

# 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 31<sup>st</sup>, 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

<b>1</b>	<b>INTRODUCTION .....</b>	<b>1</b>
1.1	CATEGORIES OF OPTICAL CHARACTER RECOGNITION SYSTEMS .....	2
1.1.1	<i>Printed and Handwritten text recognition.....</i>	2
1.1.2	<i>Online and Offline Recognition.....</i>	3
1.1.3	<i>Single and Omni Font OCRs .....</i>	3
1.1.4	<i>Isolated and Cursive script.....</i>	3
1.2	PROBLEM STATEMENT .....	3
1.3	PROPOSED METHODOLOGY AND RESEARCH CONTRIBUTIONS.....	4
1.4	THESIS OUTLINE .....	4
<b>2</b>	<b>URDU TEXT – CHARACTERISTICS AND CHALLENGES .....</b>	<b>5</b>
2.1	URDU ALPHABET .....	5
2.2	URDU SCRIPTS.....	6
2.3	CHALLENGES WITH URDU TEXT.....	7
2.3.1	<i>Cursive Text.....</i>	7
2.3.2	<i>Context Sensitivity .....</i>	8
2.3.3	<i>Bidirectional Behavior .....</i>	8
2.3.4	<i>Diagonal Writing Direction.....</i>	9
2.3.5	<i>Overlapping.....</i>	9
2.3.6	<i>Diacritics and Their Position .....</i>	9
2.3.7	<i>Absence of Fixed Baseline.....</i>	10
2.3.8	<i>Filled Loops and False Loops .....</i>	11
2.3.9	<i>Spacing .....</i>	11
2.3.10	<i>Positioning.....</i>	11
2.4	BASIC STEPS IN URDU OCR .....	12
2.4.1	<i>Image Acquisition.....</i>	13
2.4.2	<i>Preprocessing.....</i>	13
2.4.3	<i>Segmentation .....</i>	14

2.4.4	<i>Feature Extraction</i> .....	14
2.4.5	<i>Classification/Recognition</i> .....	15
2.5	SUMMARY .....	15
<b>3</b>	<b>LITERATURE REVIEW</b> .....	<b>16</b>
3.1	APPROACHES FOR URDU OCR .....	16
3.1.1	<i>Segmentation-Based Approaches</i> .....	17
3.1.2	<i>Segmentation-Free Approaches</i> .....	20
3.1.3	<i>Comparison of Reviewed Techniques</i> .....	23
3.2	SUMMARY .....	28
<b>4</b>	<b>PROPOSED METHODOLOGY</b> .....	<b>30</b>
4.1	TRAINING .....	30
4.1.1	<i>Preprocessing</i> .....	31
4.1.2	<i>Connected Components Labeling/Extraction</i> .....	32
4.1.3	<i>Feature Extraction</i> .....	34
4.1.4	<i>Clustering of Ligatures</i> .....	37
4.1.5	<i>Ligature Modeling</i> .....	47
4.2	RECOGNITION OF WORDS.....	53
4.2.1	<i>Input Word Acquisition</i> .....	54
4.2.2	<i>Baseline and Ligature Extraction</i> .....	54
4.2.3	<i>Recognition Through HMM</i> .....	55
4.2.4	<i>Association of Diacritics with Ligatures</i> .....	56
4.2.5	<i>Recognizing the Ligatures</i> .....	58
4.2.6	<i>Text Output</i> .....	63
4.3	SUMMARY .....	63
<b>5</b>	<b>EXPERIMENTS AND RESULTS</b> .....	<b>65</b>
5.1	DATA SET.....	65
5.2	PERFORMANCE OF LIGATURE CLUSTERING .....	65
5.3	PERFORMANCE OF RECOGNITION .....	65



5.3.1	<i>Ligature Recognition Rate</i> .....	65
5.3.2	<i>Word Recognition Rate</i> .....	67
5.4	LIGATURE RECOGNITION ON CLE DATASET .....	68
5.5	SUMMARY .....	69
<b>6</b>	<b>CONCLUSION AND PERSPECTIVES</b> .....	<b>70</b>
6.1	CONCLUSION .....	70
6.2	FUTURE PERSPECTIVES .....	70
	<b>BIBLIOGRAPHY</b> .....	<b>72</b>
	<b>APPENDIX</b> .....	<b>75</b>
	FEW SAMPLE CLUSTERS .....	75

## List of Figures

Figure 1.1: General OCR System .....	1
Figure 2.1: Urdu character set.....	5
Figure 2.2: Urdu Word composition a) Urdu Word ‘Pakistan’ b) Characters of ‘Pakistan’ c) Urdu word ‘Tasbih’ d) Characters of ‘ Tasbih’ (Image Source: [6]).....	5
Figure 2.3: a) Shapes of character ‘Alif’ b) Shapes of character ‘Hay’ c) Urdu word ‘Rana’ representing shapes of ‘Alif’ d) Urdu words representing different shapes of ‘Hay’ (Image Source: [4]) .....	6
Figure 2.4: Different Calligraphic Styles of Urdu script (Image Source: [7]).....	7
Figure 2.5: Ligatures and characters of word ‘Pakistan’ (Image Source: [6]).....	8
Figure 2.6: Bidirectional behavior of Urdu Script (Image Source: [4]) .....	9
Figure 2.7:Diagonal writing direction (Image Source: [4]).....	9
Figure 2.8: Examples of ligature overlapping (Image Source: [9]) .....	9
Figure 2.9:Overlapping of characters within a ligature (Image Source: [9]).....	9
Figure 2.10: Common diacritics .....	10
Figure 2.11: Uncommon diacritics .....	10
Figure 2.12: Complex placement of dots.....	10
Figure 2.13: Baselines for Naskh and Nastaliq Urdu Font (Image Source: [6]).....	10
Figure 2.14: Loops in Nastaliq and Naskh (Image Source: [10]).....	11
Figure 2.15: Example of false loop.....	11
Figure 2.16: Spacing in Urdu text.....	11
Figure 2.17: Positioning of characters in Urdu.....	12
Figure 2.18: General steps of Urdu OCR.....	13
Figure 4.1: Block diagram for Training.....	31
Figure 4.2: a) Original image b) Binarized image with global thresholding.....	32
Figure 4.3: a) 4-neighborhood b) 8-neighborhood .....	33
Figure 4.4: a) Binarized image b) Connected component labeling .....	33
Figure 4.5: a) Connected components of word ‘Mojudgi’ b) Overlapped ligatures c) Extracted individual ligatures and dot.....	34
Figure 4.6: Normalized horizontal projection of Bari-yay .....	35

Figure 4.7: Normalized vertical projection of Bari-yay.....	35
Figure 4.8: Normalized upper profile of character ‘Bari-yay’ .....	36
Figure 4.9: Normalized lower profile of character ‘Bari Yay’ .....	36
Figure 4.10: Alignments for Profiles (Image Source: [31]).....	39
Figure 4.11: a) Euclidean distance b) Warped time axis .....	39
Figure 4.12: Frequencies of 250 ligatures.....	45
Figure 4.13: Images of two ligatures “ba” and “na” (a): With dots (b): Without dots .....	45
Figure 4.14: Sliding Windows a) Overlapping in Sliding Windows b) Selected window .....	51
Figure 4.15 Projection features a) selected window b) horizontal projection c) vertical projection .....	52
Figure 4.16: Markov chain model generated for Ligature 'Umar' .....	53
Figure 4.17: Flow chart for Recognition of Urdu Words .....	54
Figure 4.18: Examples of query words .....	54
Figure 4.19: a) Input word ‘Aiwan’ b) Baseline detection .....	54
Figure 4.20: a) Binary image of word ‘Aiwan’ b) Ligatures extracted from word .....	55
Figure 4.21: Ligatures of word ‘Aiwan’ b) diacritics separated from ligatures .....	57
Figure 4.22: (a) Ligature Bay + Bay + Alif with ‘one dot above’ and ‘one dot below’ diacritic array .....	62
Figure 4.23: Output text file for word ‘Aiwan’ .....	63
Figure 5.1: Examples of query words .....	66
Figure 5.2: Recognition rate as a function of number of ligatures (CLE Database) .....	69
Figure A.1: Cluster of ligature ‘Bay + Kaaf’ .....	76
Figure A.2: Cluster of ligature ‘Choti yay’ .....	76
Figure A.3: Cluster of ligature ‘Bay + Alif’ .....	77

## List of Tables

Table 2.1: Examples of multiple shapes of a character in Urdu (Source:[8]).....	8
Table 3.1: Comparison of Segmentation-Based Approaches .....	23
Table 3.2: Comparison of Segmentation-Free Approaches .....	25
Table 4.1: Summary of Features.....	37
Table 4.2: Pseudo code for DTW algorithm.....	39
Table 4.3: Instances of 250 Clusters .....	41
Table 4.4: Character Classes and Unicode .....	46
Table 4.5: Diacritics and corresponding codes .....	53
Table 4.6: Ligatures of word ‘Aiwan’ and corresponding Unicodes after recognition .....	55
Table 4.7: Ligature and their corresponding classes.....	56
Table 4.8: Diacritics with corresponding integer values and descriptions .....	57
Table 4.9: Ligature and respective diacritic arrays.....	58
Table 4.10: Ligature and respective character classes .....	58
Table 4.11: Character class and respective class members with respect to diacritics .....	59
Table 4.12: Ligature and possible characters for word ‘Aiwan’ .....	63
Table 5.1: Types of ligatures and corresponding recognition rates .....	66
Table 5.2: Overall recognition rate .....	66
Table 5.3: Word Recognition rate.....	67
Table 5.4: Studies and their limitations .....	67

## List of Abbreviations

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