



ISRAR UDDIN
01-281121-005

Optical Character Recognition for Printed Urdu Nastaliq Font

*A thesis submitted to the Department of Computer Engineering, Faculty of Engineering Sciences,
Bahria University, Islamabad, in the partial fulfillment for the requirements of a Doctoral degree in
Computer Engineering*

Supervisor: Dr. Imran Siddiqi

Co-Supervisor: Dr. Shehzad Khalid

Department of Computer Engineering
Bahria University, Islamabad

March 2019

Abstract

Optical Character Recognition (OCR) is one of the most investigated pattern classification problems that has received remarkable research attention for more than half a century. From the simplest systems recognizing isolated digits to end-to-end recognition systems, applications of OCRs vary from postal mail sorting to reading systems in scene images facilitating autonomous navigation or assisting the visually impaired. Despite tremendous research endeavors and availability of commercial recognition engines for many scripts, recognition of cursive scripts still remains an open and challenging research problem mainly due to the complexity of script, segmentation issues and large number of classes to recognize. Among these, Urdu makes the subject of our study. More specifically, this study investigates the recognition of printed Urdu text in Nastaliq style, the most widely employed script for Urdu text that is more complex than the Naskh style of Arabic.

This work presents a holistic (segmentation-free) technique that exploits ligatures (partial words) as units of recognition. Urdu has a total of more than 26,000 unique ligatures, many of the ligatures, however, share the same main body (primary ligature) and differ only in the number and position of dots and diacritics (secondary ligatures). We exploit this idea to separately recognize the primary and secondary ligatures and later re-associate the two to recognize the complete ligature. Recognition is carried out using two techniques; the first of these is based on hand-crafted statistical features using hidden Markov models (HMMs). Features extracted using sliding windows are used to train a separate model for each ligature class. Feature sequences of the query ligature are fed to all the models and recognition is carried out through the model that reports the maximum probability. The second technique employs Convolutional Neural Networks (CNNs) to automatically extract useful feature representations from the classes and recognize the ligatures. We investigated the performance of a number of pre-trained networks using transfer learning techniques and trained our own set of networks from scratch as well.

Experimental study of the system is carried out on two benchmark datasets of Urdu text, the 'Urdu Printed Text Images' (UPTI) database and the 'Center of Language Engineering' (CLE) database. A number of experimental scenarios are considered for system evaluation and the realized recognition rates are compared with state-of-the-art recognition systems for printed Urdu text. An interesting aspect of experimental study is the combination of unique ligatures in the two datasets to generate a large set of around 2800 unique primary and secondary ligatures covering a major proportion of the Urdu corpus. The system reports high classification rates (88.10% and 94.78% on CLE and UPTI query ligatures respectively) demonstrating the effectiveness of the proposed recognition techniques which can be adapted for other cursive scripts as well. The findings of this study are expected to be useful for the document recognition community in general and researchers targeting cursive scripts in particular.

Contents

Abstract	i
1 Introduction	1
1.1 Recognition Systems - A Historical Perspective	2
1.2 Classification of OCR Systems	4
1.3 Motivation	6
1.4 Problem Statement	9
1.5 Research Objectives	9
1.6 General Steps in OCR	10
1.6.1 Image Acquisition	10
1.6.2 Image Pre-processing	10
1.6.3 Segmentation	11
1.6.4 Feature Extraction	11
1.6.5 Classification (Recognition)	12
1.6.6 Post Recognition Processing	13
1.7 Proposed Techniques	13
1.8 Thesis Contribution	14
1.9 Thesis Organization	14
2 Literature Review	15
2.1 Overview of Urdu	15
2.1.1 Urdu Alphabet and Numerals	16
2.1.2 Writing Styles	16
2.1.3 Recognition Challenges of Urdu Text	17
2.2 Datasets	23
2.2.1 CENPARMI Urdu Database	23
2.2.2 Urdu Handwritten Sentence Database (UHSD)	24
2.2.3 UCOM Offline Handwritten Dataset	24
2.2.4 Urdu Printed Text Images Database (UPTI)	24
2.2.5 Center of Language Engineering (CLE) Urdu Database	24
2.3 Recognition Techniques for Urdu Text	25
2.3.1 Analytical Techniques	27
2.3.2 Holistic Techniques	30
2.4 Summary	36
3 Data Preparation	37
3.1 Binarization	37
3.2 Data Preparation	37

3.2.1	Scenario I - Ligatures of the UPTI Dataset	38
3.2.2	Scenario II - High Frequency Complete Ligature (HFCL) Clusters in the CLE the Database	41
3.2.3	Scenario III - Ligatures extracted from images of books in the CLE database	42
3.2.4	Scenario IV - Combination of CLE and UPTI Ligatures	43
3.3	Mapping of Character Classes	44
3.4	Summary	45
4	Recognition using Hidden Markov Models	47
4.1	Overview of HMMs	47
4.1.1	Learning Problem	48
4.1.2	Evaluation Problem	48
4.1.3	Decoding Problem	48
4.2	Ligature Modeling using HMMs	49
4.2.1	Feature Extraction	49
4.2.2	Training of Models	52
4.3	Recognition of Ligatures	52
4.3.1	Grouping of Primary and Secondary Ligatures	54
4.3.2	Recognition of Primary and Secondary Ligatures	57
4.3.3	Association of Primary & Secondary Ligatures	58
4.4	System Evaluation	60
4.4.1	Results & Analysis	62
4.4.2	Performance Sensitivity to System Parameters	64
4.5	Summary	66
5	Recognition using Convolutional Neural Networks	68
5.1	Overview of CNNs	68
5.1.1	Convolutional Layer	69
5.1.2	ReLU Layer	70
5.1.3	Pooling Layer	71
5.1.4	Fully Connected (FC) Layer	71
5.1.5	CNNs and Transfer Learning	71
5.2	CNN Architectures	72
5.2.1	AlexNet	72
5.2.2	VGGNet	73
5.2.3	Proposed Architecture	73
5.3	Ligature Modeling using CNN	74
5.4	System Evaluation	74
5.4.1	Results & Analysis	75
5.4.2	Performance Sensitivity to System Parameters	77
5.4.3	Comparison of Results with Notable Studies	80
5.5	Analysis of Recognition Errors	83
5.6	Discussion	85
5.7	Summary	86
6	Conclusion and Future Work	88
6.1	Conclusion	88
6.2	Future Work	90

A	Sample Text Lines and Ligatures Clusters Images From UPTI Database	91
B	Page Image From CLE Books and Clusters Instances from Scenario-III	95
C	Research Publications	100

List of Tables

2.1	Characters having filled loops in Nastaliq and true loops in Naskh	21
2.2	An overview of databases of Urdu text	25
2.3	A summary of analytical recognition techniques	35
2.4	A summary of holistic recognition techniques	36
3.1	Overview of features to group ligatures into clusters	40
3.2	Mapping of Urdu characters to classes	45
3.3	Table of secondary ligatures with sample images and respective codes	46
4.1	An overview of features computed from each (sliding) window ($h = 32, w = 9$)	52
4.2	Look up table for character classes - Potential occurrences of secondary components	61
4.3	An example character class with corresponding dot/diacritics values forming characters	62
4.4	Summary of the four experimental scenarios	62
4.5	Over all recognition rates of the four experimental scenarios	63
4.6	Length statistics of query complete ligatures in UPTI text lines, CLE HFLs and CLE books	63
4.7	Recognition rates in four experimental scenarios as a function of ligature length	64
5.1	Network parameters for system training	73
5.2	Over all recognition rates of the four experimental scenarios	76
5.3	Post-processing error rates of the four experimental scenarios over CNN and HMM based techniques	76
5.4	Recognition rates as a function of ligature length using CNN based recognizers on Scenarios I-IV	77
5.5	Recognition rates on the four experimental scenarios using transfer learning	77
5.6	Performance comparison of proposed recognition techniques with notable studies reported in the literature	82
5.7	Examples of morphological similar ligatures causing recognition errors (HMM based recognition)	84
5.8	Examples of recognition errors resulting due to false re-association of primary & secondary ligatures	84
5.9	Examples of morphologically similar ligatures resulting in false matches with CNNs	85
5.10	Examples of morphologically similar ligatures where HMM based recognition fails but CNNs report correct recognition	85