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)

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(ط), 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

List of Figures

List of Tables

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