AN ADVANCED, ROBUST BINARIZATION AND CHARACTER RECOGNITION IN DOCUMENT IMAGES BASED ON RASPBERRY PI

Anitta Vincent and Gincy Varghese
Department of ECE, Sahrdaya college of Engineering, Thrissur, Kerala, India
E-Mail: anittavins@gmail.com

ABSTRACT

Digitization is a major scenario through which transactions and information sharing occur very fast. The notion of ‘paperless office’ is gradually coming into existence. Besides, there are a plethora of treasured documents which need to be safeguarded. In such a plot, an adequate system for reclamation and digital conservation of documents becomes indispensable. Here, a robust system for binarization and character recognition of deteriorated document images is proposed. Segmentation of text from deteriorated document images is a very tough task due to the high variation within and between the document background and the foreground text in different document images. The binarization technique employed here is a novel technique which utilizes a modified adaptive image contrast based method. This method can tolerate non-uniform background and various artifacts that creep in. Next step is the image segmentation, followed by character recognition using Artificial Neural Network. The entire system is implemented on a low cost minicomputer board named Raspberry Pi.

Keywords: adaptive thresholding, binarization, segmentation, optical character recognition, artificial neural network.

INTRODUCTION

Digital image processing is a wide and deep domain which gained wide popularity due to its diverse applications in all walks of life and living. The growth of computers and other digital technology has accelerated its growth. We deal with different kinds of document images every day. Document image processing approaches help distillate useful information from document images.

Traditionally, our main form of transmission & storage for information has been by paper documents. These documents include many common types: letters, forms, text books, technical manuals etc. Earlier many documents existed on paper. Today in the digital era, the world is going paper-less. ‘Paperless office’ concepts are seeping in every domain. This requires processing of documents using image processing techniques. The need for paper is eliminated by replacing letters and faxes by email, reference books by the cyber world and so on [1]. Another way to eliminate paper is to automate paper-predicated processes that use forms, applications etc. This paper presents an idea which can act as the stepping stone towards a paper-less office which can give faster processing and transactions. The paper gives a technique for efficient pre-processing and character recognition in documents.

Optical Character Recognition (OCR) is the Conversion of scanned images of printed/handwritten text to machine encoded/ computer-readable text. For recognizing the characters in the document image, the major pre-processing step is the binarization. Binarization is the process of converting a colour/gray-scale image into a binary image that has only two possible values for each pixel. Typically the two colours used for a binary image are black and white though any two colours can be used. The efficiency of OCR is determined by the quality of binarization.

Segmentation of text from badly degraded document images is a very challenging task due to the high variation within and between foreground and background of different document images. In contrast to conventional approach for thresholding a document like global thresholding, window based adaptive thresholding etc, a novel document image binarization technique by using adaptive image contrast is proposed [2]. The adaptive image contrast is an amalgamation of the local image contrast and the local image gradient that is tolerant to text and background variation caused by different types of document degradations. In this paper, an adaptive map showing the contrast is first constructed for an input degraded document image. The contrast map is then binarized and logical AND operation is performed with the edge map to identify the text stroke edge pixels. The document text is again segmented by estimating a local threshold. It is based on the intensities of detected text stroke edge pixels within a local window. The proposed method is of the advantage that it is robust, and involves minimum parameter tuning.

The above algorithm is evaluated on various edge detectors other than Canny and was qualitatively compared based on PSNR. The assessment revealed that some other detectors performed better in this regard and offered better results. So it can be said that the modified algorithm provides an advanced binarization technique for Optical character Recognition.

The need for a low cost hardware for binarization and OCR of document image is still a challenge. A hardware based stand alone system will be of great advantage.

There are sundry application for the above concepts such as in conversion of books to digital libraries, sorting of sizably voluminous document datasets (licit, historical, security), Search engines on the Web, Realization of 'Paper-less office' concept etc. It can be applied in PDA or tablet PC technology.

The paper is organized as follows. The chapter II gives a brief idea about the existing methods. Chapter III gives an overview of the proposed system. This is followed by description of binarization and OCR. Implementation, results and analysis, conclusion, future plans and a reference list follows.

EXISTING METHODS

Although the subject of character recognition has been the center topic for many years, it is one of the most demanding areas in pattern recognition. There are varieties of software based solutions available for OCR. However, there is little work done in the area of hardware implementation. A DSP centric OCR system based on TMS in Arabic language is described in [3]. Another hardware oriented OCR system for musical notations based on FPGA is described in [4]. The need for a low cost hardware for binarization and OCR of document image is still a challenge.

Binarization is the major pre-processing method that determines the quality of OCR. A binary image is digital image that has only two possible values of intensity for each pixel. Binarization deals with converting a given image into binary format. Popular techniques for binarization are Frequency domain techniques like Phase congruency method[5] which find locations in an image where all sinusoids in the frequency domain are in phase, Global thresholding[6], Bernsen's method[7], Local Image maximum-minimum method[8] which introduces a normalization factor to compensate the image variation within the document background.

Disadvantages of above methods are listed below. Phase congruency method is very much intensive in computations and sensitive to image noise. Degraded documents do not have a clear bimodal pattern; so global thresholding is usually not a suitable approach. Bernsen's method is simple, but cannot work properly on degraded document images with a complex document background. Local maximum-minimum method may not be fitting for document images with the bright text. This is because for such stroke edges of bright text a weak contrast will be calculated.

OCR or character recognition is the conversion of images of typewritten or handwritten documents into computer readable text. Methods extensively found in literature are: SVM [9], Template-matching, Correlation [10], etc. Template-matching makes use of the method of comparing input characters to pre-defined templates. This method identified characters either as an exact match or not at all. Other method is Recognition using Correlation Coefficients which uses cross correlation of input characters with the templates stored in database; this helps accommodate minor differences. Disadvantage of template matching is that it couldn't accommodate effects like slants and mode difference that didn't involve major shape alterations. Correlation technique introduced false or erroneous recognition among characters very similar in shape like 'B', '8' and such similar pairs [10]. SVM or Support Vector Machine is considered ideal for OCR in scenarios like license plate detection; but is not very good when it comes to document images where the text size is viable to change.

SYSTEM OVERVIEW

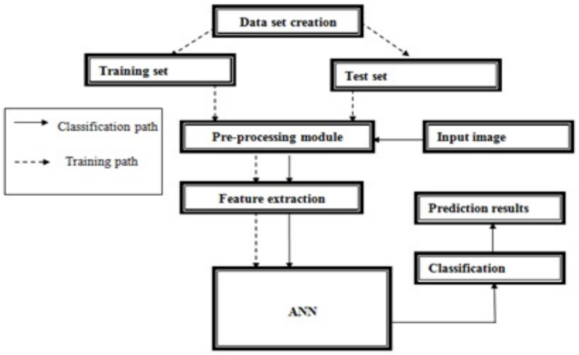

Figure-1. Block diagram.

The block diagram of the entire system is given above. First step is the creation of templates for training the ANN. The data set is created based on 17 popular fonts and each letter is cropped to a fixed standard size. Followed by is the pre-processing which binarizes the image efficiently based on adaptive contrast based method and then segments it. Binarization' is the first step which provides uniformity to all the input images. Other properties like dissimilarity, sharpness etc. can also be easily taken care of once the image is binarized. More explanations of binarization are dealt with in detail in the coming section. Segmentation can be done by calculating the edges of the character, where sum of 'black' pixels is zero, along the periphery of the character [11]. Then, each character undergoes normalization in terms of size and focus, so as to resemble the 'templates' that have been used for training the ANN.

The ANN makes use of 'Feature Vectors' as its input. The features used are listed below:

- Normalized horizontal sum
- Normalized vertical sum
- Left spacing
- Right spacing
- Top Spacing
- Bottom spacing
- No: of zero-crossings
- Central to left spacing
- Central to right spacing

Then the model of ANN is created and trained. This represents the training phase. The details of ANN are provided in the upcoming sections. Once the network is trained, it is tested on a test set and the accuracy is determined. Now the NN can incorporate incoming input and recognize the character. This phase represents the classification. The classification phase has similar steps like preprocessing, feature extraction, classification using ANN etc.

ADAPTIVE CONTRAST BASED BINARIZATION

Proposed method is adaptive image contrast based method which combines local image contrast and local image gradient adaptively [12], [2]. The method is tolerant to text and background variation caused by different types of document degradations. This technique is capable addressing various noises that crawl in. Advantages of proposed method are: they are simple and robust, capable of handling degraded images, minimum parameter tuning, and ability to deal with sudden lighting/intensity changes, absence of over-normalization problem [2].

The major steps of the proposed methods are as given below:

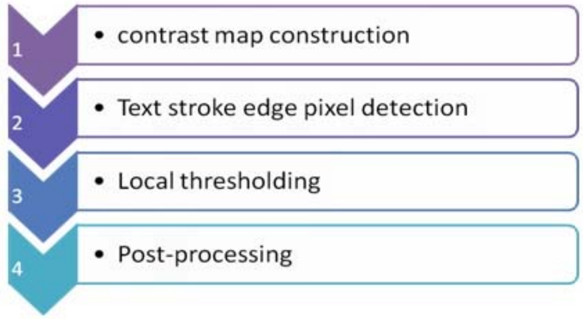
1
• contrast map construction
2
• Text stroke edge pixel detection
3
• Local thresholding
4
• Post-processing

Figure-2. Major steps in binarization.

- Construct the adaptive local image contrast map using statistical parameters of the image as given in the equation below.

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

Where,

$$C(i,j) = [I_{max}(i,j) - I_{min}(i,j)] \,/\, [\, I_{max}(i,j) + I_{min}(i,j) + \varepsilon\,] \tag{2}$$

$$\alpha = (STD/128)^{\gamma} \tag{3}$$

Here $I_{max}(i,j)$ and $I_{min}(i,j)$ are maximum and minimum intensities in a local window and Std is the standard deviation of the image. γ is a parameter which can be can be selected from [0,∞].

- Find the text stroke edge pixels
    - Binarize the contrast map by Otsu thresholding
    - Find Canny edge map of the image
    - Combine both the above steps by performing logical AND operation of their results.
- Text pixels are extracted by local threshold estimation
- Post-processing

OPTICAL CHARACTER RECOGNITION

Proposed method for OCR is Artificial Neural Network which partially imitates human thought process [13]. The ANN takes in and identifies a character based on its topological features such as form, evenness, open or closed areas etc. It gets inputs in the form of feature vectors. Every feature is assigned a value. A strong database is utilized to train the network. This helps to effectively recognize the character, by making use of its structural properties. Benefits of this method are attributed to their humanoid qualities such as adapting to changes and learning from previous experience. ANN is very much adaptable, can be thought, they can learn themselves, they are powerful [14].

Multi-layer perceptrons (MLP) are the most commonly used type of neural networks. MLP includes the input layer, output layer, and hidden layers. The hidden layers can be of any number. Each layer of MLP has one or more neurons directionally connected to the neurons from the previous and the next layer. The example below depicted in Figure-3 represents a 3-layer perceptron which have three input nodes, two output nodes, and the hidden layer having five neurons.

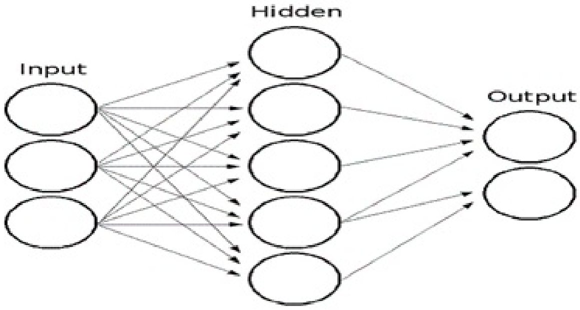

Figure-3. ANN.

All the neurons in MLP are similar. Each of them has several input links (it takes the output values from several neurons in the previous layer as input) and several output links. The values obtained from the previous layer are summed up with certain weights and added with the bias term. The sum is transformed using the activation function f that may be also different for different neurons as shown in Figure-4.

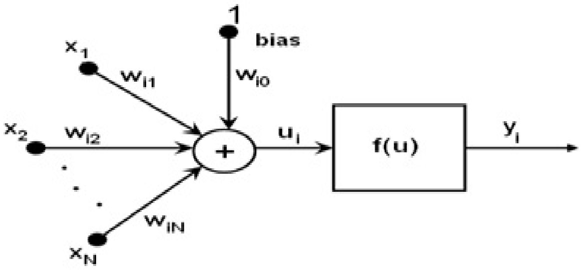

Figure-4. Neuron model.

In other words, given the outputs x_j of the layer n, the outputs y_i of the layer n+1 are computed as:

$$u_i=\sum(W_{i,j}{}^{n+1} *x_j) + W_{i,bias}{}^{n+1} \tag{4}$$

$$y_i=f(u_i) \tag{5}$$

The defined structure of the neural network is a 3 layer neural network. (253->16->26).It has 253 input nodes which is the size of feature vector, 16 nodes in the hidden layer and 26 nodes in the output layer. The optimum number of nodes in hidden layer is determined as 16 by trial and error method. The activation function used is symmetrical sigmoid. The training algorithm is back propagation.

IMPLEMENTATION

The system is implemented on the popular minicomputer board which comes in the size of a credit card-Raspberry Pi. It is a low cost board that runs on Raspbian OS. The latest model of Pi, Model B+ is chosen. Software platform chosen is C++ with OpenCV library.

Figure-5. Raspberry Pi board Image courtesy: www.raspberrypi.org.

RESULTS AND ANALYSIS

The results of binarization and OCR done on Raspberry pi are given below.

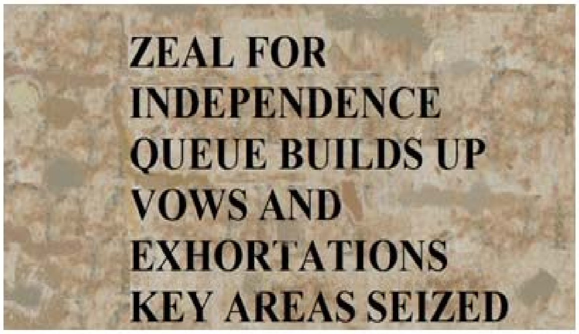

Figure-6. Input image.

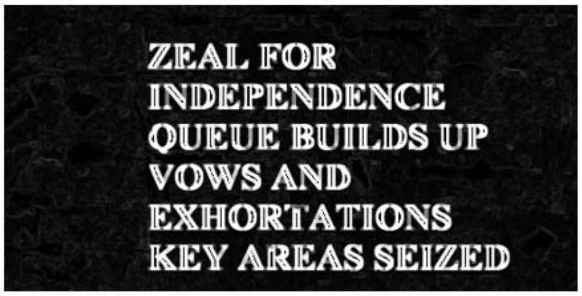

Figure-7. Contrast map.

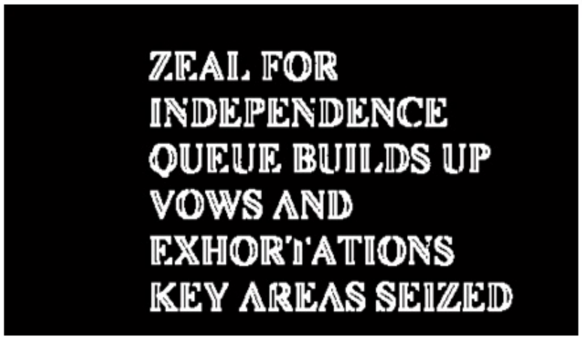

Figure-8. Binarized contrast map after applying Otsu thresholding.

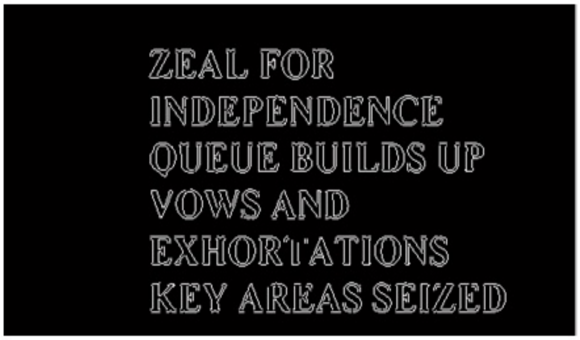

Figure-9. Canny edge map.

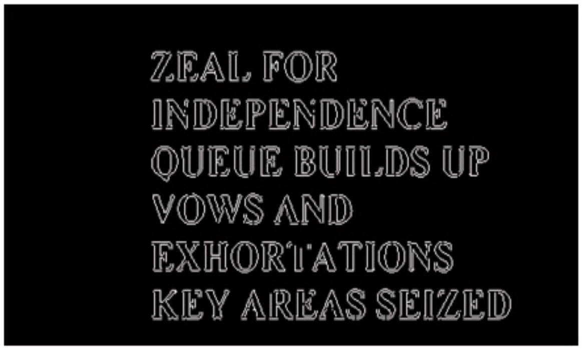

Figure-10. Combination (logical AND) of 8 and 9.

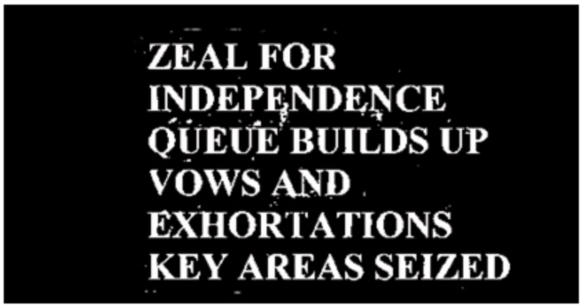
ZEAL FOR
INDEPENDENCE
QUEUE BUILDS UP
VOWS AND
EXHORTATIONS
KEY AREAS SEIZED

Figure-11. Binarized image (based on Canny detector).

The Analysis of the various edge detectors based on their PSNR (dB) and correlation is given below. From the above table, it was analyzed that detectors like Sobel and Perwitt performed better than Canny. This analysis used a non-degraded version of the input image as reference image which is of the same dimension as the input. The result using Sobel detector is given below.

Table-1. Analysis of various edge detectors.

Edge detector	Correlation value	PSNR (dB)
Canny	0.8505	17.95
Sobel	0.9136	19.82
Robert's	0.8428	17.85
Prewitt	0.9135	18.63
Laplacian	0.5070	15.1

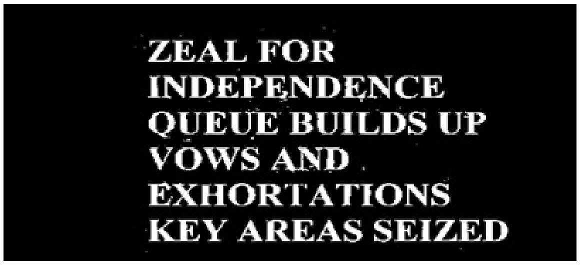
ZEAL FOR
INDEPENDENCE
QUEUE BUILDS UP
VOWS AND
EXHORTATIONS
KEY AREAS SEIZED

Figure-12. Binarization (based on Sobel detector).

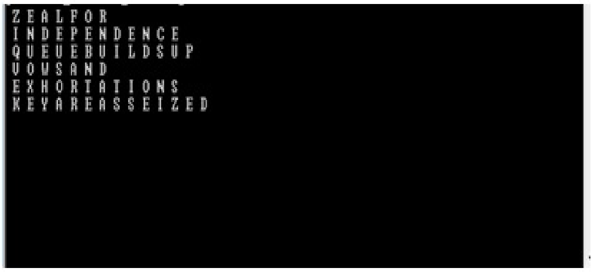
```
ZEALFOR
INDEPENDENCE
QUEUEBUILDSUP
UOUSAND
EXHORTATIONS
KEYAREASSEIZED
```

Figure-13. OCR results.

CONCLUSION AND FUTUTRE PLANS

The Document image processing is a very important area with wide applications. Binarization followed by OCR is performed to produce computer readable text. By using edge detectors like Sobel, Prewitt etc. in place of Canny edge detector the PSNR is improved thus providing better binarization. The OCR was implemented using artificial neural network. Since the pixel arrays of character couldn't give satisfactory results, some more features were extracted and trained. The final implementation was done on Raspberry Pi minicomputer board using C++ with OpenCV library and results are obtained. The ANN showed greater than 90% accuracy and gave accurate results for all the fonts trained. Raspberry Pi offered a low cost stand alone solution for binarization and OCR.

The system has its advantages such as less time complexity, very small database and high adaptability to untrained inputs using only a small number of features. Still, the system has a large scope for further developments.

The system can be made into a real time system by incorporating a camera or scanner. However, the Raspberry Pi needs to be optimized to provide faster results. Another suggestion for improvement is to have a robotic/mechanical system which can automatically scan through the lines of pages and perform OCR. Then the recognized text can be converted to speech to enable 'read aloud' mode which will be very useful for blind people.

REFERENCES

[1] Walker, Richard (2009-08-07), "Achieving The Paperless Office", Efficient Technology Inc http://docs.opencv.org/2.4.9/modules/ml/doc/neural_networks.html

[2] Bolan Su, Shijian Lu and Chew Lim Tan. 2013 " Robust Document Image Binarization Technique for Degraded Document Images", IEEE Transactions On Image Processing, Vol. 22, No. 4.

[3] Haidar Almohri, John S. Gray Hisham Alnajjar. 2008. "A Real-time DSP-Based Optical Character Recognition System for Isolated Arabic characters using the TI TMS320C6416T", Proceedings of The 2008 IAJC-IJME International Conference ISBN 978-1-60643-379-9.

[4] Salvador Espan˜a-Boquera, Maria Jose Castro-Bleda,Jorge Gorbe-Moya, and Francisco Zamora-Martinez. 2011. " Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models ", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 33, No. 4.

[5] Hossein Ziaei Nafchi, Reza Farrahi Moghaddam and Mohamed Cheriet. 2014. "Phase-Based Binarization of Ancient Document Images: Model and

Applications", IEEE Transactions On Image Processing, Vol. 23, No. 7.

[6] Yakobov V., Mash L., Thirer N. 2010. "On implementation of an OCR algorithm for musical notation", IEEE26th convention of Electrical and Electronics Engineers in Israel.

[7] J. Bernsen. 2004. "Dynamic thresholding of gray-level images," in Proc. Int. Conf. Pattern Recognit., Oct, pp. 1251–1255.

[8] B. Su, S. Lu and C. L. Tan. 2010. "Binarization of historical handwritten document images using local maximum and minimum filter," in Proc. Int. Workshop Document Anal. Syst., Jun. pp. 159–166.

[9] Christopher J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Kluwer Academic Publishers, Boston.

[10] Raghuraj Singh, C. S. Yadav, Prabhat Verma and Vibhash Yadav. 2010. "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network", International Journal of Computer Science & Communication, Vol 1.

[11] Jagroop Kaur and Dr.Rajiv Mahajan. 2014. "A Review of Degraded Document Image Binarization Techniques", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, No. 5.

[12] B. Gatos, I. Pratikakis and S. Perantonis. 2006. "Adaptive degraded document image binarization," Pattern Recognit., Vol. 39, No. 3, pp. 317–327.

[13] Vinod Chandra and R. Sudhakar. 2000. "Recent Developments in Artificial Neural Network Based Character Recognition: A Performance Study", IEEE.

[14] Vivek Shrivastava and Navdeep Sharma. 2012. "Artificial Neural Network Based Optical Character Recognition", Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.5.