

Offline Optical Character Recognition for Urdu Script



Ayesha Rafiq

Enrollment# 01-244121-002

SUPERVISED BY

Dr. Shehzad Khalid

A thesis is partial fulfillment of requirements for degree of
MS (Software Engineering)

Copyright© 2015. By Ayesha Rafiq

All rights reserved

DECLARATION

The substance of this report is the original work of the author and due references and acknowledgements have been made, where necessary, to the work of others. No part of this thesis has been already accepted for any degree, and it is not being currently submitted in candidature of any degree.

FINAL RESEARCH THESIS
MS (Software Engineering)

“Offline Optical Character Recognition for Urdu Script”



SUBMITTED BY

Ayesha Rafiq

01-244121-002

MS (Software Engineering)

SUPERVISED BY

Dr. Shehzad Khalid

Department of Computer and Software Engineering

Bahria University

ISLAMABAD

Session 2012 -2014

Acknowledgement

Countless thank to Almighty Allah, Lord of the Lords, Creator of the Universe, Worthy of all Praise, who guides in the darkness and helps in difficulties. All respects for his last Holy Prophet Hazrat Muhammad (ﷺ), who enabled me to recognize my creator.

I owe my deepest gratitude to my supervisor Engr. Dr. Shehzad Khalid, Head of Department of Computer Engineering. Despite his hectic schedule as Head of Department, he is always available for his invaluable guidance, moral support, healthy criticism and strong motivation. He has set a role model as a teacher who cares and loves his students as if they were his own kids. Thanks Engr. Dr. Shehzad Khalid.

I also want to thank my co-supervisor, Dr. Imran Ahmed Siddqui for his supportive behavior at every step of my research work.

My sincere gratitude to Dr. Shazia Noreen and Khadija Noureen, who provided encouragement, good company, and lots of good ideas. I would have been lost without them.

I am grateful to my friends Sadia Maqbool, Sunia Hassan, Nadia Hanif, Kurram Shehzad and Mazhar Iqbal Rana for all the good time we spent together and for their continued moral support thereafter. I am indebted to my family for their unflagging love and support throughout my life; this dissertation is simply impossible without them. This dissertation would not have been possible without support of Atiq Ahmad, Iftikhar Ali, Babur Shafiq and Asim Qazi who accompanied me in distant traveling.

DEDICATION

*This dissertation is dedicated to
my parents...*

Abstract:

Development of OCR system for Urdu language has been much challenging task for Urdu researchers for last few years. Intensive complex behavior of Urdu language system is one of prime reason. Urdu images are difficult to understand or manipulate properly unlike English. Retrieving text, sorting out diacritics, and more other functionalities are almost becomes impossible, until or unless they do not have satisfactory domain knowledge of the concerned field. In view of research limitations, proposed work in existing area, presents segmentation free approach using ligature base recognition for various fonts size and different writing style of Urdu. Binary image of Urdu text separates into individual lines. By using connected component labeling on segmented lines extracted ligature along with diacritics. After extraction of ligatures and diacritics, diacritics connected with their respective ligature and then these associated ligatures consider as basic recognition unit. Total 2017 clusters are used in our research; half of them serve as training data and remaining treated as test data. Discrete Fourier Transform (DFT) extracted feature vectors for data set. K-Nearest Neighbor was used to find closest node to query ligature. Our Propose system handled five type of diacritics i.e. different number and position of dots, hamza(ء), toay(ٹ), diacritics connected with haey(ہ) and gaaf(گ). The proposed system evaluated on 70595 most commonly used ligatures of Urdu script and found system is able to recognize Urdu ligature with accuracy rate 98.6%.

Table of Contents

1 Introduction.....	9
1.1 Character Recognition.....	9
1.2 Optical Character Recognition	10
1.2.1 Online Character Recognition	10
1.2.2 Offline Character Recognition	11
1.3 Application of OCR	11
1.4 Techniques for Urdu OCR	12
1.4.1 Segmentation-Based Approach	12
1.4.2 Segmentation-Free Approach.....	12
1.5 Thesis Contribution	13
1.6 Objectives of the Thesis	13
1.7 Thesis Outline	14
2 Literature Review	15
2.1 Converting Documents.....	15
2.2 Perform Online Search	15
2.3 Image Acquisition	16
2.4 Pre-Processing.....	16
2.4.1 Binarization	17
2.4.2 Thinning	17
2.4.3 Noise Removal	18
2.4.4 Smoothing.....	19
2.4.5 Unit Isolation	20
2.5 Segmentation	20
2.6 Feature Extraction	21
2.7 Classification.....	22
2.8 Related Work in Urdu OCR	23
2.9 Comparison of Related Work With Urdu OCR	25
3 Complexities of Urdu Script Writing.....	28
3.1 Urdu Character Set	28
3.2 Characteristics of Urdu Script	29
3.3 Large Number of Diacritics.....	30
3.4 Cursiveness.....	30

3.5 Context Sensitive.....	31
3.6 Bi-directional.....	32
3.7 Positioning and Spacing	33
3.8 Overlapping.....	34
3.8.1 Intra Ligature Overlapping	35
3.8.2 Inter Ligature Overlapping.....	35
3.9 Uneven Stoke Width	35
3.10 Complex Dot Position Rule.....	36
4 Proposed Methodology	37
4.1 Binarization	38
4.2 Segmentation of Image into Lines	38
4.3 Segmentation of Line into Ligatures and Diacritics.....	39
4.4 Associate Diacritics with Respective Ligature.....	40
4.5 Ligature Recognition.....	41
5 Experiments and Results.....	45
5.1 Dataset.....	45
5.2 Performance of Identification of Ligatures and Diacritics	46
5.3 Performance Associate Diacritic with Respective Ligature.....	46
5.4 Performance of Recognition.....	47
5.5 Competitor Analysis.....	48
6 Conclusion and Future Work	50
6.1 Conclusion.....	50
6.2 Future Work	50
References	52

List of Figures

Figure 1.1: Classification of Optical Character Recognition	10
Figure 2.1: Binarization Process ^[10]	17
Figure 2.2: (a) Original Text; (b) Text after skeletonization	18
Figure 2.3: (a) Skewed Document (b) De-Skewed Document	20
Figure 2.4: Segmentation of Urdu text into lines using Horizontal Projection	21
Figure 3.1: Bi-directional Text.....	28
Figure 3.2: Word, Ligature and Isolated Characters	31
Figure 3.3: Context Sensitivity; Three different shapes of bey-initial	32
Figure 3.4: Bi-directional movement	32
Figure 3.5: Bi-directional writing	33
Figure 3.6: Positioning and Spacing	34
Figure 3.7: Character Overlapping.....	34
Figure 3.8: Intra Ligature Overlapping	35
Figure 3.9: Inter Ligature Overlapping	35
Figure 4.1: Block Diagram of Proposed Method	37
Figure 4.2: Connected component b) represent ligatures c) highlight diacritics	39
Figure 4.3: Associated diacritics with their respective ligature	40
Figure 4.4: Connected Ligature.....	41
Figure 4.5: 1-D time series ligature representation	42
Figure 5.1: Example of Diacritics	45
Figure 5.2: Some Ligatures without Diacritics	46

List of Tables

Table 2.1: Comparison of Previous Research Work	25
Table 3.1: Shapes of Urdu Script	28
Table 3.2: Diacritics Type and Positioning	30
Table 5.1: Extracted of Diacritics and Ligatures	46
Table 5.2: Comparison of Proposed Work with Existing Research Work	48

List of Abbreviation

CR	Character Recognition
OCR	Optical Character Recognition
AI	Artificial Intelligence
DCT	Distinct cosine Convert
UOCR	Urdu Optical Character Recognition
DFT	Discrete Fourier Transform