AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

An Adaptive Hybrid Algorithm for Document Images Binarization subject to Complex Background

A Thesis submitted in partial fulfillment of the requirements of
Doctor of Philosophy in Electrical Engineering
(Computer and Systems Engineering)

by
Enas Mahmoud Mahmoud Mohamed Elgbbas
Master of Science in Electrical Engineering
(Computer and Systems Engineering)
Faculty of Engineering, Ain shams university, 2009

Supervised By
Prof. Hazem Mahmoud Abbas
Dr. Mahmoud Ibrahim Khalil

Cairo, 2019

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

An Adaptive Hybrid Algorithm for Document Images Binarization subject to Complex Background

by

Enas Mahmoud Mahmoud Mohamed Elgbbas

Master of Science in Electrical Engineering
(Computer and Systems Engineering)
Faculty of Engineering, Ain shams university, 2009

Examiners' Committee

Name and affiliation	**Signature**
Prof. Elsayed Eissa Abdo Hemayed Computer Engineering. Faculty of Engineering, Cairo University.	.....
Prof. Mohamed Watheq El-Kharashi Computer and Systems Engineering Faculty of Engineering, Ain shams University.	.....
Dr. Mahmoud Ibrahim Khalil Computer and Systems Engineering Faculty of Engineering, Ain shams University.	.....

Date:15/10/2019

Statement

This thesis is submitted as a partial fulfillment of Doctor of Philosophy in Electrical Engineering, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Enas Mahmoud

Enas Mahmoud

.....

Date:

Researcher Data

Name: Enas Mahmoud Mahmoud Mohamed Elgbbas
Date of Birth: 12/10/1980
Place of Birth: Cairo, Egypt
Last academic degree: Master of Science in Electrical Engineering
Field of specialization: Image processing
University issued the degree : Ain shams university
Date of issued degree : 2009

Acknowledgment

I would like to express my deep gratitude to Professor Hazem Mahmoud Abbas for his patient guidance, enthusiastic encouragement and useful critiques of this research work. I would like to express my very great appreciation to Dr. Mahmoud Ibrahim Khalil for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated.

Thesis Summary

In this thesis, we introduce two adaptive and hybrid binarization methods. The first method is proposed for solving all types of degradation, such as non-uniform background, faint text, low contrast, stain, bleed-through, and shadow. This method depends on Otsu multilevel thresholding method, the estimated stock width, and the image contrast. The second proposed method aims to binarize the document images contain normal illumination using an improved version of the spectral clustering algorithm in speed and complexity. The thesis is divided into six chapters as listed below:

Chapter 1

In this chapter an introduction of image document binarization is presented. It also introduces the objective and the thesis outline.

Chapter 2

A survey for binarization methods proposed in the literature is presented in this chapter.

Chapter 3

The method based on the Otsu multilevel for solving all types of degradation is proposed in chapter.

Chapter 4

In this chapter, the method based on spectral clustering for normally illuminated document images is proposed.

Chapter 5

The experimental results of the method based on Otsu multilevel and method based on spectral clustering are presented.

Chapter 6

The conclusions and future work are presented in this chapter.

Key words: Document images binarization, historical document images, image degradation, non-uniform background, uniform background, faint text, low contrast, thin stock width, stain, bleed-through, and shadow, Otsu multilevel algorithm, spectral clustering algorithm, adaptive algorithm, hybrid algorithm.

Abstract

In this thesis, we propose two adaptive and hybrid methods for binarization of historical document images. The first method uses Otsu multilevel algorithm, and depends on the image contrast and the estimated stroke width. This method is proposed for binarization of historical document images suffering from various types of degradation, such as non-uniform background, faint text, low contrast, stain, bleed-through, and shadow. Solving all these problems effectively is a challenge. Focus on noise elimination may cause loss of faint text. On the contrary, faint or low contrast text extraction may produce noisy images. Therefore, we classified the investigated images into two groups using a suggested factor. The first group includes uniform background images that may contain faint text or shadow, while the second group includes non-uniform background images that may suffer from stain, bleed-through, shadow or faint text. To extract this factor, the background of the investigated image is initially estimated, then global Otsu multilevel is applied and dividing it into three regions. The difference between the average intensities of the darkest and brightest regions is an indicator of the image class. For each group, an adaptive and hybrid binarization technique is suggested. For the first group, global Otsu is applied to the grayscale image and the stroke width is estimated. Areas that are more likely still contain missing text are identified adaptively

and binarized separately using a pseudo-local version of Otsu multilevel method, and lost text recovered based on the stroke width. Faint text, shadow or background noise are distinguished based on the image contrast. The clarity of text is increased by using a dynamic window size for local binarization (Niblack method is used for thin pen stroke text and Otsu otherwise). For the second group, non-uniform background and most of stain and bleeding-through are removed by normalization, then global Otsu is applied and the stroke width is estimated. Text lost during normalization is restored. The remaining stain and bleed-through objects are detected depending on estimated stroke width, then they are locally binarized. Finally, a post-processing step based on the estimated stroke width is applied to remove the shadow. The proposed method is evaluated using seven databases DIBCO09, H-DIBCO'10, DIBCO'11, H-DIBCO'12, DIBCO'13, H-DIBCO'14, and H-DIBCO'16. The average F-measure for each database 90.7%, 89.1%, 88.9%, 88.7%, 88.9%, 93.3%, and 89.4% respectively.

The second proposed method is suitable for normal illuminated document images that incorporate the advantages of Otsu and spectral clustering algorithm. To overcome the noise problem, a preprocessing step is applied to the document image. After that, the resulted image is binarized using Otsu producing a binary image. As a final step, the spectral clustering algorithm is locally applied on the original image, with the aid of the binary image, to retrieve faint text. The proposed design of spectral clustering provides a significant reduction of the similarity matrix computing time and size used, without affecting the quality of clustering. The proposed method is evaluated using the uniform background images taken from DIBCO09, H-DIBCO'10, DIBCO'11, H-DIBCO'12, DIBCO'13, H-DIBCO'14.