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

Insert photo here

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

List of Tables

List of Figures

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%.