View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by International Institute for Science, Technology and Education (IISTE): E-Journals

Computer Engineering and Intelligent Systems www.iiste.org
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol 2, No.4, 2011

Document Image Binarization Using Post Processing Method

E. Balamurugan
Department of Computer Applications
Bannari Amman Institute of Technology
Sathyamangalam, Tamilnadu, India
E-mail: rethinbs@gmail.com

K. Sangeetha
Department of Computer Applications
Bannari Amman Institute of Technology
Sathyamangalam, Tamilnadu, India
E-mail: kavigeeth@yahoo.com

Dr. P. Sengottuvelan
Department of Information Technology
Bannari Amman Institute of Technology
Sathyamangalam, Tamilnadu,, India
E-mail: sengottuvelan@rediff.com

Abstract

Binarization is the preliminary process of Document Image Analysis and Processing. Image binarization is performed through Local and Global threshold methods. In this paper local thresholding method Nilblack method with post processing was implemented. The Nilblack algorithm was implemented using Matlab and tested with a sample of ground tooth images selected from TOBACCO research database.

Keywords: Document Image, Nilblack method, Post Processing

1. Introduction

Image binarization is the process of turning a greyscale image to a black and white image. In a grey-scale image, a pixel can take on 256 different intensity values while in binary image each pixel is assigned to be either black or white. This conversion from grey-scale to black and white is performed by applying a threshold value to the image. Representation of documents in binary form is essential for further processing like finding texts, lines, graphics, logos etc. It is the primary step of document image analysis and processing research. Binarization is performed through global or local threshold values. In global approach single threshold value is applied for entire image while local thresholding apply different threshold values to different regions of images. A critical component in the binarization process is choosing a correct value for the threshold. If the threshold is set too low, then the resulting binary image will primarily be comprised of white pixels. Conversely, if the threshold is set too high, the resulting image will feature a large number of undesired black pixels. Thus, the threshold must be selected carefully to ensure the data information is preserved. Here, the section 2 has an over view of global and local thresholding methods and Tobacco research database. In section 3 about Nilblack method with post processing step. In section 4 Matlab implementation results for Nilblack method with post processing. Finally we have concluded with session 5.

2 .Local and Global Thresholding

In Global Thresholding method a single threshold value is calculated for the whole image. Pixels with a grey scale level under the threshold level are labelled as print, pixels with a grey scale level above the threshold level are labelled as background. Among many global thresholding methods, Otsu's method is one among the best method, based on an analysis of the grey scale level histogram of the whole image and selects an optimal threshold according to the discriminant theory. The computation of Otsu's method is fast and simple and efficient for large images. But, for images with uneven illuminations, the method cannot separate objects from their backgrounds effectively. In Local thresholding to calculate the threshold for a window as being the mean value of the maximum and minimum values within the window. Another local method uses gradient. Pixels are identified in, or very close to, areas where sharp changes (edges) exist in the grey level image. The areas with sharp edges are then checked for evidence labelling them as either text or background. Tobacco800, composed of 1290 document images, is a realistic database for document image analysis and retrieval as these documents were collected and scanned using a wide variety of equipment over time. In addition, a significant percentage of Tobacco800 are consecutively numbered multi-page business documents, making it a valuable tested for various content-based document retrieval approaches. Resolutions of documents in Tobacco800 vary significantly from 150 to 300 DPI and the dimensions of images range from 1200 by 1600 to 2500 by 3200 pixels.

3. Nilblack Method

Nilblack's method is based on the calculation of the local mean and of the local standard deviation. The threshold T at pixel (x, y) is determined by the following equation:

$$T(x, y) = m(x, y) + k \cdot s(x, y)$$

where m(x, y) and s(x, y) are respectively the average of a local area and the standard deviation values. The size of the neighbourhood should be small enough to preserve local details, but at the same time large enough to suppress noise. The value of k is used to determine how much of the total print object boundary is taken as a part of the given object. This method can distinguish the object from the background effectively in the areas close to the objects. But it cannot suppress background noise situated far away from the objects. Thus, if the objects are sparse in an image, a lot of background noise will be left. Suggested using the grey-level values at high gradient regions as known data to interpolate the threshold surface of image document texture features. The post processing step a threshold surface is constructed by finding the edge points of the smoothed image. The gradient magnitude image is computed and thinned to one pixel-wide line to identify edge points. An iterative interpolation process is employed to get a smooth surface passing through the edge points. The image is threshold by the constructed surface. In an iterative interpolation process, the interpolated surface is set at image grey scale at the edge points and 0 at the other points. Then the interpolation residual R(x, y) and new pixel value Pn+1(x, y) at iteration (n + 1) are calculated. The iteration will stop when the residual R(x, y) is lower than a small number. The residual Rn(x, y) at the nth iteration and the new pixel value Pn+1(x, y) are defined as follows:

$$R(x, y) = Pn(x, y+1) + Pn(x, y-1) + Pn(x-1, y) + Pn(x+1, y) - 4Pn(x, y) \quad \text{--} \quad a$$

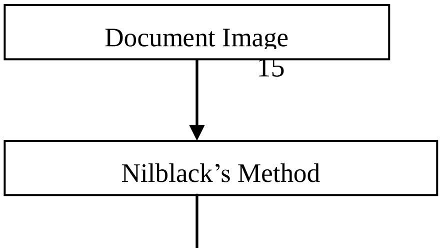
$$Pn+1(x, y) = Pn(x, y) + (\beta \cdot Rn(x, y)/4) \quad \text{--} \quad b$$

Where $1 < \beta < 2$ for fast convergence. Final false print objects are removed by the post processing step.

4. Implementation

Nilblack's method calculates the thresholding value using the local average and the local standard deviation but because of the limitation caused by the window size of [15, 15], the information outside the window is not taken into account. Therefore large areas of noise cannot be separated from their useful object information. Experiments show that Nilblack's method can separate the wide spaces from the sentences and remove the noise close to the words but cannot remove the noise far from the objects but if we are applying post processing to find the object boundary points for smoothing and averaging filter is applied to identify the edges available in the document image.

Process flow:



The binarization methods were tested on sample document images from Tobacco800 document images realistic database. We can see from the results that our implementation of fig 1. a, b,c retains most of the useful object information and suppresses the background noise effectively

5. Conclusion

In this paper, we have implemented nilblacks method with post processing step using matlab7.6 and tested random samples from the Tobacco800 database, for binarizing document images based on a criterion of iterative interpolation process. We have shown the result a sample, which is formulated in terms of contrast instead of grey values, is robust against noise variations. By testing the binarization system against various challenging document images, we have also demonstrated the effectiveness of nilblack method with post processing algorithm.

References

Nilblack, W (1986) *'An Introduction to Digital Image Processing'*, Englewood Cliffs, Prentice Hall,

pp 115-116.

Yanowitz, S D and Bruckstein, A M 1989 'A new method for image segmentation', *Computer Vision, Graphics and Image Processing*, vol. 46, no. 1, pp. 82-95.

Otsu, N 1979 'A Threshold Selection Method from Gray-Level Histograms' IEEE *Transactions on Systems,Man, and Cybernetic,* Vol.9(1) pp. 62-66.

Mehmet Sezgin and Bulent Sankur, 2004 'Survey over Image Thresholding Techniques and Quantitative Performance Evaluation', *Journal of Electronic Imaging*, Vol. 13(1) pp. 146-165.

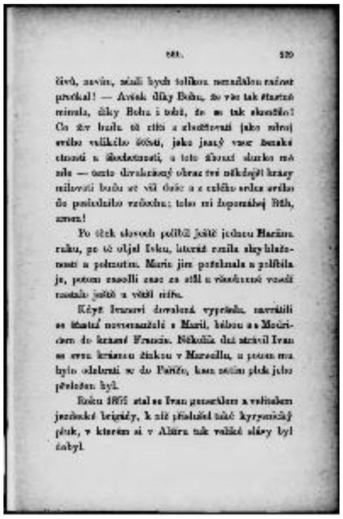
Lewis, D Agam, G Argamon, S Frieder,O Grossman, D and Heard, J 2006 'Building a test collection for complex document information processing', in Proc. Annual Int. ACM SIGIR Conference, pp.665 - C666.

Agam, G Argamon, S Frieder, O Grossman, D and Lewis, D 2006 'The Complex Document Image Processing(CDIP) test collection', Illinois Institute of Technology, http://ir.iit.edu/projects/CDIP.html

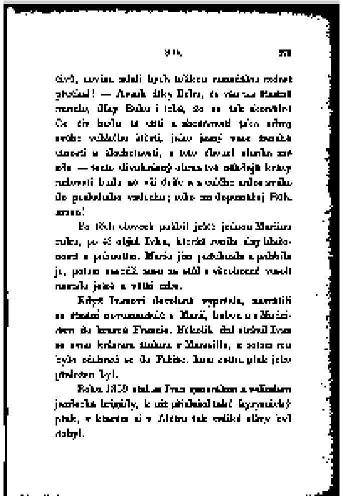
The Legacy Tobacco Document Library (LTDL) 2007, University of California, San Francisco, http://legacy.library.ucsf.edu/

Results:

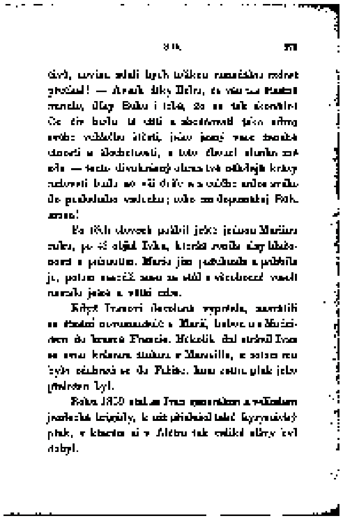
1.a Grey Scale Image



1.b. Niblacks Method



1.c. post processing step



This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage: http://www.iiste.org

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a fast manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar