ISSN: 2393-8528

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 3; May-June-2017; Page No. 117-121

Feature Extraction Based Document Image Processing For OCR

Dr. Anuj Kumar Parashar[1], Ajeet Kumar[2], Sandeep Kumar[3]

(Computer Science & Engineering) Dr.A.P.J Abdul Kalam Technical University, Lucknow (U.P), India

Assistant Professor, Department of Computer Science & Engineering, FET Agra College, Agra Uttar Pradesh, India[1]

anparashar@gmail.com[1]

Research Scholar, Department of Computer Science & Engineering, FET Agra College, Agra Uttar Pradesh, India[2]

ajeet555777@gmail.com[2]

Research Scholar, Department of Computer Science & Engineering, FET Agra College, Agra Uttar Pradesh, India[3]

sandeepk555777@gmail.com[3]

Received 14 May 2017; Accepted 10 June 2017

ABSTRACT

Image processing plays a vital role in document image processing system. For large scale of digitization process, various methods are available to provide an electronic version of a paper document, and scanning of the paper document is one of the best suitable methods. Optical scanning is the new technique applied on an image document, which converts the raw output data to the optical character recognition (OCR) system. Since the computer system cannot understand the language of the written documents, we need to convert these documents into the electronic documents, so that they can easily processed by the computer system. OCR converts the written text documents into the e- documents. In this paper we determine the threshold value of a scanned image document by using global thresholding method, which is based on the otsu's algorithm. On the basis of the threshold values obtain from the different methods, we can judge the quality of an image document and hence can improve the quality of an scanned image document.

Keywords: *Scanned documents, OCR, Thresholding and Document image processing.*

1. INTRODUCTION

Now a day's image processing plays a vital role in the field of scanned document processing. For large scale of digitization process, various methods are available to provide the electronic version of the paper document. Scanned images provide the digital record of the paper document. There are so many uses of the scanned documents in the private sector as well as in government sectors. Optical scanning is the technique which is applied on the scanned document, which forms the raw output of the optical character recognition system. The output produced by the OCR is the set of recognized characters. The Methodology employs the preprocessing of the scanned document, which improves the quality of the scanned document for the further processing of the document through OCR.

Preprocessing is the very fast step for the scanned document analysis so that the document being scanned gives more impressive results. The purpose of preprocessing is to improve the quality of the document being scanned. In this section, we first preprocess the image document the preprocessing technique employed here is the binarization. In this a scanned document is converted from color or grayscale into the bi-level representation The objective of making an image into binarized form, so that we can mark the pixels which belongs to the true foreground regions with single intensity and background region with different intensity.

Binarization of the scanned document is done by thresholding method in which the grayscale image is converted into the binary image. After the binarization of an image we cleaned the noise present in an image so that it will give the better results for OCR system.

2. Document Image Processing:

Document Image Processing is an electronic form of filing. In a DIP system, a document is passed through a

*Corresponding author: Ajeet Kumar |

scanner and a digitized image is then stored on a storage device perhaps an optical disk. This can then be retrieved and shown on a computer screen. The image of the document can include handwriting and diagrams. The process is the same as that employed in fax machine technology. That is, the image is recorded but the system does not identify the marks on the paper as letters or numbers. A scanner scans a whole page of input and records a pattern of dots, according to whether areas of the paper original are black or white.

3. Optical character recognition:

OCR is an abbreviation of optical character recognition; it is the recognition of printed or written textcharacters by a computer. This involves scanning of the text character-by-character, analysis of the scanned-in an image and then translation of the character image into character codes, such as ASCII codes, commonly used in data processing.

In OCR processing, the scanned image or bitmap is analyzed for background and foreground region in order to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code. Special circuit boards and computer chips designed expressly for OCR are used to speed up the recognition process.

4. Thresolding:

Thresholding is the technique which makes the grayscale image into the bi-level imageie.in the form of 0 or 1, in which '0' represent the black pixels and '1' represent the white pixels in the image document. The purpose of thresholding is to extract those pixels from the image document which represent the object, which can be written in either text or image.
There are different methods are available for thresholding:

(a) Global thresholding method

(b) Adaptive or local thresholding method

In **global thresholding method**,a single threshold is used for all the pixels of an image. When the pixel values of the components and that of background are fairly consistent in their respective values,over the entire image document.Each pixel in the document assigned to the whole page foregroundor background based on its gray values.Global methods are computationally inexpensive and they give good results for the scanned documents.

In **adaptive or local thresholding** method, different threshold values are used for different local regions in the image.It creates a black and white image pixels by analyzing each pixel with respect to the neighbourhood pixel.Adaptive method also give better results but this method is slower than the global thresholding method.

For binarization of the scanned document,we use the globalthresholding method,which is based on otsu's algorithm.

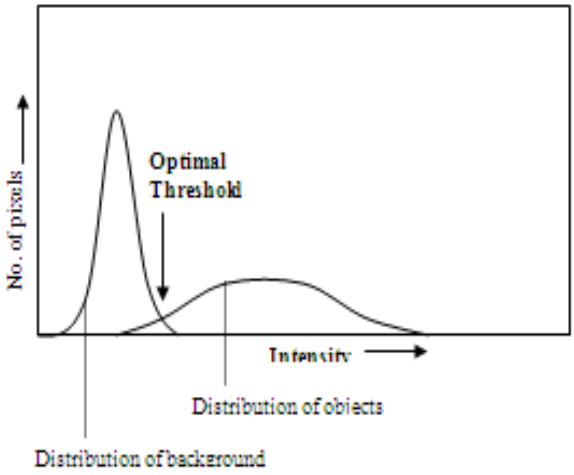

Figure 1:

5. Binarization:

A binarization method is the binarizing of an image by extracting the feature amount from the image. When a pixel is selected in an image, sensitivity is added to and/or subtracted from the value concerning the value of the selected pixel to set a threshold value range. Next, when another pixel is selected, the sensitivity is added to or subtracted from the value concerning the value of the selected pixel and a new threshold value range is set containing the calculation result and the already setup threshold value range. The pixel with the value concerning the value of any pixel in the image within the threshold value range is extracted as the same brightness as the selected pixel and the extraction result is displayed.

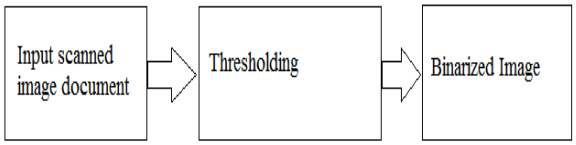

Figure 2: Binarization of scanned image document by thresholding

The above figure shows the preprocessing step which includes binarization. Binarization is done by thresholding called global thresholding method,which is based on *otsu's* algorithm.

6. *Otsu's* algorithm:

In 1979, otsu proposed an algorithm for automatic threshold selection from the histogram of an image. This is global thresholding method, which stores the

intensities of the pixels in an array. On the basis of threshold calculated by the otsu;s algorithm, each pixel is set to either 0 or 1,ie background(white) or foreground(black). Threshold of an image can be calculated by taking mean and variance of an image.

In this, the pixels are divided into 2 classes, C_1 with gray levels [1, ...,t] and C_2 with gray levels [t+1, ..,L]. The probability distribution for the two classes is given by:

$C_1: p_1/w_1(t),.....p_t/w_t(t)$ *and*

$C_2: p_{t+1}/w_2(t),.....p_L/w_2(t)$

$$\text{Where } w_1(t) = \sum_{i=1}^{t} p_i$$

$$\text{and } w_2(t) = \sum_{i=t+1}^{L} p_i$$

Also, the means for the two classes are:

$$\mu_1 = \sum_{i=1}^{t} i\, p_{i/}\, w_1(t)$$

$$\mu_2 = \sum_{i=t+1}^{t} i\, p_{i/}\, w_2(t)$$

Also using Discriminant Analysis, Otsu defined the between-class variance of the thresholded image as

$\sigma_B^2 = w_t (\mu_1 - \mu_2) + w_2 (\mu_2 - \mu_1)$

For bi-level thresholding, Otsu verified that the optimal threshold t* is chosen so that the between-class variance $_B$ is maximized; that is

$t^* = \text{Arg Max} \{ \sigma_B^2 (t)\}$

$t<L$

The advantage of otsu method is that otsu method is very simple and easy to calculate. Since it is global algorithm, thus it is well suited for the image has equal intensities.

7. Noise Cleaning:

Noise cleaning from the scanned document in this paper is done by erosion and dilation technique. Both the operations are the fundamentals to the morphological processing. Erosion is a shrinking operation, while dilation grows or thickens the objects in a binary image. In case of dilation,the value of the output pixel is the maximum value of all the pixels in the input pixel's neighborhood. In a binary image, if any of the pixels is set to the value 1, the output pixel is set to 1.While in case of erosion,the value of the output pixel is the minimum value of all the pixels in the input pixel's neighborhood. In a binary image, if any of the pixels is set to 0, the output pixel is set to 0.

8. Result and conclusion:

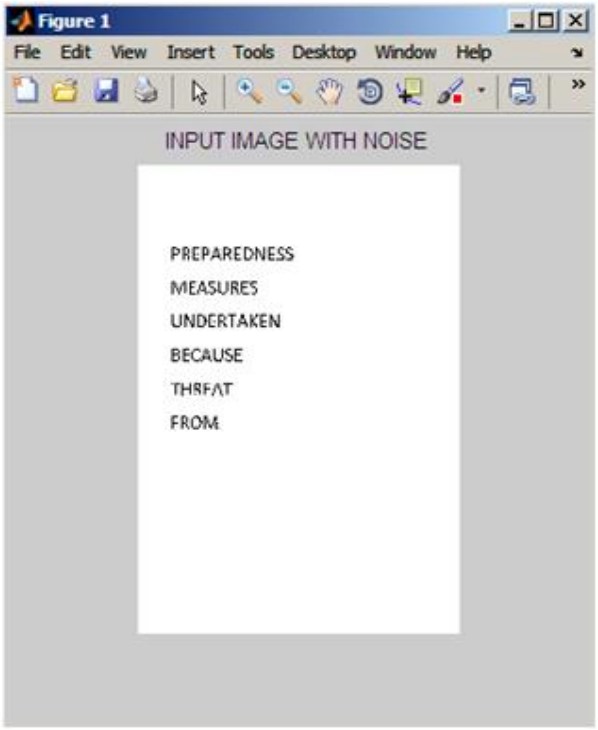

Figure 6.1.1: Input image with noise

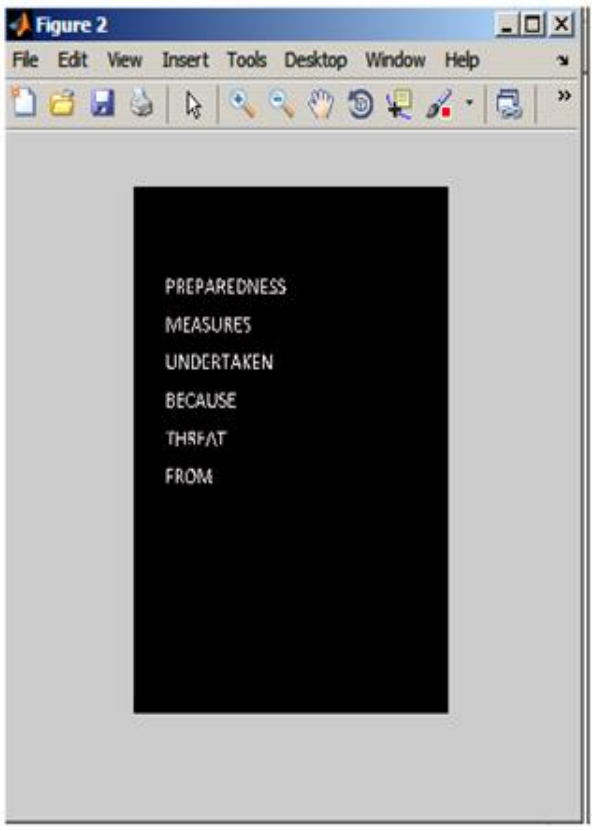

Figure 6.1.2: Preprocessing of input image

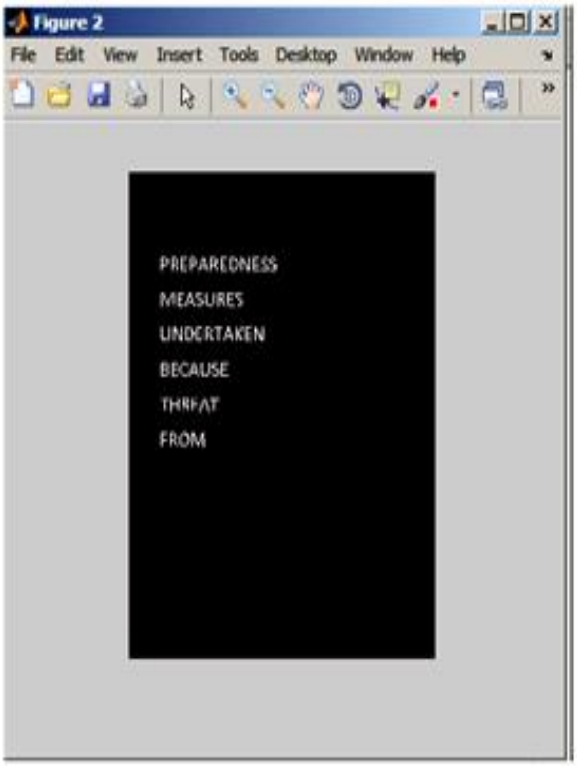

Figure 6.1.3: Image after removal noise

Figure 6.1.4: Line or string Detection from the scanned image

Fig 6.1.5: Detection of connected components of a string

Figure 6.1.6: Line or string Detection from the scanned image

Figure 6.1.7: Detection of connected components of a string

Output Result In Editable :

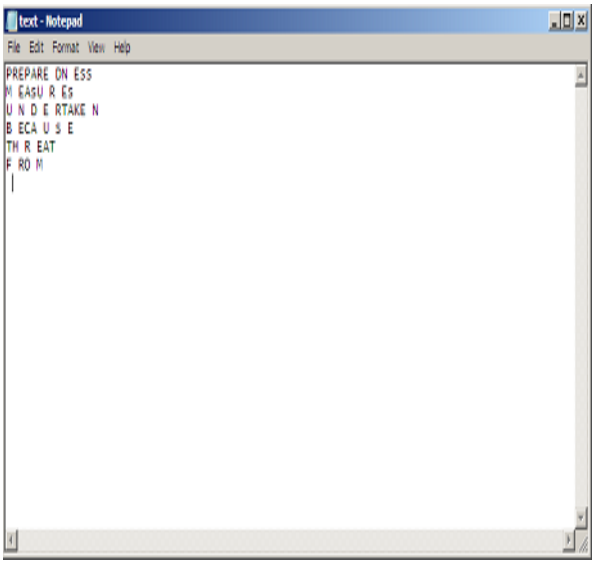

Figure 7:

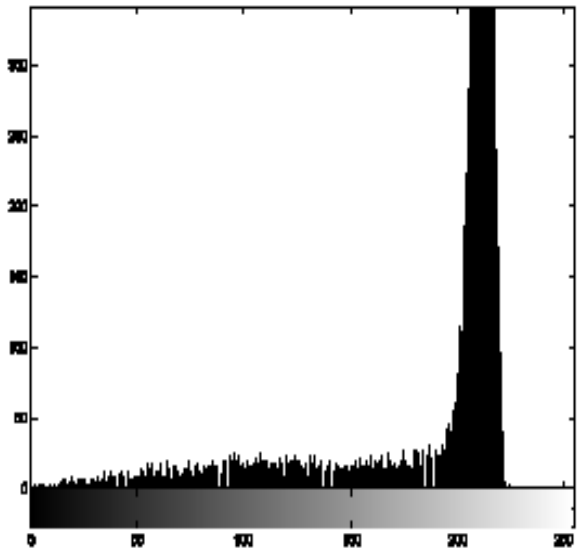

Figure7: Image histogram

Threshold value calculated using otsu's method=123

The result obtained from the various scanned document images, we had concluded that the OCR cannot process the noisy image as well, so we clean the noise present in the image, which works well on the OCR. The threshold level calculated by otsu's method is much better than the threshold calculated by other methods such as peak valley method. In this we also concluded that erosion and dilation method applied on the image give better results of noise cleaning than other methods such as n-connectivity method.

9. Future work:

In the future we will try to apply this method to the various images which are in different languages such as English text, Gurumukhi (Punjabi script),Devnagri (Hindi script) and try to achieve the best results for

these different languages and can make the text of these image into editable form.

References

1. J.Pradeep, E srinivasan and S.Himavathi , "*Diagonal Feature Extraction Based Handwritten Character System Using Neural Network*" International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010.
2. S.V. Rajashekararadhya, and P.Vanajaranjan, "*Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south-Indian scripts*" Journal of Theoretical and Applied Information Technology, JATIT vol.4, no.12, pp.1171-1181, 2008.
3. Anil.K.Jain and Torfinn Taxt, "*Feature extraction methods for character recognition-A Survey,*" Pattern Recognition,vol. 29, no. 4, pp. 641-662, 1996.
4. Nitin Khanna and Edward J. Delp "*Source Scanner Identification for Scanned Documents*" Video and Image Processing Laboratory School of Electrical and Computer Engineering Purdue University West Lafayette, Indiana USA
5. Shang Jin1, Yang You2, Yang Huafen3 "*A Scanned Document Image Processing Model for Information System*" 2010 Asia-Pacific Conference on Wearable Computing System.
6. Otsu, N., 1979. "*A threshold selection method from gray-level histograms"*. IEEE Trans. Systems, Man, and Cybernetics, 9(1), pp. 62-66.