the image variation within the document background is the denominator. For pixels within bright regions of an image, it will produce a large normalization factor to neutralize the numerator and obtains low image contrast. For the pixels within dark regions of an image, it will produce a small denominator and obtains high image contrast.

$$C(i,j) = \frac{I_{\max}(i,j) - I_{\min}(i,j)}{I_{\max}(i,j) + I_{\min}(i,j) + \epsilon} \quad (2)$$

But for the images with bright text, this method is not suitable as it produces a weak contrast due to over normalization. So in order to overcome this over normalization problem, the local image gradient is combined with the local image contrast and obtains an adaptive local image contrast as follows:

$$C_a(i,j) = \alpha C(i,j) + (1-\alpha)(I_{\max}(i,j) - I_{\min}(i,j)) \quad (3)$$

The weight between the local image contrast and local image gradient is denoted by α. The local image contrast in Equation 3 gets a high weight for the document image with high intensity variation in the document background whereas the local image gradient gets a high weight for the document image.

Independent component analysis

Independent Component Analysis (ICA) has been an active research topic because of its potential applications in signal and image processing. The goal of ICA is to separate independent source signals from the observed signals, which is assumed to be the linear mixtures of independent source components. The mathematical model of ICA is formulated by mixture processing and an explicit decomposition processing.

The most frequently considered mixing model is the linear instantaneous noise free model is described as:

$$x_i = \sum_{j=1}^{n} a_{ij} s_j \quad (4)$$

Or in the matrix notation

$$X = A.S \quad (5)$$

Where A is an unknown full rank mixing matrix, which is also called mixture matrix and it assumes that there exists a linear relationship between the sources S and the Observations X. The

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

global thresholding method is used to binarize the independent components of the image and the binarized result is combined with canny's edge map obtained by applying canny's edge detector to the contrast image.

Canny's edge detector is used to mark the real edges of the image as it has a good localization property. The fore ground text is thus extracted from the document background as the stroke edge pixels of high contrast are detected.

Local threshold estimation

In order to perform the local thresholding procedure two characteristics are to be computed. They are mean and standard deviation of intensity of the detected image in a neighbourhood window W. So through comparing the intensity of the pixel with sum of mean *Emean* and standard deviation *E*std, text can be extracted as follows:

$$R(x,y) = \begin{cases} 1 & I(x,y) \le E_{\text{mean}} + \frac{E_{\text{std}}}{2} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

So the pixel with intensity less than sum of the mean and standard deviation is assigned to 1, otherwise 0.To contain stroke edges the neighbourhood window should be larger than the stroke width..

Post processing procedure

By obtaining the result from local threshold estimation then Post-Processing can be performed. The components of text stroke image are verified for connectivity and the non-connected components are removed. For remaining components, the four point connectivity is used to connect the same components that belong to same class. So through this procedure, the unwanted pixel components are removed which adds artifacts to the image.

Application

Foreign language data acquired via Arabic OCR is of vital interest to military and border control applications. Various hardcopy paper types and machine- and environment-based treatments introduce artifacts in scanned images. Artifacts such as speckles, lines, faded glyphs, dark areas, shading, etc. complicate OCR and can significantly reduce the accuracy of language acquisition. For example, Sakhr Automatic Reader, a leader in Arabic OCR, performed poorly in initial tests with noisy document images. We hypothesized that performing image enhancement of bi-tonal images prior to Arabic OCR would increase the accuracy of OCR output. We also believed that increased accuracy in the OCR would directly correlate to the success of downstream machine translation.

We applied a wide variety of paper types and manual treatments to hardcopy Arabic documents. The intent was to artificially model how documents degrade in the real world. Four hardcopies of each document were created by systematically applying four levels of treatments. Subsequent scanning resulted in images that reflect the progressive damage in the life-cycle of each document – the Manually Degraded Arabic Document (MDAD) corpus.

Scanning, OCR and measurement

Applying the assigned image enhancement settings, three types of images were captured for each document:

- Without image enhancement,
- With Fujitsu TWAIN32 image enhancement, and
- With both Fujitsu TWAIN32 and ScanFix image enhancement.

The MDAD corpus default scans already established the images without image enhancement. The dynamic threshold capability (i.e., SDTC) was disabled in order to gain full control of the scan brightness. Discovering the ideal brightness setting involved re-scanning and reducing the brightness setting repeatedly until white pixels appeared inside glyphs. The last scan with solid black glyphs was selected as the optimal scan. The three types of images for each document were then processed through the OCR tool. CP1256 files were output and compared against the ground truth using the UMD accuracy tool. We discovered that the evaluation metrics may not be reflecting the OCR output well. We have already mentioned that the OCR tool expects clean documents and on noisy documents it attempts to recognize speckles as characters. For noisy documents, the OCR tool produced several failure characters in the output file or caused Automatic Reader to abnormally end. Since accuracy was calculated as the number of correct characters minus error characters, divided by the number of correct characters, the tool produced negative and zero values.

Figure.3 :

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

(a)

(b)

Text localization and recognition results of proposed binarization method.

Experimental results

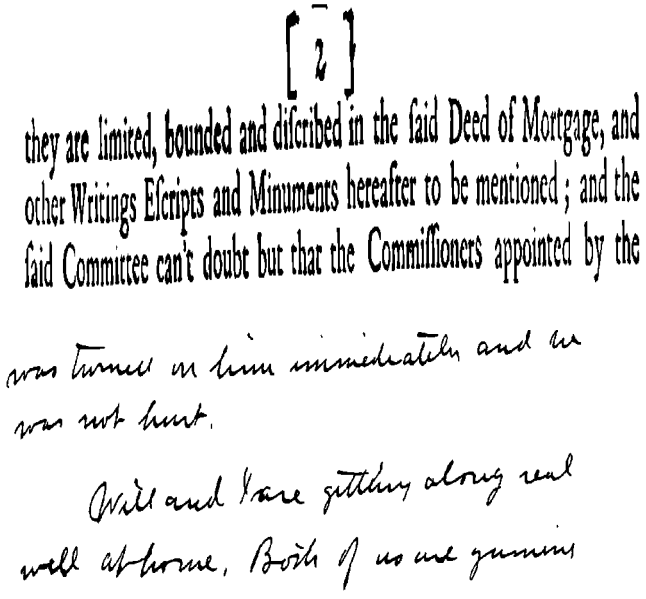
The results show that the proposed method is an effective method and performs better than other methods in the case where images have complex background. The proposed method can extract out the text embedded in complex reflective, shadowed and specular background. The proposed method obtains higher performance measures like F-measure, Peak signal to noise ratio, Negative rate metric and Mis-classification penalty matrix as shown in the table1.

Table-1: Simulation Results of the proposed method

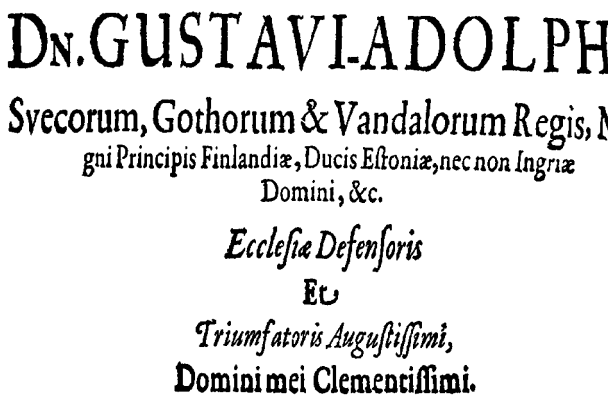
Type of Data set	F-measure	PSNR	NRM ($x10^{-2}$)	MPM ($x10^{-3}$)
DIBCO 2009	97.4023	22.50	2.597	0.928
DIBCO 2010	96.8633	24	3.136	0.922
DIBCO 2011	96.54	26	3.488	0.931

The binarized result obtained through the proposed method acquires better quality and segments the foreground text efficiently from the background of the document. The binarization results of the degraded document images that are shown in figure 1 is as shown in the figure 4. These document images suffer from different kinds of degradation, such as ink bleed-through, water stains and significant foreground text intensity and smear which are removed by the proposed method.

Figure.4:



ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114


DN. GUSTAVI ADOLPH
Svecorum, Gothorum & Vandalorum Regis, M
gni Principis Finlandiæ, Ducis Estoniæ, nec non Ingriæ
Domini, &c.
Ecclesiæ Defensoris
Et
Triumfatoris Augustissimi,
Domini mei Clementissimi.

Binarization results of the sample document images as shown in fig-1.

Conclusion

This paper presents an efficient document image binarization technique that is tolerant to different types of document degradation effects such as reflections, shadows, specularity, smear and uneven illumination. The proposed technique uses the Independent component analysis that is combined with the canny edge map to obtain text strokes for different kinds of degraded document images. The proposed method has tested on the various datasets. Experiments conducted show that the proposed method achieves efficient performance measures like F-measure, NRM, PSNR and MPM.

References

[1]. Bolan Su, Shijian Lu, and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images," IEEE transactions on image processing, vol. 22,no. 4,April 2013.

[2]. B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," International Workshop on Document Analysis Systems, pp. 159–166, June 2010.

[3] S. Lu, B. Su, and C. L. Tan, "Document image binarization using back-ground estimation and stroke edges,"Int. J. Document Anal. Recognit.,vol. 13, no. 4, pp. 303–314, Dec. 2010.

[4]. M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," IEEE Transactions on Image Processing, pp. 918–921, June 1998.

[5]. Siddharth Kherada, Anoop M. Namboodiri "An ICA based Approach for Complex Color Scene Text Binarization," pattern Recognition, 2nd IAPR Asian conference on Nov 2013.