Threshold Approach to Handwriting Extraction in Degraded Historical Document Images

Sangeeta Lalwani
M.Tech (CS&E)
Amity University
Noida, India

Piyush Saxena
M.Tech (CS&E)
Amity University
Noida, India

Amarpal Singh
M.Tech (CS&E)
Amity University
Noida, India

ABSTRACT

Handwriting extraction is the skill of a system to get and translate comprehensible hand written input via sources such as document, photos, tough screen and other devices. The picture of the written document is used to detect written text by the use of optical scanning i.e. known as optical character recognition. Handwriting extraction basically uses optical character recognition. Conversely, an absolute hand writing extraction process that handles format and perform correct segmentation into typescript and searches for the most reasonable terms.

Handwriting extraction is a process of automatic typesetting of text from a picture to letter sets that are exploitable by a system or a computer by the use of text- processing software. The information received via this method form is treated as static illustration of hand writing. Off line handwriting recognition is relatively complex due to the reason that different persons have differences in the handwriting styles.

Today, Optical Character Recognition engines mainly focus on instrument printed text and Intelligent Character Recognition for hand written text.

The proposed system uses the above mentioned key features with going one step further. One of the most impressive aspects of human visual processing is the ability to recognize objects despite severe degradations in image quality. The paper focuses on the recognition of impoverished handwritten documents.

General Terms

Optical Character Recognition (OCR), Intelligent Character Recognition (ICR), Binarization, Ocular Scanning (optical character extraction)

Keywords

Optical Character Recognition, Intelligent Character Recognition, Rank Conditioned Rank Selection.

1. INTRODUCTION

Handwriting extraction is the capability of a system to accept and deduce hand written input from sources such as photographs, paper documents, touch-screens and other devices. The picture of the typed text may be sensed offline via a portion of printed paper by ocular scanning (optical character extraction) or intelligent word extraction. Handwriting extraction mainly entails optical character recognition. Besides this, an handwriting extraction system is also used to perform formatting, accurate segmentation into characters and finds the most conceivable words.

The proposed system involves three main steps which are necessary before handwriting extraction. The loaded image is firstly binarized. A binarized image is an image which has maximum two possible values for each pixel. Characteristically, a binary image usually has two colors that are black and white. Foreground color is being used for the object pixels in the picture whereas background color is used for the rest of the image pixels. This is known as bi-tonal image in document scanning industry. Successful completion of this step leads to the removal of noise from the image. Third step is to segment the image. The process of segmentation involves partitioning of a digital image into several segments. The aim of splitting into segments is to make things easier and modify the illustration of a picture as something that is more significant and simpler to investigate. After the segmentation part is over, the individual characters are recognized from a pre-processed training data and the result is produced [1].

2. LITERATURE SURVEY

Historical credentials have significant and appealing information. A number of methods have in advance been projected for thresholding document images. Thresholding chronological handwritten document image converts the gray scale image to binary plan by unscrambling the functional font and information from the background [3]. Old versions desired to be programmed with images of each character, and converted on one font at a time. "Intelligent" system is a scheme with a high level of correctness for the recognition of the majority of fonts.

Few systems are skilled for reproducing formatted result that directly approximates the original scanned page including images, columns and other non-textual parts [9, 12].

Intelligent Character Recognition (ICR) is an advanced optical character recognition (OCR) technique that allows diverse handwriting styles and fonts to be erudite via a computer while processing to improve a variety of recognition levels and accuracy [8, 9]. Many of the ICR software have a self-learning system known as neural network, with the capability to expand the helpfulness of scanning device which is used for document processing, and for printed character recognition to hand-written material detection and updates the acknowledgment database for new hand writing patterns. As this process is involved in recognizing hand writing therefore correctness levels may not be very high-quality but can achieve 97% or more accuracy rates in reading handwriting in ordered form. Often few read engines use this program to attain high recognition rates and each is given elective voting rights to resolve the true conversion of characters. In alpha fields, engines are intended to read hand written characters and have higher elective rights while in numeric fields, engines are intended to read numbers.

3. OBJECTIVE

The vision involves three steps which are necessary before handwriting recognition / extraction. The original user given image is firstly binarized. A binary image is a digital image that has only two possible values for each pixel [6]. Typically the two colors used for a binary image are black and white though any two colors can be used. The color used for the object(s) in the image is the foreground color while the rest of the image is the background color. In the document scanning industry this is often referred to as bi-tonal. Successful completion of this step leads to the removal of noise from the image.

This second step is crucial so as to remove the minute particles from the image for accurate recognition [4]. Images taken with both digital cameras and conventional film cameras will pick up noise from a variety of sources. And historical documents are often degraded due to time which attracts noise in various forms.

Third step is to section the image. Segmentation is the method of partitioning a digital image into multiple sections. The goal of segmentation is to make simpler and/or alter the depiction of an image into something that is more significant and easier to examine. After the segmentation part is over, the individual characters are recognized from the dictionary and the result is produced.

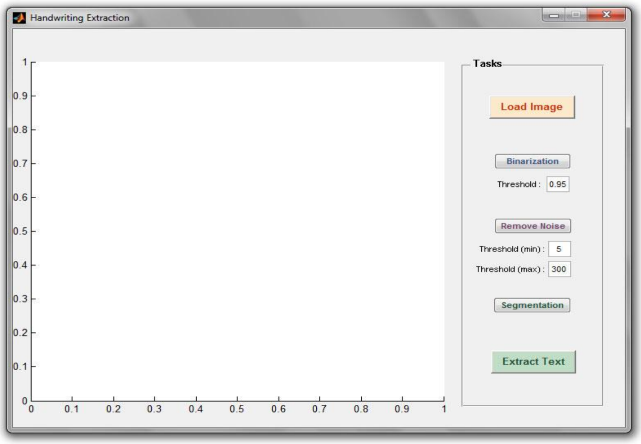
4. RESEARCH METHODOLOGY

The research methodology comprises of full and in depth study and analysis of all the algorithms [3] such as image capturing and image extraction using image segmentation and binarization and others as and when required. As historical documents contain important information and facts, their preservation is necessary. Character recognition frequently involves scanning a structure or document sometime in the past. Tools exist that are proficient of performing the step however, a number of frequent imperfections in this step. The frequent are characters that are coupled together are returned as a solitary sub-image having both characters. This causes a chief difficulty in the recognition stage. Once the withdrawal of individual characters occurs, recognition engine is used to identify the corresponding computer character.

ICR is an advanced OCR or handwriting recognition system which allows fonts and different handwriting styles to be trained by a system throughout processing to get better correctness and identification levels. A number of techniques that have been formerly projected for thresholding document images. All formerly reported thresholding methods have been established to be efficient for certain classes of document images. Along with this a lexicon is also created. Knowledge of what all algorithms can work upon for simple text images and applying them on a given text image is the major goal. The final motive is to get a noise free text image.

5. IMPLIMENTATION OF THE APPROACH

The main interface of the system represents the outlook of the whole system. Please refer to fig.1 and fig.2



Threshold : 0.95

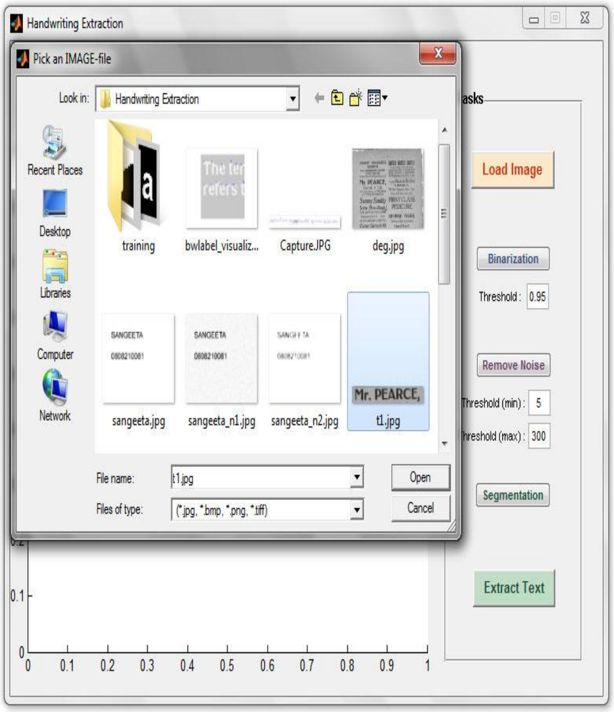
Threshold (min) : 5

Threshold (max) : 300

Fig1: Main interface

The proposed system contains 5 buttons namely

- Load Image
- Binarization
- Remove Noise
- Segmentation
- Extract Text



File name: t1.jpg

Files of type: (*.jpg, *.bmp, *.png, *.tiff)

Threshold : 0.95

Threshold (min) : 5

Threshold (max) : 300

Fig2 :Image Loading interface

1. Image Acquisition:

The first stage of any vision system is the image acquisition stage. After the image has been obtained, various methods of processing can be applied to the image to perform many different vision tasks required today. Image acquisition involves capturing the image from any source. After the image has been obtained, various methods of processing can be applied to the image to perform many different vision tasks required today [5]. However, if the image has not been acquired satisfactorily then the intended tasks may not be achievable, even with the aid of some form of image enhancement. In this implementation the scanned images or pictures captured from digital cameras are taken into consideration. The images are loaded into the primary screen as shown in Fig.2. Various formats are supported for this which includes JPEG, BMP, PNG and TIFF. After this the image contrast can then be increased by using histogram equalization. Histogram equalization adjusts the contrast of an image by using image's histogram as shown in figure 3 below. Through this adjustment, the intensities can be better distributed on the histogram.

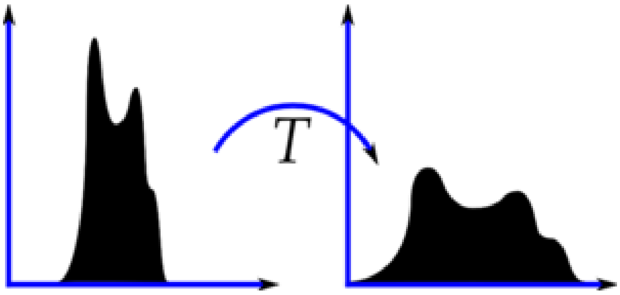


Fig 3: Histogram Equalization

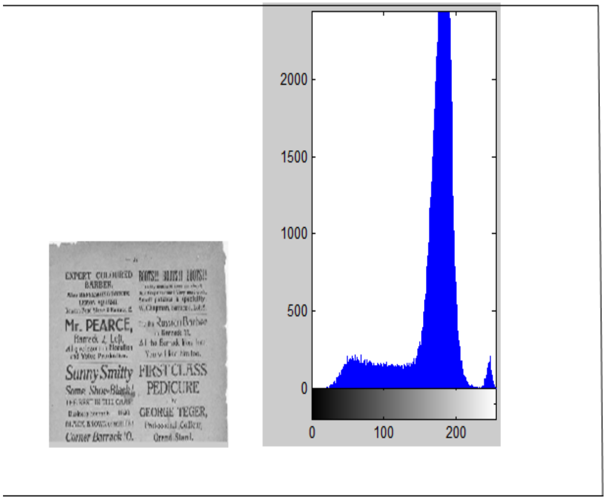


Fig 4: Image after Applying Histogram Equalization

2. Image Binarization:

Document image binarization [6] is an important basic task needed in most document analysis systems. A binary picture is an image that has maximum two possible values for each pixel that are black and white. The binarization is done on the basis of a threshold value given by the user. Binary images are obtained in digital image processing as masks or as the result of certain operations such as segmentation, thresholding, and dithering. In this module, the image is first converted into grayscale and later converted into binary.

The binarization is done on the basis of a threshold value given by the user. The threshold value determines the quantization level over which binarization is done. During the thresholding process, individual pixels in an image are marked as "object" pixels if their value is greater than some threshold value (assuming an object to be brighter than the background) and as "background" pixels otherwise. This convention is known as threshold above. Variants include threshold below which is opposite of threshold above, threshold inside, where a pixel is labeled "object" if its value is between two thresholds and threshold outside, which is the opposite of threshold inside the image. Typically, an object pixel is given a value of “1” while a background pixel is given a value of “0.” Finally, a binary image is created by coloring each pixel white or black.

There are several different methods for choosing a threshold exist, users can manually choose a threshold value, or a thresholding algorithm can compute a value automatically, which is known as automatic thresholding. A simple method would be to choose the mean or median value, the rationale being that if the object pixels are brighter than the background, they should also be brighter than the average. In a noiseless image with uniform background and object values, the mean or median will work well as the threshold, however, this will generally not be the case [9]. The image after binarization has only two colors i.e. black and white.

3. Noise Removal:

There are two types of recording devices, i.e. analogue and digital and they both have features that make them vulnerable to risks like noise. The user is required to input the minimum and maximum threshold values between which the noise will be removed. The pixel values lying below the minimum threshold value and the pixel values lying above the maximum threshold values are removed from the resulting image [5]. The user is required to input the minimum and maximum threshold values between which the noise will be removed. The pixel values lying below the minimum threshold value and the pixel values lying above the maximum threshold values are removed from the resulting image. Through noise removal unwanted particles (pixels) are removed from the binarized image so that the segmentation could be done properly without taking tiny pixels into consideration. In this task, the pixels are traversed one by one, line by line.

4. Image Segmentation:

It is the method of dividing a digital picture into various parts. The aim of segmentation is to make simpler and to change the illustration of a picture into something that is further significant and simpler to investigate [8]. Image segmentation is characteristically used to situate stuffs and borders in images. The method of image segmentation assigns a tag to every pixel in an image in a manner that pixels with the same tag share definite image uniqueness.

The outcome of image segmentation is a set of segments that jointly wrap the whole image, or a set of contours extracted from the image [14]. All the pixels in a region are alike wrt some attribute or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same attribute(s).

The simplest process of image segmentation is called the thresholding process. This process is based on a clip-level (or a threshold value) to turn a gray-scale image into a binary image. The process is to select the threshold value. Several popular methods are used in industry including the maximum entropy method, Ostu's method (maximum variance), niblack method etc. K-means clustering can also be used.

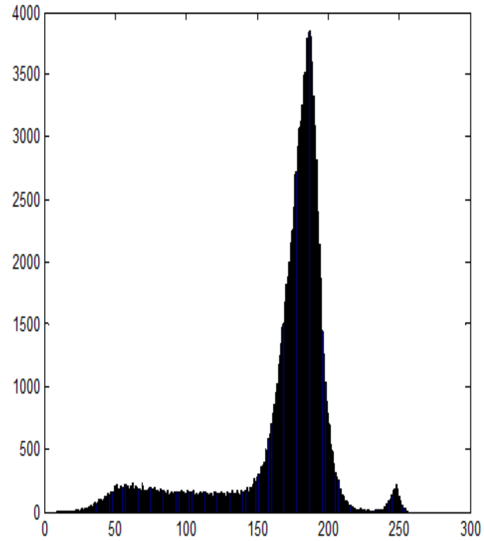


Fig 5: Histogram after applying niblack's algorithm

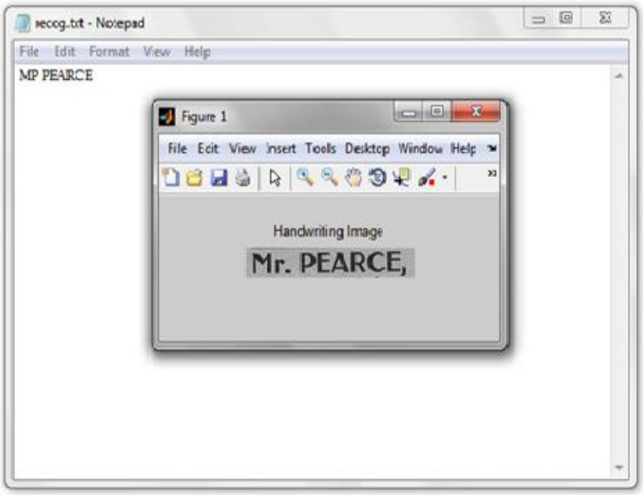
Edge detection is a multifaceted job within image processing. Region boundaries and edges are closely related, since there is often a sharp adjustment in intensity at the region boundaries. The edges identified by edge detection are often disconnected. To segment an object from an image however, one needs closed region boundaries. The desired edges are the boundaries between such objects. Edge detection techniques have therefore been used as the base of another segmentation technique. These are the characters which will be matched with the existing training set for character recognition. Segmentation methods can also be applied to edges obtained from edge detectors. After segmentation, the individual characters are distinguished by colored rectangular boxes. These are the characters which will be matched with the existing training set for character recognition.

5. Text Extraction:

The final and the most important module of the implementation is handwritten text extraction. Before commencing with the text extraction, we need to create a training data set against which the characters will be matched [9]. The training data set is created using 26 capital alphabet and 10 numeral images. Although the accuracy is not 100% but it surely recognizes maximum handwritten characters accurately. There are two important functions which are used namely lns() and ltr(). The lns() function is used to iterate for every line in the document and the ltr() function is used to iterate for every character in a line. It matches each and every character from the training data set. After the successful creation of the training data set and segmentation of the text, each character (connected components) is virtually labeled and is matched with the training data set created earlier.

6. RESULT OF THE APPLICATION

For the input shown in figure 6 below, the output is around 95% accurate.


recog.txt - Notepad
File Edit Format View Help
Figure 1
File Edit View Insert Tools Desktop Window Help
Handwriting Image

MP PEARCE

Mr. PEARCE,

Fig 6: Text recognition by the application from the image.

The flow chart below shows the desired output:-

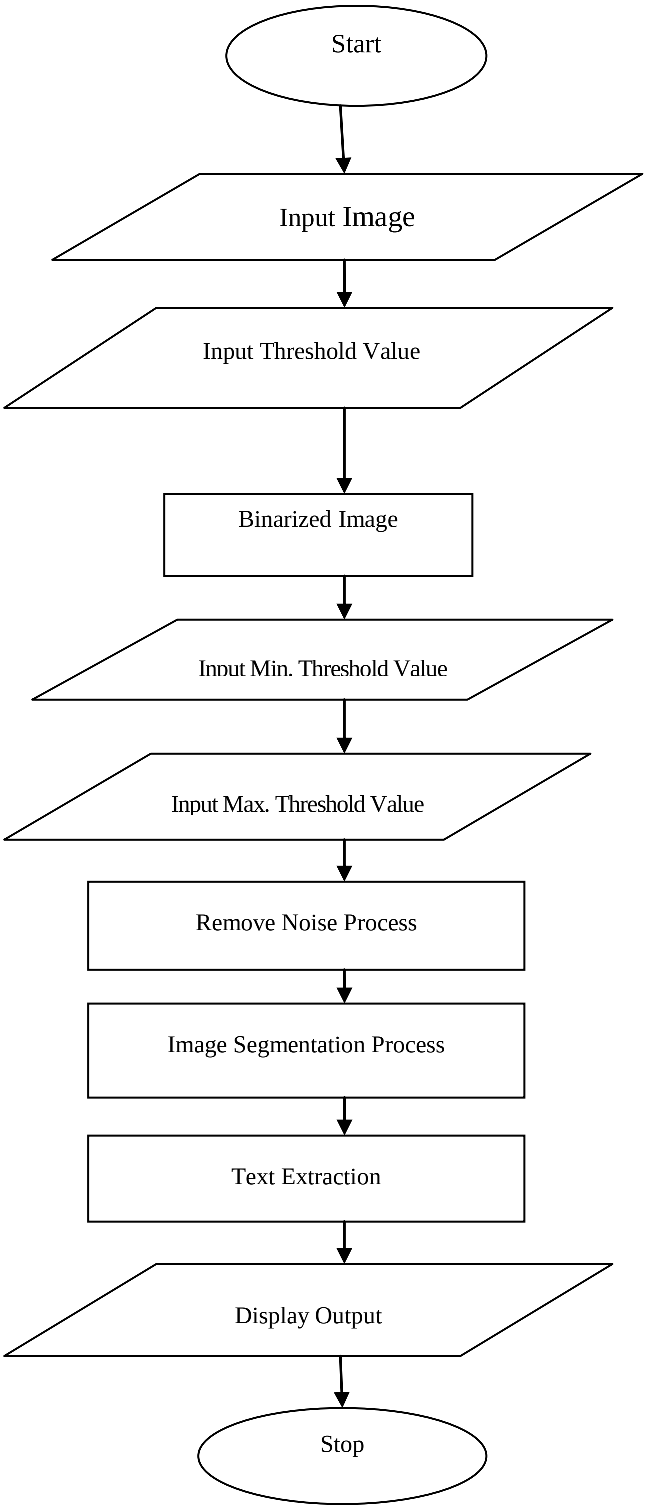


Fig 7: Flowchart Representation of Text Extraction

7. CONCLUSION

This paper has presented a comparison of several binarization algorithms by measuring their end-to-end word recognition performance on archive document word images. Described algorithms utilize spatial structure, global and local features or both. Many algorithms require extensive pre-processing steps in order to obtain useful data to work with because document image and data mining classification techniques is still in infancy.

The conclusion is that no single algorithm works well for all types of image but some work well than others for particular types of images suggesting that improved performance can be obtained by automatic selection or combination of appropriate algorithm(s) for the image that is being under examination. Also, researcher could improve the post processing step such as by adding edge detection techniques and further enhanced by an innovative image refinement technique and a formulation of a class proper method.

Somehow, other alternative technique also could be tested like binarization using directional wavelet transforms hybrid thresholding and high performance adaptive binarization. Furthermore, to enhance the evaluation method, researcher should take into consideration for using Jawi Optical. Character Recognition rather than human assessment due to human limits and view error. Historical images mostly exhibit degraded characteristic qualities after years of storage. Satisfactory thresholding results can rarely be obtained if same process is applied to the entire image. The threshold approach is effective at resolving this problem[9].

It uses local feature vectors for analysis and hence to find the best approach for thresholding local area. Appropriate algorithm(s) is selected or combined for specific types of document image under investigation automatically. The original image is recursively broken down into sub-regions using quad-trees until an appropriate thresholding method can be applied to each of the sub-region.

The approach outperforms existing single methods by measurement of ‘recall’ value. The future application of this technique can contribute to other difficult document images, such as newspaper images and cheques.

8. FUTURE SCOPE

The scope is very wide. The characters from the historical document image are matched accurately and the result is found to be approx. 95% accurate. In most of the cases this process of text extraction yields correct result.

The process of text extraction is limited to examine small alphabets and the algorithm is implemented to yield correct results. Also spacing between two letters has not been considered in the paper.

Hence, there are a lot of things to be worked out in this paper like-

1. Recognizing Capital Alphabets.
2. Since the correct result varies in some cases, hence there can be a dictionary which can be created in MS-Word which could recognize the correct spelling of any word which has not been recognized accurately.

9. REFERENCES

[1] Chanda H. and Dutta D.,“Digital image processing and analysis”, Prentice Hall of India, 2005.

[2] Eikvil L., Taxt T.and ,Moen K., “An adaptive method for binarization of grey level images”, NOBIM National Conference on Image Processing and Pattern Recognition ,June 1991, pp 123-131.

[3] Esakkirajan S,Veerakumar V., “Digital image processing “,Tata McGrawHill publications,third edition.

[4] Gonzalez R., Woods R. ,” Digital Image Processing Using MATLAB” second edition.

[5] Jain K., “Fundamentals of digital image processing” ,Prentice Hall of India, 2002, fifth edition.

[6] Kavallieratou E. and Stathis S.(2006).Adaptive Binarization of Historical Document Images. Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), pp 742-745.

[7] Kenneth R.,”Digital image processing”,New Jersey: Prentice Hall, 1996, second edition.

[8] Meyer U.,” Digital signal processing using programmable gate arrays”,second edition.

[9] Nick E.,“Digital Image Processing”,Pearson Education Asia, 2000.

[10] Pratt W. “Digital image processing using matlab technology” ,tata mc-graw hill,college edition.

[11] Rafael C. and Enrich R.,“Digital image processing” Pearson Education, 2002., second edition.

[12] Retsch G.,”Solutions in Particle Size- and Shape-Analysis “,third edition.

[13] Sezgin M., Sankur B., “Survey over image thresholding techniques and quantitative performance evaluation.” Journal of Electronic Imaging 13(1), 146– 165 (January 2004)”.

[14] Shapiro L., Stockman G., “Computer Vision”.publisher Prentice Hall1, first edition (February 2, 2001).