

RECTIFICATION OF IMAGES FOR CAMERA BASED OCR SYSTEMS

By

Mohamed Fawzi Abdalrasoul Afifi

A Thesis submitted to the

Faculty of Engineering at Cairo University

In partial Fulfillment of the

Requirement for the degree of

MASTER OF SCIENCE

In

ELECTRONICS AND COMMUNICATIONS ENGINEERING

FACULTY OF ENGINEERING, CAIRO UNIVERSITY

GIZA, EGYPT

2015

RECTIFICATION OF IMAGES FOR CAMERA BASED OCR SYSTEMS

By

Mohamed Fawzi Abdalrasoul Afifi

A Thesis submitted to the

Faculty of Engineering at Cairo University

In partial Fulfillment of the

Requirement for the degree of

MASTER OF SCIENCE

In

ELECTRONICS AND COMMUNICATIONS ENGINEERING

Under the supervision of

Prof. Dr. Mohsen Abdelrazik Rashwan

Professor of Electronics and Communications Engineering

Electronics and Communication Department

Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY

GIZA, EGYPT

2015

RECTIFICATION OF IMAGES FOR CAMERA BASED OCR SYSTEMS

By

Mohamed Fawzi Abdelrasoul Afifi

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

Approved by the
Examining committee

Prof. Dr. Mohsen Abdelrazik Rashwan, Thesis Main Advisor

Prof. Dr. Samia Mashaly, External Examiner
Electronics Research Institute

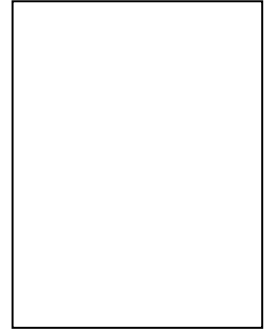
Prof. Dr. Sherif Mahdy Abdou, External Examiner
Faculty of Computers and Information, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY

GIZA, EGYPT

2015

Engineer: Mohamed Fawzi Abdelrasoul
Date of Birth : 24/ 6 / 1987
Nationality : Egypt
E-mail : m.fawzi1987@gmail.com
Phone. : +201227789886
Address: kobba-Cairo
Registration Date : 1/10 /2009
Awarding Date : / /
Degree: Master of Science
Department: ELECTRONICS AND COMMUNICATIONS
ENGINEERING



Supervisor: Prof. Dr. Mohsen Abdelrazik Rashwan

Examiners: Prof. Dr. Mohsen Abdelrazik Rashwan
Prof. Dr. Samia Mashaly
Prof. Dr. Sherif Mahdy Abdou

Title of Thesis: RECTIFICATION OF IMAGES FOR CAMERA BASED OCR SYSTEMS

Key Words: Camera – OCR – Distortion – Scanner – Arabic language

Summary:

Optical character recognition (OCR) works with scanners to convert printed characters into digital text. Current advance in camera's lens industry introduced the cameras as an alternative to scanner for document image capturing. Compared to scanners, cameras are considered to be more convenient way to capture document images. However, document images captured by camera suffer from problems such non uniform illumination, resolution and distortion which affect the OCR performance. We introduce new techniques to overcome the distortion combined with capturing Arabic document image by camera.

Acknowledgements

I would like to gratefully acknowledge the enthusiastic supervision and guidance of Prof. DR. Mohsen Rashwan during this work. His kind inspiration, invaluable guidance and encouragement from time to time had kept me going in spite of the adversities confronted and have culminated in the successful completion of this thesis.

I would also like to thank all of my family for their unfailing support and love all through. Their help can never be penned in words.

I am grateful to my dear friend Dr. Ahmed Elmoslimany for his great support during all this thesis phases. He as a good friend was always willing to help and give his best suggestions.

I am also grateful to all MOD-CDM elite team for their cooperation and friendly moral support throughout. I will always cherish the good times we have had. Thanks to my dear friends Heba Shalaby, Mohamed Sayed, Ahmed Hussien and Mehased Ahmed for their endless support and kind help. My research would not have been possible without their helps.

Above all, I thank Allah for bringing me this far.

Mohamed Fawzi...

CONTENTS

LIST OF FIGURES.....	IVIII
LIST OF TABLES.....	XI
LIST OF ABBREVIATION	XII
ABSTRACT.....	XII
CHAPTER 1.....	1
INTRODUCTION.....	1
1.1 CAMERA CAPTURED DOCUMENT PREPROCESSING CHALLENGES	2
1.2 THESIS OBJECTIVES.....	6
1.3 THESIS CONTRIBUTION.....	8
1.4 THESIS OUTLINE.....	10
CHAPTER 2.....	11
RELATED WORK.....	11
2.1 SKEW DETECTION AND CORRECTION.....	11
2.2 PERSPECTIVE DISTORTION DETECTION AND CORRECTION.....	12
2.3 GEOMETRIC DISTORTION DETECTION AND CORRECTION.....	14
CHAPTER 3.....	16
SKREW AND PERSPECTIVE DISTORTION RECTIFICATION FOR CAMERA CAPTURED DOCUMENTS.....	16
3.1 INTRODUCTION	16
3.2 THEORY	17
3.2.1 <i>planar projective transformation(Homography)</i>	18
3.2.2 <i>Hough Transform</i>	20
3.3 <i>PROPOSED Algorithm</i>	23
3.3.1 PREPROCESSING.....	23
3.3.2 <i>Text line detection</i>	26
3.3.3 <i>Page frame detection</i>	28
3.3.4 <i>Rectification</i>	33
3.4 RESULTS	36
CHAPTER 4.....	38
ARABIC DOCUMENT IMAGE DEWARPING BASED ON TEXT LINES CORRECTION	38
4.1 INTRODUCTION	38
4.2 OVERVIEW	40
4.3 PROPOSED ALGORITHM.....	41
4.3.1 <i>Preprocessing</i>	41
4.3.2 <i>Connected component detection</i>	44
4.3.3 <i>Non linear rectification</i>	47
4.4 RESULTS	49

CHAPTER 5.....	54
ARABIC DOCUMENT IMAGE DEWARPING BASED ON TEXT LINES	
CORRECTION.....	54
5.1 INTRODUCTION	54
OVERVIEW	55
5.2 ALGORITHM	56
5.2.1 <i>Estimation of document image warping</i>	56
5.2.1.1 Horizontal text line detection	56
5.2.1.1.1 Line tracing using self-similarity measure	56
5.2.1.1.2 Resampling traced lines	57
5.2.1.1.3 Line refinement	58
5.2.1.2 Text orientation estimation using local stroke statistics	58
5.2.2 <i>Reconstruction from a single image</i>	59
5.2.3 <i>Image rectification</i>	61
5.3 FAILURE CASES	62
5.4 OUR CONTRIBUTION	62
5.4.1 <i>Text lines tracing problem</i>	62
5.4.2 <i>Non text region rectification</i>	64
5.5 RESULTS	67
CHAPTER 6.....	83
CONCLUSIONS AND FUTURE WORK	83
4.1 CONCLUSION	83
4.2 SCOPE OF FUTURE WORK	84
REFERENCES.....	85

List of Figures

1-1 Challenges with camera-captured images.....	6
1-2 High level illustration of geometric document image rectification	8
1-3 Flow Chart for Skew and Perspective Distortion Rectification	9
1-4 Flow Chart of Arabic Text Lines Rectification	9
1-5 Flow Chart 3D Document Rectification	10
3-1 System flow diagram	17
3-2 Planar projective transformation	18
3-3 Cartesian coordinate	21
3-4 Geometric interpolation of parameters θ and ρ	22
3-5 Failure of Hough transform with unpreprocessed image	24
3-6 Camera captured document image	25
3-7 Camera captured document image after applying Sobel horizontal kernel	26
3-8 Camera captured document after applying canny edge detector	26
3-9 Camera captured document image in rho and theta plan	27
3-10 Hough peaks	27
3-11 Horizontal text lines detection	28
3-12 General document with non uniform paragraphs	28
3-13 Estimated text lines	29
3-14 Hough lines slope histogram	30
3-15 Hough lines length histogram.....	30
3-16 Vertical sides estimation after using RANSAC	31
3-17 Frame detection	32
3-18 The four point in image plan and in the fronto-parallel plan.....	34
3-19 Original image.....	35
3-20 Rectified image	35
4-1 Base line connecting letters in English language	35
4-2 Base line detection of Arabic language	40
4-3 system flow diagram	41
4-4 Binarization using global threshold across the image	42
4-5 Output of Binarization process under different squares sizes (16, 32, 64,128 and 256) ..	43
4-6 Noise removal process.....	44
4-7 Text line with texture noise from the lower text line.....	44

4-8 Text line after removing small connected components	44
4-9 Segmentation of connected components.....	45
4-10 Centriod of connected components.....	45
4-11 Residual of curve fitting of curled text line	46
4-12 Norm value of residual	44
4-13 Cubic line curve fitting for text line	45
4-14 Text line after removing texture noise	45
4-15 Estimation of word skew angle.....	45
4-16 Perspective distortion rectification.....	48
4-17 Image for book captured by mobile camera	52
5-1 System flow diagram	55
5-2 Line tracing	57
5-3 Line resampling.....	57
5-4 Vertical direction estimation	59
5-5 3D document shape construction	61
5-6 incomplete seeds lines	63
5-7 bad resampling due to incomplete seed lines.....	Error! Bookmark not defined.
5-8 Reconstruction of seed lines.....	64
5-9 example of document having figure inside	65
5-10 Projection profile for the document image.....	66
5-11 Histogram of inter-peaks spacing.....	66
5-12 Camera captured document image (A)	68
5-13 Rectification with original algorithm for image (A)	69
5-14 rectified Image with modified algorithm for image (A)	70
5-15 Camera captured document (B)	77
5-16Rectification with Original algorithm for image (B).....	78
5-17 Rectification with modified algorithm for image (B).....	79
5-18 Camera captured document image (C)	74
5-19 Rectification using the original algorithm for image (C).....	75
5-20 Rectification using the modified algorithm for image (C)	76
5-18 Camera captured document image (D)	77
5-19 Rectification using the original algorithm for image (D)	78
5-20 Rectification using the modified algorithm for image (D).....	79
5-18 Camera captured document image (E).....	80
5-19 Rectification using the original algorithm for image (E).....	81
5-20 Rectification using the modified algorithm for image (E)	82

List of Tables

1-1 Camera captured documents challenges	5
3-1 Skew and Perspective Distortion Rectification result with RDI	37
3-2 Skew and Perspective Distortion Rectification result with ABBYY	37
4-1 Example of Arabic language letter figures	39
4-2 Curled lines rectification result with RDI.....	49
4-3 Curled lines rectification result with ABBYY.....	50
4-4 Results of Arabic Document Image Dewarping Based on Text Lines Correction	50

List of Abbreviation

Abbreviations

OCR	Optical character recognition
PDA	Personal digital assistant
VPM	Vertical paragraph margins
PF	Paragraph formatting
SVD	Singular value decomposition
PSF	Point Spread Function
CCD	Charge coupled device
PF	Paragraph formatting
HDB	High-contrasted document boundary
RANSAC	Random Sample Consensus
MPI	Mean pixel intensity

Abstract

Due to the rapid progress in digital cameras industry, document images that were captured with digital camera become as another choice for document optical character recognition (OCR). In contrast with traditional scanners, digital cameras introduce a proper, light weight and mobile image capture, which enables many new applications and breathes new life into existing ones. However, image quality degradation arising from the image acquisition process has a severe effect on these applications.

Many algorithms and techniques were introduced to solve skew, perspective distortion and distortion caused by non planar document shapes. Most of the efforts were directed toward documents written in English and similar languages and the performance of these techniques is very good, conversely, algorithms and techniques handle such problems for documents written in Arabic and similar languages are still not mature.

We present rectification frameworks for restoring the frontal-flat view for planar and curved Arabic document from a single camera-captured image. The first approach proposes skew and perspective distortion rectification for planner document based on frame detection. The approach is introduced mainly for Arabic language, but gives also good results with English language. The second approach aims to rectify the curved layout of Arabic document images. This approach depends on the Arabic characters structure to rectify the curved text lines.

We also proposed some enhancements for a published language independent approach that restores the flat shape of document image for documents that take curved layout. This approach depends on texture flow information obtained directly from the image to rectify the curved shape. In contrast with other approaches, this approach does not depend on any prior assumptions about document shape which improves the rectification performance. The approach also does not use any additional tools like scanners and projectors in the rectification process.

Experiments show that our methods produce images that are significantly more OCR compatible than the original images. Experimental results on 100 image set show that the

proposed algorithms bootstrap the performance of OCR commercial packages with digital cameras which opens the road to new useful applications. Experimental results show that ABBYY FineReader12 performance is enhanced by 10~16%, while RDI performance is enhanced by 40~52%.

Chapter 1

Introduction

Traditionally, scanners are used to digitize textual content in books, newspapers and articles and read through optical OCR. OCR community started to consider digital cameras as an alternative to scanners for document acquisition due to the recent advances in digital cameras industry. Clearly capturing document images is easier using a digital camera rather than a traditional scanner. Digital cameras can be easily integrated in modern mobile devices such as personal digital assistant (PDA), cell phones and media players. As a result, camera based document analysis attracted more research efforts.

Employing character recognition techniques along with text detection algorithms on portable devices assist users in understanding or gathering useful information around them. The rapid progress in the portable devices operating systems introduced many functional applications, which enrich users experience and make life easier. For example, mobile applications like the translation tools and application that help blind people to listen to book's contents instead of reading them using Braille tactile writing system depend mainly on the mobile camera picture as an input to OCR packages.

However, convenience of using the digital camera introduces some image quality problems. When a user captures a photo of a document page, the user may often find that the document is warped or takes curved layout in the captured images. This is especially true when the user aims to capture images for pages from books. Normally, the books are thick and strongly bounded which forces the pages to take the curved layout.

OCR packages introduce very good results for scanned documents as their images were taken without any distortion, but the imperfections of images captured by digital cameras lead to degradation on OCR performance. In particular, skew, perspective projection and non-planar document shape which are common in camera-captured images are not expected at all by traditional OCR algorithms. As a result, the performance of some of the state-of-the-art OCR packages on document image captured by digital cameras is inappropriate for end users.

As a result, preprocessing for a camera captured documents images become a must. The target of this step is the rectification of images captured by digital cameras to restore the front-parallel view of the document and make document lines straight as skewed characters make both segmentation and recognition difficult.

1.1 Camera Captured Document Preprocessing Challenges

OCR technologies can produce excellent results from clean documents as these document images are captured without any distortion. The techniques assume high resolution, high-quality document images with fairly simple structure (black text on a white background). Unfortunately, these assumptions are not typically valid for images captured by a digital camera. The authors in [1] illustrate the major challenges for document images captured by camera as below:

1. Low resolution

Images taken with cameras suffer from the low resolution problems. Most of OCR engines are tuned to handle text between 150 and 400 dpi; however, the same text in a video frame may not exceed 50 dpi which makes simple tasks such as segmentation difficult. The great progress in the camera technologies introduces cheap, small and high resolution output cameras. As a result, we cannot consider this as real problem.

2. Uneven lighting

Opposite to scanners, Camera has fewer control lighting conditions on objects. Due to the physical environment and bumpy response from the objects, uneven lighting is a common case for all cameras captured images. Additional complications occur if we try to use artificial light sources. For example, in case that on-camera flash is used, the center of the view is the brightest and lighting decays outward.