

بِحَمْدِهِ تَعَالَى



DOCUMENT PREPROCESSING MODULE FOR OPTICAL CHARACTER RECOGNITION

By

Shaimaa Samir Abu-Elela Mohamed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND COMMUNICATION ENGINEERING

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT

DOCUMENT PREPROCESSING MODULE FOR OPTICAL CHARACTER RECOGNITION

By

Shaimaa Samir Abu-Elela Mohamed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND COMMUNICATION ENGINEERING

Under the Supervision of

Prof. Mohsen Abdelrazik Ali Rashwan

Professor

Electronics and Communication Engineering Department

Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT

DOCUMENT PREPROCESSING MODULE FOR OPTICAL CHARACTER RECOGNITION

By

Shaimaa Samir Abu-Elela Mohamed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS AND COMMUNICATION ENGINEERING

Approved by the Examining Committee:

Prof. Mohsen Abdelrazik Ali Rashwan, Thesis Main Advisor

Prof. Omar Ahmed Nasr, Internal Examiner

Prof. Sherif Mahdi Abdou, External Examiner

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT

Engineer's Name: Shaimaa Samir Abu-Elela Mohamed
Date of Birth: 18/1/1990
Nationality: Egyptian
E-mail: eng.shaimaa.samir@gmail.com
Phone: 01220429316
Address: Postal Address
Registration Date: 20/9/2012
Awarding Date: //2017
Degree: Master of Science
Department: Electronics and Communication Engineering



Supervisors:

Prof. Mohsen Abdelrazik Ali Rashwan

Examiners:

Prof. Sherif Mahdi Abdou

(External examiner)

Prof. Omar Ahmed Nasr

(Internal examiner)

Prof. Mohsen Abdelrazik Ali Rashwan

(Thesis main advisor)

Title of Thesis:

Document preprocessing module for optical character recognition

Key Words:

Line segmentation; Image enhancement; Sparse Modeling; Image Preprocessing; Optical Character Recognition.

Summary:

Optical Character Recognition (OCR) is the process of converting document images into editable text. This enables us to edit and search documents very easily. Document preprocessing of a document image is a very important stage prior to OCR. This thesis presents new line segmentation, image enhancement, and character reconstruction algorithms to improve document preprocessing. Consequently, we can get better results from the OCR system.

Acknowledgements

Praise be to God for his kindness and mercy. Which without his will I would not have done, or completed this work. Many words of thanks and appreciation to my thesis supervisor Dr.Mohsen for his guidance, precious advices, and the trust he put in me to do the work in this thesis. Many and great thanks for my family, their support and help always push me forward. Special thanks to my dear friend Wafaa, for her concern and help in writing this thesis. Finally I would like to thank all my friends, and my colleges at the Engineering Company for the Development of Computer Systems (RDI).

Dedication

For all those who believe in me more than I believe in myself, my family and my friends.

Table of Contents

Acknowledgements	i
Dedication	ii
Table of Contents	iii
List of Tables	vi
List of Figures	vii
Abstract	x
1	
INTRODUCTION	1
1.1 Image preprocessing applications	1
1.2 Image preprocessing for OCR applications	1
1.3 Challenges of preprocessing	1
1.4 Historical documents challenges	2
1.5 Arabic language challenges	2
1.6 Thesis objectives	2
1.7 Thesis contributions	2
1.8 Organization of the thesis	3
2	
LITERATURE REVIEW	4
2.1 Document denoising and background enhancement	4
2.1.1 Filter based algorithms	5
2.1.1.1 Natural Images	5
2.1.1.2 Document images	5
2.1.2 Dictionary based algorithms	5
2.1.2.1 Natural Images	6
2.1.2.2 Document images	6
2.1.3 Transform based algorithms	6
2.1.3.1 Natural Images	6
2.1.3.2 Document images	7
2.1.4 Binarization and segmentation algorithms	9
2.1.4.1 Document images	9
2.1.5 Deep learning algorithms	9
2.1.6 Other denoising algorithms	9
2.2 Image binarization	10

2.3	Skew detection and correction	11
2.4	Frame detection and line removal	11
2.5	Layout analysis	11
	2.5.0.1 Top-down	11
	2.5.0.2 Bottom-up	12
2.6	Line Segmentation	12
2.7	Character reconstruction	14
	2.7.1 Active contours algorithms	16
	2.7.2 Rule based algorithms	16
	2.7.3 In-painting algorithms	16
	2.7.4 Filter based algorithms	17
	2.7.5 Segmentation and binarization algorithms	17
	2.7.6 Build a robust recognition system	18

3

SYSTEM ARCHITECTURE 19

3.1	Denoising	19
	3.1.1 Image denoising using sparse redundant representations	19
	3.1.2 Image denoising using neural networks (Denoising auto-encoders)	21
3.2	Orientation detection and correction	21
3.3	Binarization	23
3.4	Frames and line removal	23
3.5	Line segmentation	25
	3.5.1 Line segmentation using projection profiles	25
	3.5.2 Segmentation using projection profile, and connected components analysis	30
3.6	Character reconstruction	32
	3.6.1 Character reconstruction using sparse redundant representation	37
	3.6.2 Character reconstruction using neural networks (Denoising auto-encoders)	37

4

SPARSE MODELING 40

4.1	Sparse modeling	40
	4.1.1 Image denoising using sparse redundant representation	41
	4.1.2 Image in-painting using sparse redundant representation	41
4.2	Sparse auto-encoders	45
	4.2.1 Neural networks	45
	4.2.2 Auto-encoders	45
	4.2.2.1 Denoising auto-encoders	46
	4.2.2.2 Stacked denoising auto-encoders	47
	4.2.2.3 Sparse auto-encoders	47
	4.2.3 Image denoising using auto-encoders	48

5

EXPERIMENTAL RESULTS AND EVALUATION	51
5.1 Denoising experiments	51
5.1.1 Computer generated data experiments	51
5.1.1.1 Noisy data generation	51
5.1.1.2 Sparse redundant representation experiments	53
5.1.1.3 Denoising auto-encoder experiments	53
5.1.1.4 State of the art denoising experiments	59
5.1.2 Real data Experiments	64
5.1.2.1 Historical manuscripts experiment	64
5.1.2.2 Ordinary pages experiment	64
5.2 Line segmentation experiments and results	71
5.2.1 Line level accuracy	72
5.2.2 Recognition results	73
5.3 Character reconstruction experiments	73
5.3.1 Computer generated data experiments	74
5.3.1.1 Broken characters computer generated data	74
5.3.1.2 Character reconstruction using sparse redundant representations	74
5.3.1.3 Character reconstruction using denoising auto-encoders	80
5.3.2 Real data experiment	82
5.3.2.1 Historical manuscripts experiments	82
5.3.2.2 Ordinary pages experiment	82

6

CONCLUSION AND FUTURE WORK	85
6.1 Conclusions	85
6.2 Future work	86

References	87
-------------------	-----------

List of Tables

5.1	Average recognition results for sparse redundant representations denoising.	53
5.2	Recognition results at different patch size.	58
5.3	Effect of the layer size, the network sparsity, and the neuron activation function on the recognition result	58
5.4	Recognition results for different number of layers.	58
5.5	Average recognition results for auto-encoder denoising.	59
5.6	Recognition results for state of the art denoising algorithms.	60
5.7	Mean square error for state of the art denoising algorithms versus alpha.	63
5.8	Mean square error for BM3D, and the proposed algorithms.	63
5.9	Signal to noise ratio for state of the art denoising algorithms versus alpha.	63
5.10	Signal to noise ratio for BM3D, and the proposed algorithms.	63
5.11	Recognition accuracy for the denoising of historical pages	64
5.12	Recognition results for 100 ordinary pages	71
5.13	Line segmentation results	72
5.14	Average recognition results of RDI engine	73
5.15	Recognition accuracy of sparse coding character reconstruction algorithm	80
5.16	Recognition results at different patch size.	80
5.17	Effect of layer size, network sparsity, and neuron activation function on recognition result	80
5.18	Recognition accuracy of denoising auto-encoders character reconstruction algorithm	81
5.19	Recognition results for historical documents.	82
5.20	Recognition results for 100 ordinary pages	84

List of Figures

2.1	Sparse coding document denoising	6
2.2	Example of wavelets diffusion denoising	7
2.3	Example of wavelets diffusion denoising	8
2.4	Curve-lets document denoising	8
2.5	Projection profile example	13
2.6	Attractive repulsive method example	14
2.7	Seam carving binarization free line segmentation example	15
2.8	Example for the character reconstruction algorithm output	17
2.9	Example for character reconstruction algorithm using PDE	18
3.1	Example of the dictionary used in the denoising process	20
3.2	Example of encoder weights of an auto-encoder trained on Arabic text images	22
3.3	Hough transform	23
3.4	Savoula binarization	24
3.5	Frame removal example	26
3.6	overlapped character example	27
3.7	Projection profile example	28
3.8	Projection profile example	29
3.9	Components classification example	31
3.10	Separated line extraction example	33
3.11	Assign component to line example	34
3.12	Final segmentation result example	35
3.13	Final segmentation of the whole page	36
3.14	Example of the dictionary used in character reconstruction	38
3.15	Example of the weights learned by an auto-encoder trained on Arabic text images	39
4.1	Example for denoising of natural images using sparse representation	43
4.2	Example for learned global dictionary	43
4.3	Example for in-painting of natural images using sparse representation	44
4.4	Traditional auto-encoder	46
4.5	Denoising auto-encoder	47
4.6	Stacked denoising auto-encoder	47
4.7	Comparison between image denoising using sparse auto encoders and other denoising algorithms	49
4.8	Comparison between image in-painting using sparse auto encoders and other KSVD algorithm	50
4.9	Learned weights for denoising auto-encoders at different noise levels, trained at natural images	50
5.1	Example for noisy images at different degradation levels	52

5.2	Recognition accuracy for sparse representation algorithm	54
5.3	Example for the denoised algorithms results at low degradation levels	55
5.4	Example for documents before and after applying the denoising algorithm	56
5.5	Dictionary used in sparse denoising	57
5.6	Stacked denoising auto-encoders (Deep feed-forward neural network)	59
5.7	Denoising auto-encoder	59
5.8	Denoising auto-encoder learned weights	60
5.9	Recognition accuracy for denoising auto-encoder algorithm	61
5.10	Example for document before and after applying the denoising algorithm	62
5.11	Recognition accuracy for Median filtering algorithm	65
5.12	Recognition accuracy for Weiner filtering algorithm	66
5.13	Recognition accuracy for BLS-GSM algorithm	67
5.14	Recognition accuracy for phase binarization algorithm	68
5.15	Recognition accuracy for BM3D algorithm	69
5.16	Example 1 for the denoising of historical documents	70
5.17	Example 2 for the denoising of historical documents	70
5.18	Example 3 for the denoising of historical documents	70
5.19	Example of documents used in line segmentation	71
5.20	Example for manually segmented page	72
5.21	Example for manually segmented page	73
5.22	Example for a line resulted from different segmentation algorithms	74
5.23	Example for a line resulted from different segmentation algorithms	75
5.24	Example for a line resulted from different segmentation algorithms	76
5.25	Example for a line resulted from different segmentation algorithms	76
5.26	Example for a line resulted from different segmentation algorithms	77
5.27	Different types of masks that are used to induce broken characters	77
5.28	Different types of image degradation	78
5.29	Example of character reconstruction algorithm	79
5.30	Example for document before and after applying the algorithm	81
5.31	Structure for auto-encoder used in character reconstruction	82
5.32	Example 1 for character reconstruction of historical documents	83
5.33	Example 2 for character reconstruction of historical documents	83
5.34	Example 3 for character reconstruction of historical documents	84

List of Symbols and Abbreviations

BLS-GSM : Bayes Least Squares Gaussian Scale Mixture

BM3D : Block-matching and 3D filtering

DCT : Discrete Cosine Transform

DA : Denoising Auto-encoder

GVF : Gradient Vector Flow

HMM : Hidden Markov Model

K-SVD : An Algorithm for Designing Over-complete dictionaries for Sparse Representation

KNN : K nearest neighbors

LASSO : Least Absolute Shrinkage and Selection Operator

MNIST : National Institute of Standards and Technology

MLP : Multilayer Perceptron

NP : Nondeterministic Polynomial Hard

OCR : Optical Character Recognition

QCRI : Qatar Computing Research Institute

SVD : Singular Value Decomposition

SDA : Stacked Denoising Auto-encoder

SSDA : Sparse Stacked Denoising Auto-encoder

3D : Three dimensional

Abstract

The process of converting a document image into an editable text has become a very spread process nowadays because, at the era of technology, everything is handled by computers. The process of converting a scanned image into an editable text is called Optical Character Recognition (OCR). This process consists of a lot of stages, and document preprocessing is one of those stages. This stage is very essential in any OCR system because the scanned documents are usually not ideal. The preprocessing stage is responsible for preparing the document to be recognized by a recognition system.

The preprocessing stage is concerned with everything before the character recognition stage. The tasks, which are performed in the document preprocessing stage, are listed as follows. Denoising cleans the document from any unwanted objects, that are not part of the original document. Background enhancement algorithms are sometimes applied, if needed, to enhance the background of the document. Binarization converts the document into black and white in order to facilitate feature extraction process at the recognition stage. Document layout analysis extracts document blocks and classifies them into text and image regions. Some documents are rotated during scanning, their rotation angle must be detected and fixed, which is called document deskewing. Afterwards, segmentation divides text blocks into smaller regions (lines, words, or character). Finally, some documents may contain broken characters, which need to be fixed through character reconstruction.

This thesis proposes new techniques in document preprocessing for Arabic OCR systems. The proposed techniques can handle both early printed manuscripts and ordinary Arabic documents. Moreover, those techniques are general and can handle any other language. The thesis focuses on the preprocessing functions that may cause great enhancement in the recognition accuracy.

Denoising and background enhancement can greatly improve the results of an OCR system. In the field of natural images, denoising of an image using its sparse representation over a learned dictionary shows the state of the art results. Denoising auto-encoders show good results too in natural images, but there are only few experiments on these algorithms in document images. The work in this thesis uses the same algorithms in denoising of documents and background enhancement, but with one of the most complicated noise types, which is the show through noise. The show through noise results from double sided pages when one of those pages is transparent. Experiments on this part show great improvement in the recognition of a highly degraded document. The enhancement in recognition accuracy ranges from 13% to 26% calculated using three different OCR engines.

The segmentation of blocks of text into lines is also an essential stage. Segmentation using projection profile is a very common method. The results show that projection profile line segmentation, with some modifications, is able to segment text lines of early printed Arabic manuscripts. Moreover, the proposed algorithm shows its superiority in the recognition accuracy in comparison with the other algorithms. The algorithm can reach 68.5% as the OCR recognition accuracy, while on line level accuracy has a precision equals to 95.18

The problem of broken characters is also one of the critical aspects. A document that contains such bad characters may not be recognized properly at all. The algorithms used in denoising

can be used in character reconstruction as well. Applying the algorithm and comparing the recognition results of the three different engines, the recognition accuracy enhancement ranges from 4.2% to 26.5%.