Optical Character Recognition System For Urdu Words in Nastaliq Font

Safia Shabbir

Enrollment #: 01-244112-023

Supervised By

Dr. Imran Ahmed Siddiqi

Department of Computer and Software Engineering

Bahria University, Islamabad Campus

2011-2014

Optical Character Recognition System For Urdu Words in Nastaliq Font

A THESIS SUBMITTED TO THE BAHRIA UNIVERSITY IN THE PARTIAL FULFILLMENT OF REQUIREMENTS FOR THE DEGREE OF MS SOFTWARE ENGINEERING

Safia Shabbir

Enrollment #: 01-244112-023

Supervised By

Dr. Imran Ahmed Siddiqi

Department of Computer and Software Engineering

Bahria University Islamabad

2011-2014

CERTIFICATE OF ORIGINALITY

I certify that the intellectual contents of the thesis

"Optical Character Recognition System for Urdu Words in Nastaliq Font"

is the product of my own research work except, as cited properly and accurately in the acknowledgment and references, the material taken from any source such as research papers, research journals, books, internet, etc. solely to support, elaborate, compare and extend the earlier work, Further, this work has not been submitted previously for a degree at this or any other University.

The incorrectness if the above information, if proved at any stage, shall authorize the university to cancel my degree.

Signature: _____. Dated: January 31st , 2014 .

Name of the Research student: Safia Shabbir .

ACKNOWLEDGMENTS

First of all, I thank to Allah Almighty for giving me the opportunity and potential to complete this dissertation. I would like to express my deepest gratitude to my supervisor Dr. Imran Ahmed Siddiqi, for the courage, endless support and guidance throughout my research. It has been an honor to work in his supervision. I am grateful for his precious time, ideas and knowledge that has made my research an unforgettable experience for me. Without his motivation and guidance it would have been impossible to remain firm in obscure situations. His enthusiasm towards research was motivational for me during tough times in my research. His courageous behaviors always boosted up my morale. Above all I salute to his patience and care at every stage of my research.

Last, but not the least I would like to thank my beloved parents and brothers for their continuous support, love and encouragement for the completion of this dissertation. I am also grateful to my friends specially Azra Batool for the concern, help and motivation regarding this research.

DEDICATION

This thesis is dedicated to my beloved parents, Shabbir Hussain and Fiaz Bibi, for being role models for me, and my brothers Zahid Shabbir, M. Asim Shabbir and M. Waqas Shabbir for their encouragement and support to achieve my goals.

ABSTRACT

Optical Character Recognition (OCR) has been an attractive research area for the last three decades and mature OCR systems reporting near to 100% recognition rates are available for many scripts/languages of the world today. Despite these developments, research on recognition of text in many languages is still in its early days, Urdu being one of them. The limited existing literature on Urdu OCR is either limited to isolated characters or considers limited vocabularies in fixed font sizes. This research presents a segmentation free and size invariant technique for recognition of Urdu words in Nastaliq font using ligatures as units of recognition. Connected component labeling is applied to binarized images of Urdu text to extract ligatures which are separated into primary ligatures and diacritics. Ligatures extracted from a set of documents are represented by profile and projection features and grouped into clusters using Dynamic Time Warping (DTW) as the (dis)similarity measure. A total of 250 clusters of frequent Urdu ligatures are considered in our study. These clusters serve as training data to train a separate right-to-left Hidden Markov Model (HMM) for each ligature. Ligatures (main body as well as diacritics) of the query word are recognized by their respective HMMs. Using position information; diacritics are associated with their corresponding ligatures which are then validated by a dictionary. Unicode of the complete word is finally written to a text file. The proposed system evaluated on 100 query words realized promising results at ligature and word level recognition.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

OCR	Optical Character Recognition
DTW	Dynamic Time Warping
HMM	Hidden Markov Model
CLE	Centre of Language Engineering
UPTI	Urdu Printed Text Image Database
KNN	K Nearest Neighbors
TTF	True Type Font
CENPARMI	Centre for Pattern Recognition and Machine Intelligence
IFN	Institute of Communications Technology
ENIT	Ecole Nationale d'Ingénieurs de Tunis