



SHAHBAZ HASSAN
01-243172-029

Handwriting Recognition for Cursive Scripts:A Case Study on Urdu Text

A thesis submitted in partial fulfillment of the requirements for the Master degree in Computer Science at Bahria University, Islamabad Pakistan.

Supervisor: Dr. Imran Ahmed Siddiqi

Department of Computer Science
Bahria University, Islamabad

June 2019



Bahria University
Discovering Knowledge

MS-13

Thesis Completion Certificate

Scholar's Name: SHAHBAZ HASSAN Registration No. 01-243172-029

Programme of Study: MS-CS

Thesis Title: Handwriting Recognition for cursive
Scripts: A case study on Urdu Text.

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for Evaluation. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index at 17% that is within the permissible limit set by the HEC for the MS/MPhil degree thesis.

I have also found the thesis in a format recognized by the BU for the MS/MPhil thesis.

Principal Supervisor's Signature: 

Date: 05/08/2019 Name: Dr. Imran Ahmed Siddiqi



Bahria University
Discovering Knowledge

MS-14A

Author's Declaration

I, SHAMBAZ HASSAN hereby state that my MS thesis titled
" Handwriting Recognition for cursive scripts: A case study
on Urdu Text
"

is my own work and has not been submitted previously by me for taking any degree from this university

Bahria University Islamabad

or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduate the university has the right to withdraw/cancel my PhD degree.

Name of scholar: SHAMBAZ HASSAN
Date: 05/08/2019



Bahria University
Discovering Knowledge

MS-14B

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled
"Handwriting Recognition for cursive Scripts: A case
Study on Urdu Text."
is solely my research work with no significant contribution from any other person.
Small contribution / help wherever taken has been duly acknowledged and that complete
thesis has been written by me.

I understand the zero tolerance policy of the HEC and Bahria University towards
plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my
thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even
after award of PhD degree, the university reserves the right to withdraw / revoke my PhD
degree and that HEC and the University has the right to publish my name on the HEC /
University website on which names of students are placed who submitted plagiarized
thesis.

Student / Author's Sign:

Name of the Student:

SHAHBAZ HASSAN

Abstract

Recognition of cursive handwritten text is a complex problem due challenges like context-sensitive character shapes, non-uniform inter and intra word spacings, complex positioning of dots and diacritics and very low inter-class variation among certain classes. This research study presents an effective technique for recognition of cursive handwritten text using Urdu as a case study (though findings can be generalized to other cursive scripts as well). We present an analytical approach based on implicit character segmentation where CNNs are employed as feature extractors and LSTM network is used as a classifier. The proposed technique is validated on UNHD and UHTI (custom generated data set) and reported promising recognition results.

Acknowledgments

First of all, I would thank Allah Subahanhu-Wa-Ta-Ala, the most merciful and beneficent, for giving me the strength and courage to accomplish this research and to achieve this milestone - the MS dissertation.

It is my pleasure to acknowledge my deepest thanks and gratitude to the polite and friendly person, my supervisor, Dr. Imran Ahmed Siddiqi for his guidance and kind supervision during my research.

I am also very thankful to Muhammad Atif and Mr. Ali Mirza for providing an opportunity to work in a research lab to carry out simulations of my experiments.

I would like to thank my friends Mr. Aaqib Mehran, Syed Abdul Basit and Mr. Haider Ali Khan for their encouragement and moral support throughout my MS work.

Finally, I would like to express my very profound gratitude to my parents and family members for always believing in me and encouraging me to reach higher in order to achieve my goals.

SHAHBAZ HASSAN
Bahria University Islamabad, Pakistan

2019

“There are no secrets to success. It is the result of preparation, hardwork and learning from failure.”

Colin Powell

Contents

Abstract	i
1 Introduction	1
1.1 Historical Background of Recognition Systems	2
1.2 Classification of OCR Systems	3
1.3 Background	4
1.4 Problem Statement	4
1.5 Research Objectives	4
1.6 General Steps in OCR	5
1.6.1 Image Acquisition	5
1.6.2 Image Preprocessing	5
1.6.3 Segmentation	6
1.6.4 Feature Extraction	7
1.6.5 Classification (Recognition)	7
1.6.6 Post Recognition Processing	7
1.7 Proposed Technique	7
1.8 Thesis Contribution	8
1.9 Thesis Organization	8
2 Literature Review	9
2.1 Overview of Urdu	9
2.1.1 Urdu Alphabet and Numerals	9
2.1.2 Writing Styles	9
2.1.3 Recognition Challenges of Urdu Text	11
2.2 Datasets	13
2.2.1 Printed Text Data Sets	13
2.2.2 Handwritten Text Data sets	14
2.3 Recognition Techniques for Urdu Text	16
2.3.1 Analytical Approaches	16
2.3.2 Holistic Approaches	18
2.4 Summary	19
3 Methods	21
3.1 Dataset	21
3.1.1 Urdu Handwritten Text Images Database (UHTI)	21
3.2 Proposed Methodology	23
3.2.1 Data Preprocessing	23
3.2.2 Feature Extraction	24

3.3	Classification	28
3.4	Conclusion	31
4	Results and Discussion	33
4.1	Experimental Protocol and Results	33
4.1.1	Character recognition rate on UNHD data set	34
4.1.2	Character recognition rate on UHTI data set	35
4.2	Performance Analysis and Discussion	36
4.2.1	Character recognition using BLSTM with pixel values	36
4.2.2	Character recognition using transfer learning	37
4.3	Comparison	38
5	Conclusion and Perspectives	41
5.1	Conclusion	41
5.2	Perspectives	42
6	Research Publication	43
6.1	Accepted for Publication	43

List of Figures

1.1	(a) An example of a complete Urdu ligature (b) Main body of ligature (Primary ligature) and (c) Secondary ligatures	2
1.2	Optophone - Reading by ear for the blind (Image Source: [4])	2
1.3	Battery operated Optacon reader for the blind (Image Source: [7])	3
1.4	General steps in an (offline) OCR system	5
1.5	OCR techniques from the view point of segmentation	6
1.6	A Workflow of proposed Methodology	8
2.1	Urdu characters written right-to-left	10
2.2	Urdu numbers written left-to-right	10
2.3	Popular writing styles for cursive scripts (Image Source: [36])	10
2.4	A complete ligature (a) With primary and secondary ligatures (b) Without secondary ligatures	11
2.5	Urdu character set (a) Joiners and (b) Non-joiners	12
2.6	Urdu character "Sheen" with possible shapes	12
2.7	Bidirectional behaviour of Urdu text.	13
2.8	Text Overlapping	13
2.9	Variable Spacing	13
2.10	Sample text line from UPTI dataset	14
2.11	Instances in an example cluster from CLE dataset	14
2.12	Statistics of UCOM-Data set	15
2.13	A sample Text line from UCOM-Offline data set	15
2.14	Statistics of UNHD-Data set	15
3.1	A sample Image from UHTI offline data-set	22
3.2	Statistics of UHTI- Dataset	22
3.3	An overview of the recognition system	24
3.4	(a) Grayscale image (b) Binarized image with Otsu's global thresholding	24
3.5	The architecture of Convolutional Neural Network	25
3.6	Convolution of an image(50 X 50 X 3) with a filter(5 X 5 X 3)	26
3.7	Convolution of an image with five filters to produce the output volume	27
3.8	Representation of simple RNN cell	29
3.9	(a): A standard Vanilla RNN Network (b): An LSTM unit based RNN Network	29
3.10	(a): BLSTM Architecture	31
3.11	CTC computing the probability of an output sequence "PAKISTAN" (written in Urdu)	31
4.1	The training loss of a model on UNHD as a function of the number of epochs	34
4.2	Recognition rate as a function of the size of training data	35
4.3	The training loss of a model on UHTI as a function of the number of epochs	35

4.4	Recognition rate as a function of the size of training data	36
4.5	UPTI-Model Training	38
4.6	The training loss of a model during fine-tuning on UNHD data set as a function of the number of epochs	38
4.7	The training loss of a model during fine-tuning on UHTI data set as a function of number of epochs	39
4.8	Examples of recognition errors	40

List of Tables

2.1	Summary of explicit segmentation based techniques	17
2.2	Summary of implicit segmentation based techniques	18
2.3	Summary of holistic segmentation based techniques	19
2.4	Summary of Urdu handwriting recognition systems	20
3.1	Summary of convolutional and Pooling layers	27
3.2	Summary of convolutional and Recurrent layers	30
4.1	UNHD Data set division	33
4.2	UHTI Data set division	34
4.3	BLSTM with pixel value based recognition rate	37
4.4	UPTI Data set division	37
4.5	UNHD and UHTI Datasets division for fine-tuning	37
4.6	Summary of Results	39
4.7	Recognition rates of notable studies on (printed and handwritten) Urdu text	40

Acronyms and Abbreviations

CLE	Center of Language Engineering
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
GPU	Graphical Processing Unit
OCR	Optical Character Recognition
RNN	Recurrent Neural Network
LSTM	Long-Short Term Memory
UPTI	Urdu Printed Text Images
UNHD	Urdu Nastalique Handwritten Dataset
UHTI	Urdu Handwritten Text Images
CENPARMI	Center of Pattern Recognition and Machine Intelligence.
FCC	Freeman chain codes
HMM	Hidden Markov Model
MDD-RNN	Multi-Dimensional Recurrent Neural Network
GPU	Graphical processing unit.

Chapter 1

Introduction

Offline handwritten character recognition has become a challenging area in the field of pattern recognition and image processing. Due to rapid growth in digital technology, there is a need for having mature recognition systems for cursive handwritten documents. The requirement of these recognition systems is to convert the handwritten information into the machine-understandable format. History of character recognition is spanned over many decades and after significant efforts, today mature recognition systems are available. Among this OCRs, the recognition systems for non-cursive language are considered very mature.

Non-cursive scripts, for example, Latin have characters printed and handwritten in isolated form and offer less complexity. Characters in cursive scripts (Urdu, Arabic or Persian) contain complex shapes which are highly challenging to recognize. Several studies have been carried out to recognize cursive scripts that report high recognition rates in printed and handwritten text. Regardless of these efforts and studies, many cursive languages could not gain the attention of researchers due to complexity in writing style and lack of standard datasets, Urdu language is one of them. The Urdu language is mostly influenced by Arabic, Sanskrit, and Persian. Urdu has more than 100 million native speakers all over the world with major share from Pakistan, India and the Middle East. In these regions, a lot of information in manuscripts is available in handwritten form but rarely available in digital format. It has become necessary to digitize this information in order to make it recognizable and searchable through machines. In this regard, handwritten character recognition is an attractive option to achieve the desired goal [2].

Characters in cursive scripts appear in different forms either isolated or are joined with other characters to form words/partial words. This combination of characters is called a ligature. A ligature can be divided into two further components i.e primary ligature (main body) and secondary ligature (dots and diacritic). An example of handwritten ligature along with primary and secondary components is presented in Figure 1.1

Cursive languages are written in different writing styles such as Naskh, Koffi, and Nastalique. Arabic and Pashto are written in Naskh while Urdu is mostly used Nastalique style. The Nastalique writing style offers more complexity. From the viewpoint of recognition, the challenges offered by Nastalique style include variable spacing between words, text overlapping, no fixed baseline and

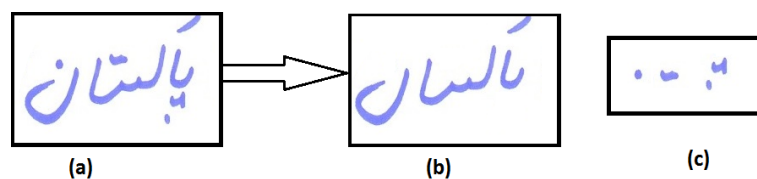


Figure 1.1: (a) An example of a complete Urdu ligature (b) Main body of ligature (Primary ligature) and (c) Secondary ligatures

filled or false loop of characters. More challenges associated with Urdu handwriting recognition are presented later in the document.

1.1 Historical Background of Recognition Systems

From many decades, printed and handwritten information is being used as a major source of communication as well as gathering and preserving important information. Humans always had the desire to design a machine that can read the text. In this regard, the first retinal images scanner was proposed for classification of photocells. In 1912 the first optophone (Figure 1.2) came into the scene which was used to read characters when moved over the printed text. Another important milestone was achieved by Gustav Tauschek in 1929 [3]. He invented a first photo-sensor based "Reading Machine" that can read a printed text letter by letter and produce the output in the form of speech.



Figure 1.2: Optophone - Reading by ear for the blind (Image Source: [4])

With further of technological development in 1954, an American Magazine (Reader's Digest) first time employed an OCR to convert their handwritten reports into punch cards [5]. In 1962 another milestone was achieved by Louis H Goldish. He invented the electromechanical reading device "Optacon" (Figure 1.3) that can read printed text for blind people. In 1965 Reader's Digest enhanced the functionality of their OCR machines in order to recognize the numbers. The second generation reading systems appeared in the late 1960s and early 1970s [6]. In this duration, an effort was made to standardize a font for automatic recognition of text. The third generation recognition system appeared in the 1970s. The main objective of these recognition systems was to address the

problems of low-quality images and large handwritten and printed characters sets. In the same era, the Omni OCRs were proposed to handle the multiple fonts in a text.

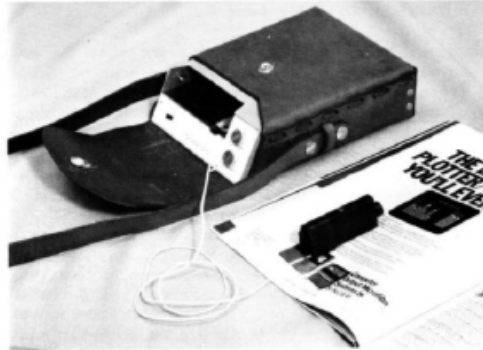


Figure 1.3: Battery operated Optacon reader for the blind (Image Source: [7])

In the 1990s number of business migrated to automatic recognition systems due to the combination of Artificial intelligence with image processing operations. The recognition systems were employed to read passport number, price, and credit card. With the tremendous growth in smartphones and mobile devices, sophisticated recognition systems were developed in the year 2000.

The recent development of databases for printed as well as handwritten text and availability of high-speed graphical processing units allowed to apply deep learning algorithms for character recognition. The latest recognition system report high recognition rates for cursive and non-cursive scripts.

1.2 Classification of OCR Systems

OCR systems are categorized with respect to image acquisition (online [8] or offline [9]), type of text (printed or handwritten) and types of scripts (independent or dependent). Each of these categories is discussed in the following.

- **Online vs. Offline OCRs:** Offline OCRs [9] is used to recognize text in scanned or camera captured images. Online recognition systems [8], on the other hand, take input through digitizing devices. Offline recognition is applicable for both printed as well as handwritten text while Online recognition is only meaningful for handwritten text. Offline recognition is performed to learn the final shape of printed or handwritten characters whereas online recognition is based on dynamic information of characters (strokes, angles, etc).
- **Handwritten vs. Printed Text OCRs:** As per type of input text, recognition systems can be divided into printed [10, 11] and handwritten [12, 13] OCRs. Printed text is mostly associated with digitized images of books, magazines, and other documents. The handwritten text includes lecture notes, manuscripts, and dictations. Most of the time handwriting

recognition can be performed using digital tablets called online recognition. In offline mode, handwriting recognition is more difficult task as compared to printed text because the printed text contains same font size, same printing style within a document while handwritten document contains variations in writing style and variable spacing between characters. In some cases the document may contain a mix of printed and handwritten text therefore, a segmentation step is added to discriminate the handwritten and printed text. Each type of text can be separately recognized according to their recognition systems.

- **Single vs. Multi-font OCRs:** Initial recognition systems were applicable to recognize a text only in single font [14, 15]. The significant efforts of researchers made recognition systems mature in order to recognize text in multiple fonts [16]. Multi-font recognition systems are applicable only for printed text.
- **Cursive vs. Non-cursive Scripts:** In non-cursive scripts, Latin, for example, [2] characters are printed and written in an isolated form which offers less complexity as compared to cursive scripts. Characters in cursive scripts [16, 17] (Arabic, Persian, Pashto, and Urdu) contain complex shapes which are challenging to recognize. In cursive scripts, ligature or partial words are considered as a recognition unit.

1.3 Background

The problem of Urdu OCR attained significant research attention during the last ten years. In this domain initial research studies were carried out on the recognition of isolated characters [18–20] that reported significant recognition results but far away from practical situations. Later, a number of robust recognition solutions were reported [11, 21–24] where recognition techniques employed ligatures or characters as a recognition units. Ligature based techniques require a huge number of ligature classes for recognition and character-based techniques need to address the challenging task of segmentation of cursive Urdu text into characters.

1.4 Problem Statement

The aim of this study is to develop a robust technique for offline handwriting recognition of cursive scripts including dots and diacritics using Urdu as a case study.

1.5 Research Objectives

The objectives of the current study are discussed in the following.

- To propose robust recognition techniques for offline handwriting recognition of cursive scripts (Urdu as a case study).
- To recognize both primary and secondary ligature of Urdu text.

- To investigate the effectiveness of machine-learned features in characterizing Urdu text.
- To evaluate the proposed techniques on custom developed as well as publicly available benchmark datasets.

1.6 General Steps in OCR

Optical character recognition systems consist of a number of steps. Initially acquired image is preprocessed where different preprocessing methods are employed. Then segmentation is carried out to segmented out the characters from text lines. Later, handcrafted or machine learned features are extracted for the classification process. Finally, few post-processing techniques are carried out to enhance the recognition rate. The general steps of OCR are presented in Figure 1.4.

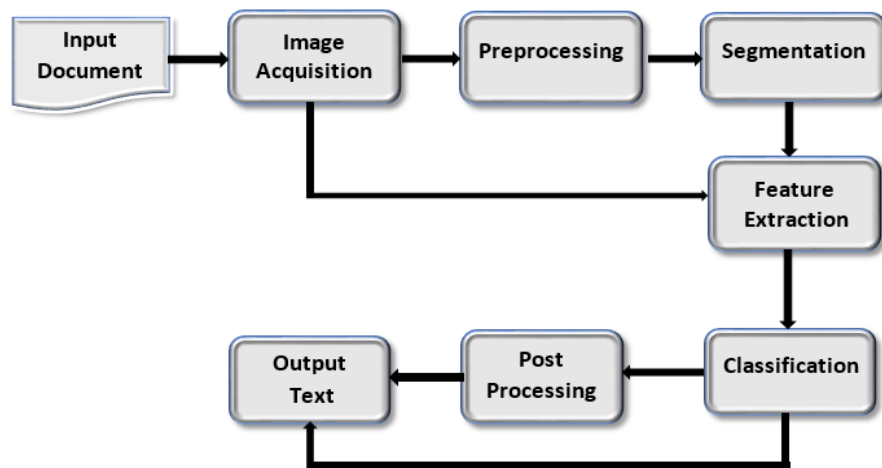


Figure 1.4: General steps in an (offline) OCR system

1.6.1 Image Acquisition

The digitized version of the document is obtained in an image acquisition process. In offline recognition systems, document images are scanned using a high-resolution scanner or captured from the camera. In online recognition, writing images or other attributes (stroke, angle) are acquired through digital devices. This research focused on offline recognition therefore remaining steps are discussed in the context of offline recognition systems.

1.6.2 Image Preprocessing

Image preprocessing is a critical step in offline recognition systems. The aim of the preprocessing step is to eliminate unwanted information from images to enhance the recognition rate. The type of preprocessing is carried out according to the input image. In general, preprocessing steps

include image binarization, skew detection, and noise removal. In case of handwriting recognition slant removal, skeletonization and baseline removal, etc are also carried out to remove the writer dependent variations.

1.6.3 Segmentation

It is a process in which the digital image is divided into different segments. From the viewpoint of text recognition, image is segmented into text lines, words, subwords or characters. In non-cursive scripts, words or characters can be easily segmented out from text lines but it is much difficult and challenging to find the boundaries of characters/ words in cursive scripts. The recognition techniques that consider words or subwords as a recognition unit are known as holistic techniques while those based on characters are known as an Analytical technique. There are two subtypes of Analytical technique i.e implicit segmentation and explicit segmentation. In an explicit segmentation based methods text is either divided into characters/sub-words or isolated characters are assumed for recognition. For cursive scripts, explicit segmentation is considered a complex problem. In implicit segmentation based techniques segmentation and classification is achieved at the same time by using machine or deep learning classifiers. The OCR techniques from the viewpoint of segmentation are presented in Figure 1.5.

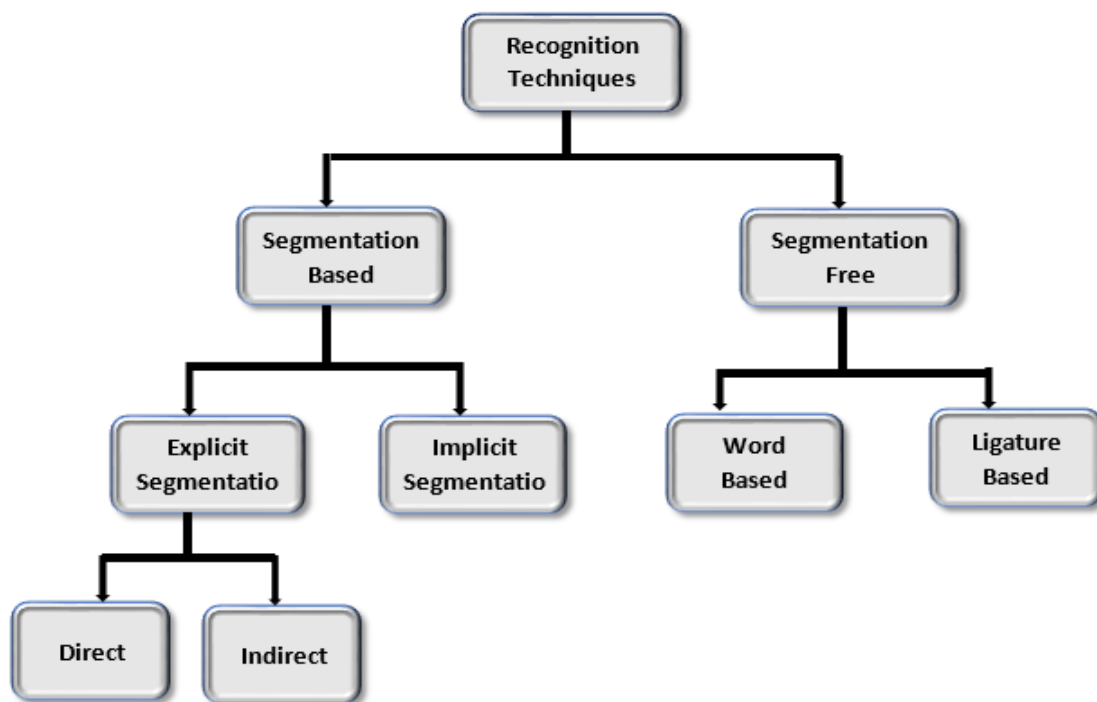


Figure 1.5: OCR techniques from the view point of segmentation

1.6.4 Feature Extraction

After segmentation of an input image into characters/sub-words or words features are extracted. Typically, statistical or structural features are computed from the images of words and subwords. The statistical features represent the statistical values that can compute from the images while structural features based on topological attributes of the object. From the viewpoint of recognition, structural features are considered more suitable. Recent studies in recognition employed a CNN for feature extraction. Major limitations in these techniques are the availability of large database and GPUs.

1.6.5 Classification (Recognition)

In the classification step, the classifier is trained by using features, extracted through the handcrafted approach or machine learning approach in order to discriminate between character/sub-words, words classes. The initial studies employed simple classifier like nearest neighbour [25], Hidden Markov Models (HMMs) [14, 26, 27] and Support vector machine(SVM) [28] for recognition task. In some studies, feature extraction and classification is achieved at the same time by using Convolutional or Recurrent Neural Networks (RNNs) [11, 29, 30].

1.6.6 Post Recognition Processing

In post-processing step recognized units are grouped (for example characters or subwords into words) and validated through the dictionary. These steps are helpful to increase the overall performance of recognition systems.

1.7 Proposed Technique

In this study, We employed a hybrid network of CNN and RNN for character recognition. We extracted features from the Urdu text line image by using CNN and these features were fed to BLSTM for classification. RNN is a very popular model that has been successfully employed for the recognition of sequential data. The RNN typically contains input, hidden and output layer where neurons of each layer are interconnected in such a way provides help to trace back the previous computation [31]. Unfortunately, the standard RNN architecture cannot handle the sequential data with more time stamp delays and faces a problem called vanishing gradient. The LSTM is used to tackle this problem. The LSTM architecture consists of memory cells, an input gate, the output gate and forget gate. In text recognition sequence prediction in both direction are useful for correct transcription of the text. Therefor Bi-directional LSTM is created by combining the forward and backward LSTM. The proposed technique is evaluated on custom generated data set as well as the publically available standard benchmark. A workflow of the proposed methodology is illustrated in Figure 1.6.

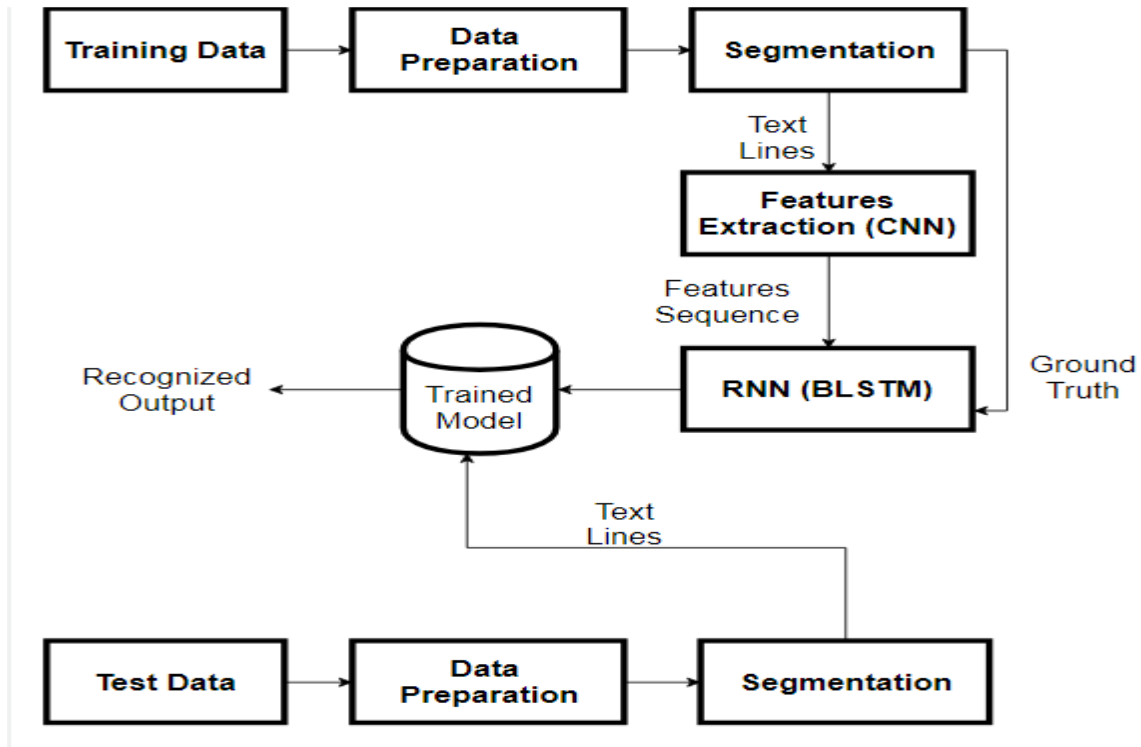


Figure 1.6: A Workflow of proposed Methodology

1.8 Thesis Contribution

The key contributions of the present study are listed in the following.

- A novel, implicit segmentation based recognition system for handwritten Nastaliq Urdu is proposed.
- Recognition is carried out through machine-learned features (using CNN) that are fed to RNN for sequence modeling.
- A custom data set is generated along with transcriptions.
- The proposed technique is validated on a custom generated data set and on publicly available offline handwritten data set (UNHD).

1.9 Thesis Organization

The thesis is organized as follows. Chapter 2 provides an overview of Urdu text and recognition challenges associated with Urdu recognition. Chapter 3 introduces the data set preparation and proposed methodology of the present study. Chapter 4 presents a comprehensive discussion of recognition results. Conclusion and perspectives are presented in chapter 5. The research publication is listed in chapter 6.

Chapter 2

Literature Review

The history of recognition systems is spanned over many decades. Due to significant efforts of research endeavors, now mature recognition systems are available for many languages. Among these, the recognition systems for non-cursive languages are considered very mature. Cursive languages are highly challenging to recognize. A number of recent studies on cursive language like Arabic reported high recognition rate on printed as well as handwritten text [32–34]. Regardless of these developments, recognition systems of many cursive languages could not gain the attention of the researcher due to lack of standard dataset and the complexity of writing styles. The Urdu language is one of them. The focus of this study is on handwriting recognition of Urdu text. An overview of Urdu, recognition challenges, handwriting recognition techniques and well-known datasets of printed as well as handwritten is discussed in subsequent sections.

2.1 Overview of Urdu

The word Urdu is derived from the Turkic word 'Ordu' meaning Army [35]. The characters of Urdu language are derived from Arabic, Sanskrit, and Persian. There are more than 100 million native Urdu speakers with a major share of Pakistan, India and the Middle East. The salient characteristics of Urdu language including characters, numerals and writing style are discussed in later sections.

2.1.1 Urdu Alphabet and Numerals

Urdu language consists of 39 isolated characters (Figure 2.1) and 10 numerals (Figure 2.2). Character shape depends on their position in ligatures because characters can appear in the start, middle, or last. Urdu is a bidirectional script because ligatures and numbers are written in two different directions.

2.1.2 Writing Styles

The writing style is the way to represent a word, sentence or paragraph. There are different writing styles used for cursive scripts e.g Nastalique, Kofi, and Naskh. Arabic, Pashto, and Persian are

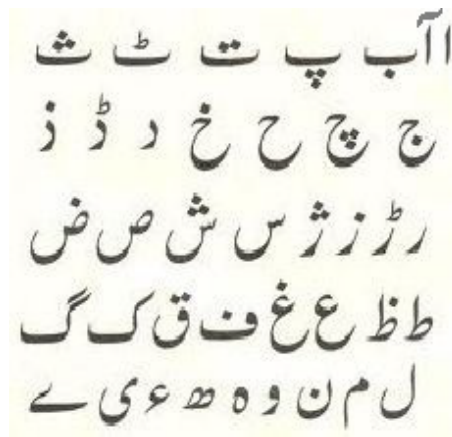


Figure 2.1: Urdu characters written right-to-left

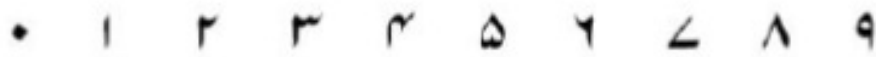


Figure 2.2: Urdu numbers written left-to-right

mostly printed and written in Naskh style while Urdu uses Nastalique. From the viewpoint of recognition, Nastalique style offers more complexity and challenges as compared to Naskh. Popular writing styles of cursive scripts are presented in Figure 2.3. Recognition challenges of Urdu handwritten text are presented in the next section.

و سخر الشمس والقمر	نستعلیق
وسخر الشمس والقمر	کوفی
و سخر الشمس والقمر	ثلث
و سخر الشمس والقمر	دیوانی
وسخر الشمس والقمر	رقاع
و سخر الشمس والقمر	نسخ

Figure 2.3: Popular writing styles for cursive scripts (Image Source: [36])

2.1.3 Recognition Challenges of Urdu Text

This section presents the various challenges of Urdu Nastalique text. The major challenges include text cursiveness, ligature overlapping, the variable spacing between words, no fixed baseline and bidirectional text.

2.1.3.1 Text Cursiveness

The major cause of cursiveness in Urdu text is calligraphic nature of Nastalique font. In Urdu language, characters are joined with other characters to form a ligature. A ligature may consist of two sub-components, primary ligature (main body) and the secondary ligature (dots and diacritic) as presented in Figure 2.4. Two types of characters exist in ligatures i.e joiner and non-joiner characters as presented in Figure 2.5. The joiner characters are the major source of cursiveness in text lines. The explicit segmentation of Urdu text lines into characters, words, and ligature is a highly challenging task therefore, ligatures are mostly considered as a recognition unit in holistic techniques.

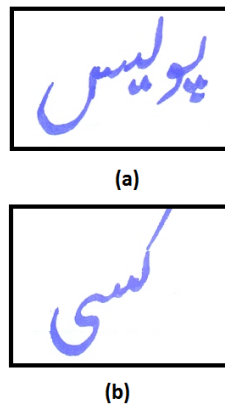


Figure 2.4: A complete ligature (a) With primary and secondary ligatures (b) Without secondary ligatures

2.1.3.2 Shape Variation

In Urdu text the joiner characters may appear into two to four different shapes i.e isolated, initial, middle, end according to their position in ligature. The shape of character also varies according to preceding and following joining character. The variations in character shape offer more complexity in recognition. An example of Urdu character "sheen" with possible shapes is presented in Figure 2.6.

ص	ش	س	خ	ح	چ	ج	ث	ط	ت	پ	ب
12	11	10	9	8	7	6	5	4	3	2	1
ن	م	ل	گ	ک	ق	ف	غ	ع	ظ	ط	ض
24	23	22	21	20	19	18	17	16	15	14	13
									ی	ھ	ہ
									27	26	25

(a)

		ے	و	ژ	ز	ڑ	ر	ذ	ڈ	د	ا
12	11	10	9	8	7	6	5	4	3	2	1

(b)

Figure 2.5: Urdu character set (a) Joiners and (b) Non-joiners





			
End	Middle	Initial	Isolated

Figure 2.6: Urdu character "Sheen" with possible shapes

2.1.3.3 Bidirectional Text

Urdu text lines may comprise characters and numerals together, where Urdu characters start from the right side of the page and numbers are from the left side. A text line with bidirectional behavior is presented in Figure 2.7

2.1.3.4 Text Overlapping

Urdu text lines consist of multiple numbers of ligature. The characters of these ligatures may overlap with other characters within ligature as well as with neighboring ligatures. The overlapping of characters creates segmentation issues (Figure 2.8).

Figure 2.7: Bidirectional behaviour of Urdu text.

Figure 2.8: Text Overlapping

2.1.3.5 Variable spacing

Text lines of the cursive script may contain variable spaces between words and ligatures. These variable spaces between words are the main reason due to which segmentation of Urdu text line is not possible and resultant ligatures or partial words are employed as a recognition unit. An example of variable spacing in Urdu text is presented in Figure 2.9.

Figure 2.9: Variable Spacing

After discussion of recognition challenges, few databases of Urdu text for printed and handwritten text are discussed in the following section.

2.2 Datasets

Availability of labeled datasets is the fundamental requirement in the development of recognition systems. Due to the significant efforts of document recognition community, a number of datasets have been developed and made available for the evaluation of printed and handwritten Urdu text. The databases have been developed at character/digit [37], word [38], sentence [25] and paragraph levels [39, 40]. A detailed discussion of well-known Urdu databases for printed and handwritten text is presented below.

2.2.1 Printed Text Data Sets

2.2.1.1 Urdu Printed Text Images Database (UPTI)

UPTI [25] is a familiar database of printed Urdu text that has been utilized for the evaluation of the recognition systems. The database contains 10,063 lines of text generated in Nastalique font and

also contains three versions of images (line level, ligature level and degraded noisy). A sample text line from UPTI data set is presented in Figure 2.10

پاکستان ترقی کی راہ پر گامزن ہے۔

Figure 2.10: Sample text line from UPTI dataset

2.2.1.2 Center of Language Engineering (CLE) Urdu Database

The CLE database [41] is another familiar database of printed Urdu text which is more comprehensive and more difficult as compared to UPTI. The CLE database contains 2,017 HFL cluster and each one is comprises 35 instances (Figure 2.11). To generate a cluster, ligatures were typed on computer software, printed on a page and then scanned. The second set contains scanned images of printed Urdu books which are more challenging to recognize as compared to HFL images.

ترقی ترقی ترقی ترقی ترقی
 ترقی ترقی ترقی ترقی ترقی
 ترقی ترقی ترقی ترقی ترقی
 ترقی ترقی ترقی ترقی ترقی
 ترقی ترقی ترقی ترقی ترقی
 ترقی ترقی ترقی ترقی ترقی
 ترقی ترقی ترقی ترقی ترقی

Figure 2.11: Instances in an example cluster from CLE dataset

2.2.2 Handwritten Text Data sets

2.2.2.1 CENPARMI Urdu Database

They generated a database for Urdu handwritten words, characters, digits, and numerical strings. The database consists of 318 date samples, 60,329 isolated digits, 12,914 strings of numerals, 1,705 occurrences of 4 special symbols, 14,890 samples of 37 basic characters and 19,432 examples of 57 words related to finance[39]. The data is collected from 343 different native Urdu speakers around the world.

2.2.2.2 UCOM Offline Handwritten Data set

UCOM offline handwritten dataset [42] consists of 600 filled forms with 6,200 handwritten words. Data is collected from 100 native Urdu writers where each writer contributed 6 pages and each page contains 8 text lines. The labels of the dataset include a transcription of text (UTF-8 encoding) that supports the evaluation of recognition systems. The number of unique text lines in the database, however, is very small (only 48). A sample image of UCOM offline data is presented in Figure 2.12 and the statistics table of UCOM data set is presented in Figure 2.13. The UCOM data set was later extended to UNHD.

Detail	Statistics
No. of text lines per page	8 Text lines
No. of words written by writer	620
No. of characters written by writer	6656
Total number of characters	53248
Total number of words	62000
Total text lines	6400
Total number of writer	100

Figure 2.12: Statistics of UCOM-Data set

Figure 2.13: A sample Text line from UCOM-Offline data set

2.2.2.3 UNHD Offline Handwritten Data set

UNHD offline handwritten dataset [43] is a relatively recent Urdu handwritten dataset. The dataset consists of 10,000 text lines and 31,200 handwritten characters. The data is collected from 500 native Urdu writers where each writer contributed 5 to 8 lines per page. The UNHD data set contains only 700 unique text lines. We acquired the UNHD database, however, it was observed that the number of samples provided was fewer than what was claimed (only 4500 text lines are available). Furthermore, since the number of unique text lines in this database is only 700, it does not truly match the real world recognition scenarios. Consequently, for a robust evaluation of the proposed technique, we generated our custom data set which is explained in the next chapter. The complete statistics table of UNHD data set is presented in Figure 2.14.

Detail	Statistics
Total number of writers	500
No. of text lines per page	5 to 8 text lines
Total no. of text lines	10000
No. of words written by each writer	624
Total number of words	312000
Total number of characters	187200

Figure 2.14: Statistics of UNHD-Data set

In the next section, we present recognition techniques for Urdu text.

2.3 Recognition Techniques for Urdu Text

In Urdu and other cursive scripts, text can be recognized through two different approaches i.e., analytical & holistic. The analytical techniques are based on the segmentation of text into characters while the holistic techniques are based on a larger unit of text (words/sub-words). Initial research studies for recognition of Urdu text mainly focused on isolated characters with the assumption that text is already segmented [17]. Few studies have also focused on the recognition of online Urdu handwritten text [44]. The reported recognition rates of isolated and pre-segmented characters are much higher but far from practical situations. The details of analytical and holistic techniques are presented in the following.

2.3.1 Analytical Approaches

In analytical techniques, characters are considered as recognition units. Such techniques deal with a smaller number of unique classes for characters and their context-dependent shapes. The key challenge of this approach is the segmentation of text into characters. The detail of analytical approaches is presented in subsequent sections.

2.3.1.1 Explicit Segmentation Based Techniques

In these techniques, the text is either divided into characters or isolated characters are assumed for recognition. One of the early studies is for Urdu text recognition conducted by Pal & Sarkar [18] proposed a recognition system for isolated characters. The document image was digitized through a flatbed scanner and segmented into text lines by horizontal projection method. Topological, contour and water reservoirs features were obtained for classification. Later, Decision Tree was used for character recognition. The system performance was evaluated on 3,050 isolated characters and reported 99.8% accuracy.

In another study, Hussain et al. [45] proposed segmented characters based recognition system. The system takes an image of the segmented character and passes through pre-processing steps to make it more valid for extraction of features and recognition. Initially, 104 segmented characters were classified into 33 different clusters using Kohonen self-organizing map (SOM). In the second step, additional features (height, width, loop, cross, and curve) were extracted from each character for final classification. The system reported an 88% recognition rate. In the cursive script, text segmentation is considered a challenging task. Several techniques have been proposed for text segmentation only. Among these techniques, authors in [46] proposed a structural feature-based character segmentation technique for printed Urdu text. To detect the text line, the image was scanned horizontally and vertically in search of text pixels. The FCC of the ligature was calculated to determine whether the pixels belong to the primary or secondary ligature. Later, the text

Table 2.1: Summary of explicit segmentation based techniques

Paper	Technique	Data-Base	Experiment	Result
Pal & Sarkar [18]	Structural features with decision Trees	Private	3050 isolated character	98.7 %
Hussain et al [45]	Topological features with SOM	Private	104 segmented characters	88 %
Ahmed et al [17]	Artificial Neural Network	Private	56 character classes	93.40 %
Akram et al [47]	DCT with HMMs	CLE	224 Document images	86.15 %

image was segmented into characters on the basis of primary ligature. Over-segmentation and under-segmentation issues were also addressed for enhancement of segmentation steps. System performance was tested on images of Batool font and achieved 99.4 % segmentation accuracy.

Ahmed et al. [17] proposed a novel recognition technique. Few pre-processing steps were carried out to eliminate the overlapping issues between cursive characters. Later, neural network was employed for classification of Segmented characters and realized 93.40% recognition rate on a dataset comprising scanned and computer-generated images. The system only performs segmentation and recognition when the script is diacritic free and has a fixed size. The same study was later extended [14] where system consisted of two phases i.e. segmentation and classification. Segmented characters were used to train a neural network on 56 different classes and reported 70% classification accuracy.

In another comprehensive study [47] authors proposed a recognition technique for Urdu Nastalique text based on explicit segmentation. The HMM was used as a classifier. A novel sliding window based segmentation technique was used along thinned images of the primary ligatures to extract the character boundaries. The system reported 97.11% recognition accuracy when tested on 79,093 instances of 5,249 main body ligature classes and reported 87.44% recognition rate.

The explicit segmentation based recognition techniques are employed less often due to the involvement of the segmentation step. Furthermore, the high rate of errors in the segmentation phase also affects the overall performance of the system. Implicit segmentation based techniques are considered more suitable and well known for recognition task due to less complexity and high accuracy. An overview of these techniques is presented in Table 2.1

2.3.1.2 Implicit Segmentation Based Techniques

Generally, in these techniques segmentation and classification are achieved at the same time by using machine learning algorithms. These techniques are known to be robust and report low error

Table 2.2: Summary of implicit segmentation based techniques

Paper	Technique	Data-Base	Experiment	Result
Hussain et al. [48]	Raw pixels with BLSTM	UPTI	Training : 4000 lines Testing:2000	94.5 %
Ahmed et al. [15]	Raw pixels with BLSTM	UPTI	Training:12415 lines Testing: 2836 lines	96 %
Naz et al. [30]	Statistical features with MD-RNN	UPTI	Training: 6800 lines Testing:1600 lines	94.97 %
Naz et al. [30]	CNN with MD-RNN	UPTI	Training: 6800 lines Testing:1600 lines	98.12 %

rates. On the other hand, these techniques require large training data and more computational power. These techniques, in general, employ different variants of Recurrent Neural Networks(RNNs). Among one of the earlier work in this area, Hassan et al. [48] proposed a recognition system for offline printed Urdu Nastaleeq script. In this system, BLSTM model of RNN is used for text classification. BLSTM network was evaluated for two cases. In the first case, network performance was measured without considering the character shape variations while they were considered in the second case. Experiments conducted on UPTI dataset reported 94.85 % recognition rate without character shape variations and 86.43% with character shape variations. In another similar work, Naz et al. [11] proposed a statistical feature based recognition system for printed Urdu Nastalique script. In this system, statistical features were extracted by overlapping sliding windows on Urdu text lines. The extracted features of text lines were fed to a MD-RNN for final classification. The system performance was evaluated on UPTI data set that reported 94.97 % recognition accuracy. The same study was later extended by Naz et al. [30] to propose a machine learning based approach for feature extraction, where CNN was employed for feature extraction and MD-RNN for classification. This combination of CNN and MD-RNN enhanced the classification rate up to 98.12%. An overview of these techniques is presented in Table 2.2

2.3.2 Holistic Approaches

Holistic or segmentation free techniques overcome the complexity of the segmentation process in recognition problems. These techniques employ ligatures or sub-words as a recognition unit therefor, segmentation is not required. However, a huge number of ligature classes are required. The number of ligature classes can be reduced by splitting the ligature into primary and secondary components. Among significant holistic techniques, Sardar and Wahab [49] proposed a recognition system for offline and online Urdu characters. Offline images were converted into binary images in a pre-processing step and text lines were extracted from binary images through horizontal projection method. As features, Hu's invariant moments of each ligature were calculated and matched with stored features to recognize the ligature. K-Nearest neighbor was employed for recognition purpose.

Table 2.3: Summary of holistic segmentation based techniques

Paper	Technique	Data-Base	Experiment	Result
Sardar and wa-hab. [49]	Hu's moments with KNN	Private	1050 ligatures	97.12 %
Ahmed et al. [10]	Autoencoder	UPTI	3732 ligatures	96 %
Ahmed et al. [29]	Raw pixel with LSTM	UPTI	3604 ligatures	96.71 %

The system was evaluated on 1,050 ligatures and achieved 97.12% recognition accuracy.

Among one of the recent studies, Ahmed et al. [10] proposed a stacked denoising autoencoder based recognition system for Nastalique Urdu script. Different number autoencoders were trained on 178,573 ligatures with 3,732 classes from UPTI data set. The experimental study reported 96% recognition accuracy. The same work was later extended in [29] to propose Gated BLSTM (GBLSTM) for Urdu character recognition using raw pixel values. The system was evaluated on UPTI dataset that reported 96.71% recognition accuracy. An overview of holistic techniques is presented in Table 2.3.

While notable work is reported for printed Urdu text, limited work has been carried out on Urdu handwriting recognition due to its cursive nature. Sagheer et al. [50] proposed a system for recognition of handwritten, pre-segmented characters. In the pre-processing phase, handwritten images were converted into grayscale binary images. Two types of feature sets were computed from each image i.e. gradient and structural. The Gradient features were computed from grayscale images and structural features were computed from binary images. Support vector machine (SVM) was employed for classification purpose. The performance of the system was evaluated on the CENPARMI Urdu word database and achieved 97% recognition accuracy.

In another study, Saad et al. [42] proposed a handwriting recognition system based on RNN. Handwritten images were segmented into text lines using their ground truth in a pre-processing phase. Fixed size sliding windows were traversed over text lines to extract pixel values as features, which were fed to RNN for learning. The system was evaluated on UCOM data set and reported high recognition accuracy. The same study was later extended by the same authors [43] to evaluate the system on UNHD data set. A summary of Urdu handwriting recognition systems is presented in Table 2.4

2.4 Summary

In this chapter, we presented an overview of Urdu language, recognition challenges and databases of printed and handwritten text developed by the relevant research community. The two different recognition techniques i.e analytical and holistic are also discussed. A number of studies reported

Table 2.4: Summary of Urdu handwriting recognition systems

Paper	Technique	Data-Base	Segmentation	Experiment	Result
Sagheer et al. [50]	Gradient and structural features with SVM	CENPARMI	Holistic	1817 Hand-written words	97 %
Saad et al. [42]	RNN	UCOM	Implicit	Training:50 text lines Test:20 text lines	94%
Saad et al. [43]	RNN	UNHD	Implicit	Training:50% Validation:30% Test:20%	92.7%

UPTI have been widely employed database for printed Urdu text recognition. As handwriting recognition is relatively less explored area therefore, UNHD is the only standard benchmark for handwriting recognition.

In the next chapter generated the dataset, model training and model evaluation is presented.

Chapter 3

Methods

This chapter presents the detail of proposed methodology carried out for handwritten character recognition. We employed CNN for feature extraction and BLSTM model of RNN for feature classification. We first present the detail of our custom developed data set. We then discuss each phase of the proposed methodology in detail i.e data preprocessing, feature extraction and classification.

3.1 Dataset

As discussed earlier, for printed Urdu text two benchmark datasets have been employed by researchers, the UPTI database [25] and the CLE database [41]. For Urdu handwriting text recently, Ahmed et al. [42] introduced a UCOM offline dataset. The data set was acquired from 100 Urdu writers. The dataset contains only 48 unique text lines. Later UCOM was extended into UNHD where 500 writers wrote only 700 unique text lines. The UNHD dataset contains 10,000 text lines in total. We acquired the UNHD database, however, it was observed that the number of samples provided was fewer than what was claimed (only 4500 text lines are available). Furthermore, since the number of unique text lines in this database is only 700, it does not truly match the real world recognition scenarios. Consequently, for a robust evaluation of the proposed technique, we collected 6,000 unique text lines from 600 writers as a part of this study. The detail of the developed dataset is presented in a given section.

3.1.1 Urdu Handwritten Text Images Database (UHTI)

UHTI handwritten data set is the recent contribution in Urdu handwritten standard benchmark. The data set contains 6000 unique text lines which are collected from 600 Urdu native writers including male and female. In order to have more variations in writing samples, we obtained data from school, college and universities students as well as from professionals and housewives. Each individual was asked to write in a natural way, where each writer wrote 10 text lines. The text lines contain ligatures and each ligature comprised of five to six characters. The acquired data set can be used for different research scenarios such as character recognition and for writer identification. We tried to

cover all characters and ligatures with their possible variations according to their position in Urdu text. There is a plan to increase the data set up to 2500 writers later with more variations. A sample image of UHTI is illustrated in Figure 3.1 and the statistics table is mentioned in Figure 3.2 .

رہے ہیں جو لاپتا افراد کا کیس سن رہا ہے۔
 اور یہ بہت حساس نوعیت کا کیس ہے۔ قانونی حلقوں
 کے مطابق تفتیشی ایجنسیاں تحقیقات کے دوران اس پہلو پر
 بھی خصوصی توجہ دیں گی کہ کہیں اس قتل کا
 تعلق لاپتا افراد کے کیس سے تو نہیں ہے۔

Figure 3.1: A sample Image from UHTI offline data-set

Detail	Statistics
Total number of writers	600
No. of text lines per page	10
Total no. of text lines	6000
No. of text lines written by each writer	10
Total number of words written by each individual	144
Total number of words	86400
Total number of characters	4320000

Figure 3.2: Statistics of UHTI- Dataset

3.2 Proposed Methodology

Thanks to the recent advancements in deep neural networks, the last few years have resulted in a paradigm shift from traditional classification pipeline (involving pre-processing, feature extraction and classification) to end-to-end trainable systems [51]. Hand-engineered features are being replaced with data-driven machine-learned features. These developments have had a significant impact on document and handwriting recognition community as well. CNN(s) are known to be state-of-the-art feature extractors while RNNs (and their variants) have been effectively employed to sequence modeling problems. Handwriting represents a sequence of strokes that needs to be mapped to the corresponding transcription. Recurrent nets, hence, offer an attractive choice for handwriting recognition. The effectiveness of recurrent nets has been validated on printed Urdu text where windows sliding over lines of text are employed to feed pixel values [11] or statistical features [30] to the network to learn character shapes and segmentation points. A similar technique [43] was also applied to handwriting images where raw pixel values from columns of text line images are fed to a 1D-LSTM for learning and classification. A step further to this is to replace the raw pixel values (or hand-engineered features) by machine learned features extracted using CNN(s). This combination of CNN and LSTM has been previously investigated where the features extracted by CNN(s) are fed to the recurrent layers for classification [51] and has proved to be an effective solution for recognition tasks. In our study, we adopt the same combination (CNN+LSTM) for recognition of Urdu handwriting. An overview of the proposed recognition technique is presented in Figure 3.3.

Real data is often inconsistent, incomplete and noisy therefor, few preprocessing steps are involved to resolve such issues. From the viewpoint of recognition systems typically preprocessing steps included image binarization, grayscale conversion, skew and slant correction. The detail discussion of the preprocessing step carried out on input handwriting image is discussed in the given section.

3.2.1 Data Preprocessing

The data set contains scanned images in different colors therefor grayscale conversion is applied as a first step of preprocessing. To separate out the foreground and background region from the grayscale image, binarization is carried out. Image binarization is segmentation of pixels into foreground and background using thresholding methods. There are two types of thresholding methods are used i.e. local and global thresholding. In global thresholding, a single threshold value is used for the complete image while in local thresholding a separate threshold value is used for different regions of the image. The local thresholding is suitable for images which suffer the problem of degradation and non-uniform illuminations. In the proposed study, Ostu's binarization algorithm [52] is carried out for segmentation of text and background. Grayscale and binarized image is illustrated in Figure 3.4. Our proposed study is based on character recognition therefor after binarization text segmentation is performed.

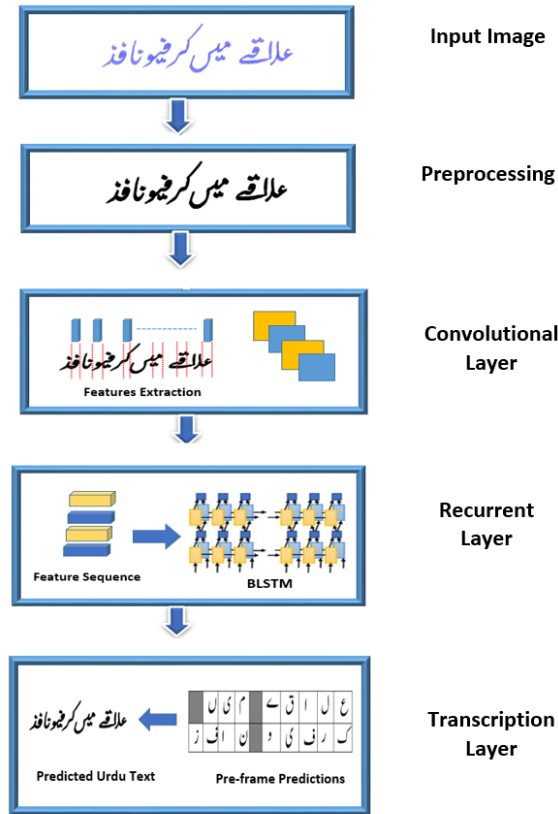


Figure 3.3: An overview of the recognition system

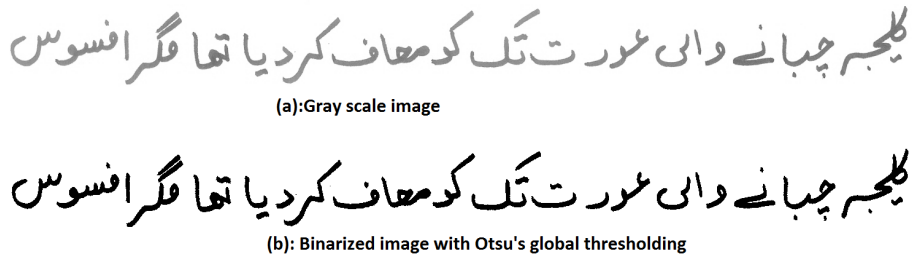


Figure 3.4: (a) Grayscale image (b) Binarized image with Otsu's global thresholding

The height of segmented text lines is normalized and fed to Convolutional layers for feature extractions. The detail discussion of the feature extraction process is illustrated in the next section.

3.2.2 Feature Extraction

A key component for classification problems where meaningful and robust features are extracted from an input image. The classifier used these sets of features to perform classification. This process varies according to the nature of the problem because the feature set of one problem might fail in another problem. The structural features are the most common type of features used for text recognition. Some common features for our Urdu text are a number of dots and diacritics, weight

of stroke, shape of a false and filled loop, etc. Statistical features based on statistical quantities of input pixels. This traditional feature extraction approach is often expensive. The domain expert is required to identify the most applied and meaningful features from input data. On the other hand, the machine learned features eliminates the need for domain expert and hardcore feature extraction process. Recently, the machine learned features are extracted by using deep learning algorithms. These algorithms are used to learn high-level features from input data in an incremental manner. In the proposed technique we employed CNN for feature extraction from segmented text lines. The detail of CNN is illustrated in the given section.

3.2.2.1 Overview of CNNs

CNN belongs to a class of deep learning algorithms which is mostly used for image analysis. CNN was first time introduced in 1990s [53], but due to nonavailability of high-performance machines and large datasets, it could not gain the attention of the research community. In recent years, the development of large dataset particularly imageNet [54] and remarkable improvements in hardware technology enhanced the performance of CNN on learning tasks. As compared to traditional approaches, CNN based methods provide remarkable results and also reduce error rates. CNN(s) have been widely used in character recognition [55], object detection [56] and face recognition [57] problems. The traditional neural network takes input in the form of a single vector where fully connected architecture requires a large number of weights per neuron. Such types of networks are not suitable for images. CNN(s) takes images as an input where each neuron is connected to the local region of the image and hence lesser number of weights are associated with each neuron. The neurons have three dimensions. i.e height, width, and depth. The architecture of CNN comprises a convolutional layer, pooling layer, Relu layer, and fully connected layer. Each layer is discussed in the following sections. The architecture of CNN is presented in Figure3.5.

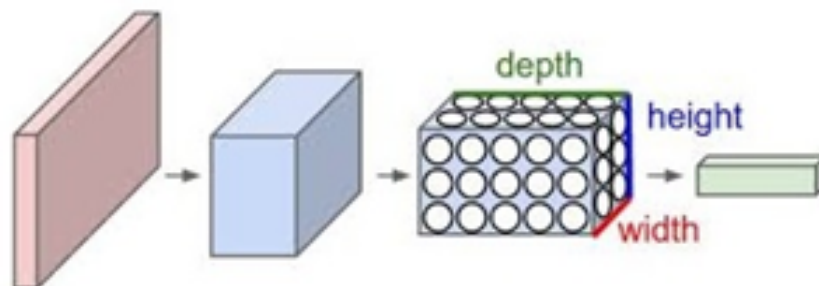


Figure 3.5: The architecture of Convolutional Neural Network

Convolutional Layer The convolutional layer contains a number of filters which are used for feature extraction. The convolutional layers extract features from the previous layer and map their appearance into a feature map. The initial layers are used to extract low-level features such as lines, curves, and dots, the subsequent layers combine these simpler features in order to extract high-level features. There are multiple types of filters can be used in convolutional layer to perform different operations such as edge detection, image, and image sharpening. Each filter in the Conv layer generates a separate feature map as presented in Figure 3.6. The feature maps of all filters are combined together against the depth of CNN for computation of output volume. In Figure 3.7, $50 \times 50 \times 3$ image is convolved with five $5 \times 5 \times 3$ filters which produce 5 activation map. The number and size of the filter at each convolutional layer may vary but the depth of filter remains the same as input volume. The size of output volume at each layer is based on three parameters i.e padding, depth, and stride. Stride represents a number of steps that filter takes after each convolution and padding refers to the number of extra rows and column added at the border of an image to accomplish convolution operation at border pixels. The parameter depth, define the number of filters in a layer. In our proposed methodology we employed 7 convolutional layers with a different number of filters and in size.

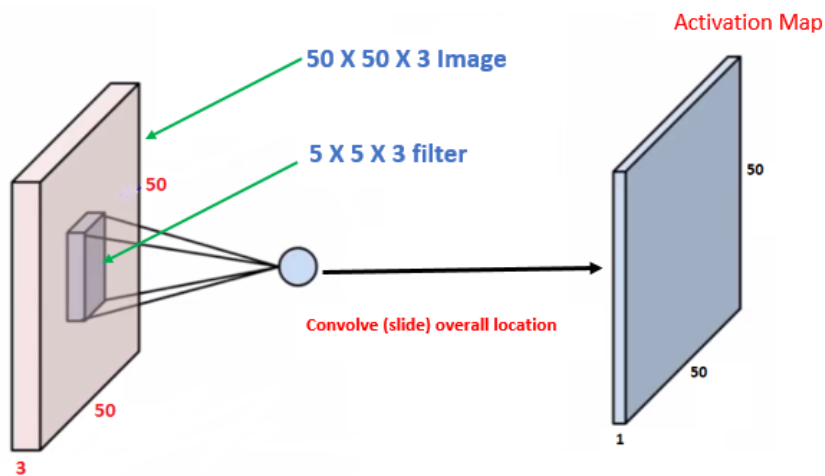


Figure 3.6: Convolution of an image($50 \times 50 \times 3$) with a filter($5 \times 5 \times 3$)

ReLU Layer It is an activation function which is applied on a feature map to add non-linearity. There are various types of nonlinear function such as tanh or sigmoid, but the main reason to employ ReLU function is, it effectively avoids the vanishing gradient problem and performs better. Mathematical notation of ReLU function is $r(x) = \max(0, x)$. In our model, we employed ReLU activation function after each convolutional layer.

Pooling Layer It is used periodically after multiple stages of other layers. The prime responsibility of this layer is to normalize the size (parameters) of the activation map. This downsampling

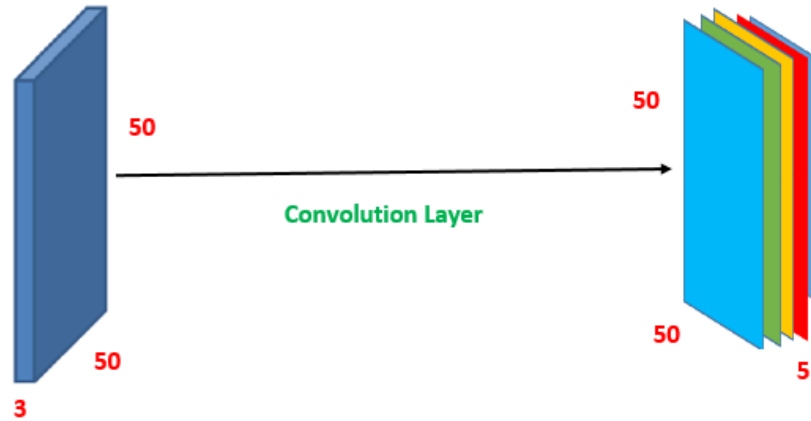


Figure 3.7: Convolution of an image with five filters to produce the output volume

operation of pooling layer controls the model from the over-fitting problem. Different types of pooling can be applied i.e Sum pooling, Max pooling, and Average pooling. Among these, Max pooling operation is most common. Typically 2 X 2 filter is used to perform max pooling. We carried out max pooling after convolutional layers. The detail of applied pooling operation is presented in Table 3.1.

Table 3.1: Summary of convolutional and Pooling layers

Layers	Filter Size	No.of Filters
CNN Layer 1	3×3	64
ReLU Layer		
Pooling Layer	Max Pooling 2 X 2	
CNN Layer 2	3×3	128
ReLU Layer		
Pooling Layer	Max Pooling 2 X 2	
CNN Layer 3	3×3	256
ReLU Layer		
CNN Layer 4	3×3	256
ReLU Layer		
Pooling Layer	Max Pooling 2 X 1	
CNN Layer 5	3×3	512
ReLU Layer		
CNN Layer 6	3×3	512
ReLU Layer		
Pooling Layer	Max Pooling 2 X 1	
CNN Layer 7	2×2	512
ReLU Layer		

Fully Connected Layer In this layer, each neuron is completely connected with every element of the previous layer. The convolutional and pooling layers generate feature maps which contain

high-level features of the input image. This layer is used to classify the input image into different classes according to training data. The number of neurons in the FC layer is equal to the number of unique classes in the problem. In our model, we performed classification using BLSTM model of RNN. Therefore the FC layer is not part of the proposed architecture.

3.3 Classification

It is a supervised learning technique which is used to assign a label or class to new data instance on the basis of training data. The classification stage requires training data along with their labels for correct mapping and this trained model is further used to assign the correct class for the new input sample. The most common classification methods are KNN, SVM, ANN, and Hidden Markov Models. In the proposed methodology, the features map extracted from convolutional layers is fed to the classification stage to assign class or label from a given set of classes. We employed BLSTM model of RNN for classification. The detail of BLSTM Architecture is illustrated in a given section.

3.3.0.1 BLSTM

The traditional neural network does not perform well where the sequence of data is important, e.g. language translation, text recognition, and sentimental analysis. To overcome this limitation, the Recurrent Neural Networks were proposed. The fundamental building block of RNN is neuron whose strength is encoded by weights. The RNN typically contains input, hidden and output layers where neurons of each layer are interconnected with each other that provides help to trace back the previous computations. RNNs perform very well on sequential data where a sequence of data is very important. The RNN cell not only considers the input of the cell but also considers the output of current cell to maintain the sequence. The representation of the RNN cell is illustrated in Figure 3.8. The simple RNNs mostly used tanh as an activation function. During backpropagation, these small numbers squeezed the final gradient and do not make any change in the weights. This problem slows down the model training (called vanishing gradient). To overcome this problem memory cells are introduced in the hidden layers of RNN and modify the vanilla RNN architecture into LSTM. The LSTM architecture consists of memory cells, an input gate, the forget gate and an output gate. The input gate of the LSTM decides when the current activation of the cell should be changed by adding new information from the input network in the present cell. Similarly, the output gate is responsible to select the useful information from the current cell and propagate to the next cell. The forget gate take the input from the previous state, resets the activation values and helps to decide what must be removed from the previous state and keep only relevant information. These gates help to reduce the vanishing gradient problem in LSTMs. A visual representation of Vanilla RNN and LSTM is illustrated in Figure 3.9. The LSTM is directional, it uses the past context which stored by forget gate. For text recognition, correct transcription can be achieved by using this context in both directions. Therefore we combine two LSTMs (forward and backward) into bidirectional LSTMs (Figure 3.10). Multiple such BLSTMs layers can be stacked together to make it deeper and to achieve more abstraction. In the proposed methodology we employed 2

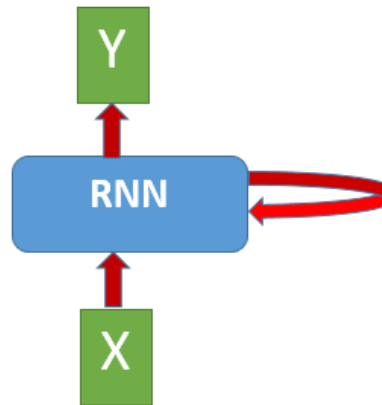
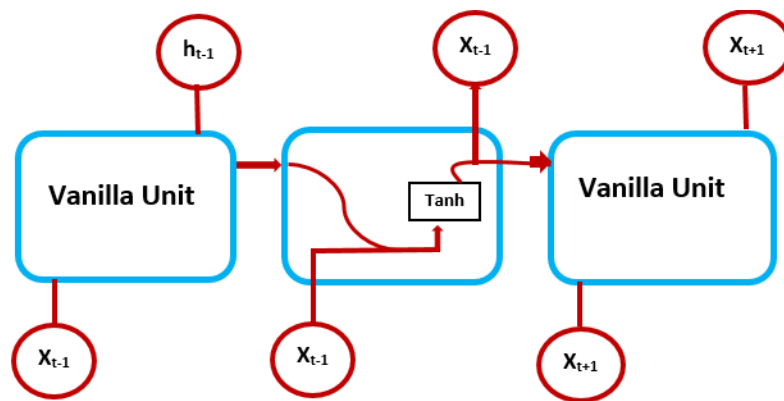
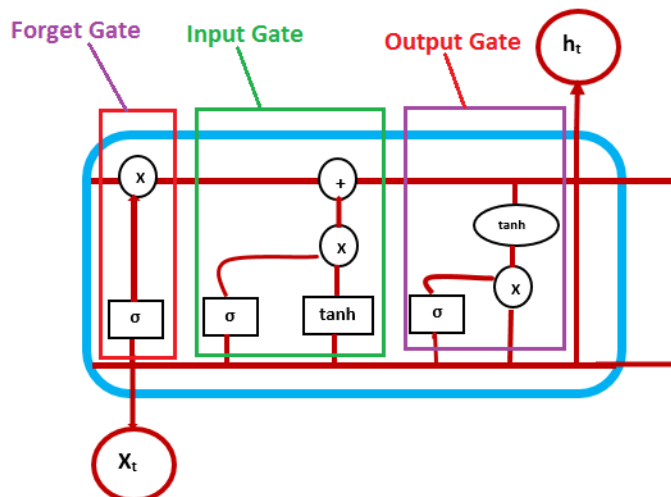


Figure 3.8: Representation of simple RNN cell



(a): A standard (Vanilla RNN Network



(b): An LSTM Unit based RNN Network

Figure 3.9: (a): A standard Vanilla RNN Network (b): An LSTM unit based RNN Network

BLSTM layers (one for forward and second for backward) where each layer contains 256 hidden cells. The output of these recurrent layers is fed to CTC layer to translate the predictions. The arrangement of BLSTM layers along convolutional layers is presented in Table 3.2

Table 3.2: Summary of convolutional and Recurrent layers

Layers	Filter Size	No.of Filters
CNN Layer 1	3×3	64
ReLU Layer		
Pooling Layer	Max Pooling 2×2	
CNN Layer 2	3×3	128
ReLU Layer		
Pooling Layer	Max Pooling 2×2	
CNN Layer 3	3×3	256
ReLU Layer		
CNN Layer 4	3×3	256
ReLU Layer		
Pooling Layer	Max Pooling 2×1	
CNN Layer 5	3×3	512
ReLU Layer		
CNN Layer 6	3×3	512
ReLU Layer		
Pooling Layer	Max Pooling 2×1	
CNN Layer 7	2×2	512
ReLU Layer		
BLSTM Layer 1	Hidden Units: 256	
BLSTM Layer 2	Hidden Units: 256	

3.3.0.2 Connectionist Temporal Classification (CTC)

In cursive languages, segmentation of text into their relevant classes is a challenging task. Therefore, we employed a connectionist temporal classification layer to interpret the network output into their relevant classes. It is used as an output layer with LSTM for sequence labeling. It computes the conditional probabilities of labels on the basis of an input sequence. A CTC layer has one extra unit than the number of labels in an input sequence. To produce the continuous output of the network, CTC uses softmax activation. The activation of the first label is considered as the probability of observing the corresponding labels at the specific time stamp. The activation of extra output is considered as the probability of observing a "blank" or "no-label". The aggregate probability of any specific label can be computed by summing the all probability values in their corresponding alignments. The working of CTC is represented in Figure 3.11.

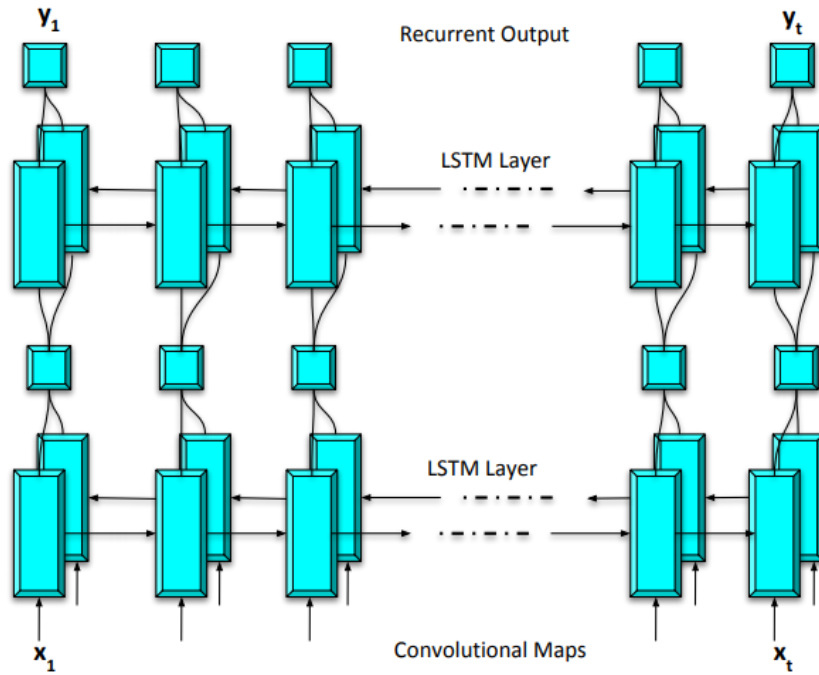


Figure 3.10: (a): BLSTM Architecture

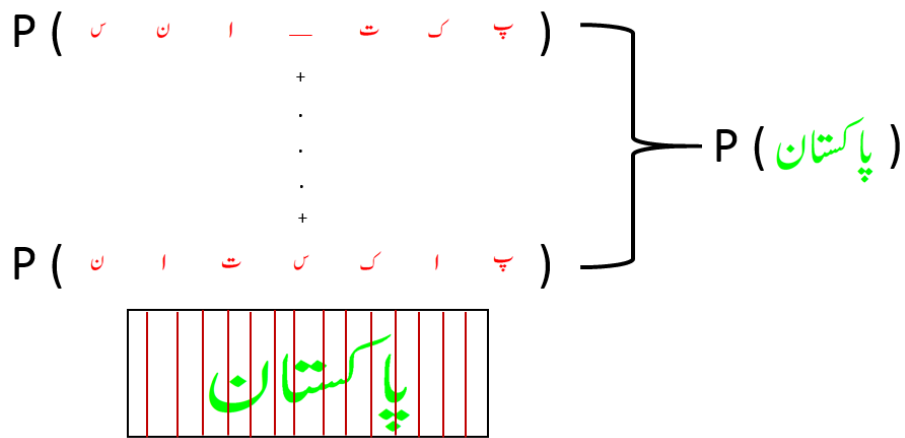


Figure 3.11: CTC computing the probability of an output sequence "PAKISTAN" (written in Urdu)

3.4 Conclusion

After a detailed discussion of each phase of the proposed methodology here, we present the conclusion of our solution architecture. The input handwriting image is binarized as a pre-processing step, the height of the text line is normalized and the resulting image is fed to the convolutional layers. The convolutional layers produce a volume of feature maps which is converted to feature sequences using sliding windows and these sequences are fed to the recurrent net. The recurrent net is a bi-directional LSTM. LSTMs are known to outperform the vanilla recurrent nets which suffer from the vanishing gradient problem once attempting to model long term dependencies. The LSTM

layers are followed by the CTC layer which aligns the feature sequences with the ground truth transcription during training and decodes the output of the LSTM layer to produce the predicted transcription during the evaluation phase. The system is trained in an end-to-end manner by feeding it with text lines images and the respective ground truth transcriptions (in UTF-8). The network architecture comprises seven convolutional layers followed by two B-LSTM layers.

Chapter 4

Results and Discussion

This chapter presents the details of experiments carried out to evaluate the effectiveness of the proposed recognition system. We employed UNHD as well as UHTI Urdu handwritten dataset to accomplish this study. We first present the experimental protocol and results. Later, we present computed recognition rate in different experimental scenarios. Finally, we summarize a few of the recent studies on recognition of printed and handwritten Urdu text for comparison purpose.

4.1 Experimental Protocol and Results

To evaluate the effectiveness of recognition system we employed two different Urdu handwritten dataset i.e. publicly available data set (UNHD) and our custom developed dataset (UHTI). As explained earlier, the UNHD dataset contains 10,000 Urdu text lines but there are only 700 unique text lines that were collected from 500 Urdu writers. Though authors claimed UNHD is publicly available for experiment purpose, in fact, the only subpart of UNHD i.e. (only 4200 Urdu text lines) is available. We acquired these text lines and apply different preprocessing steps (as already explained in chapter 3). Initially, we employed UNHD data set and computed character recognition rate. We employed 3200 Urdu text lines in the training set and 500 each in the validation and test set. The UNHD data set division is illustrated in Table 4.1.

Table 4.1: UNHD Data set division

Training Set	3200 Urdu text lines
Validation Set	500 Urdu text lines
Test Set	500 Urdu text lines

The UNHD data has few limitations and drawbacks, due to which the computed results on UNHD data set cannot depict the practical and actual situation. Therefore, in order to realize the character recognition rate that can depict the actual scenario we developed our custom data set named UHTI. We also evaluated the effectiveness of our recognition system on UHTI and reported promising results. The division of UHTI data set is illustrated in Table 4.2

Table 4.2: UHTI Data set division

Training Set	4000 Urdu text lines
Validation Set	1000 Urdu text lines
Test Set	1000 Urdu text lines

We trained our model by using UNHD as well as UHTI data set. The detail description of model training and computed recognition rate on UNHD and UHTI is illustrated in subsequent sections.

4.1.1 Character recognition rate on UNHD data set

Initially, we evaluated the effectiveness of our recognition system on a subpart of UNHD handwritten dataset. The input image of UNHD text line is binarized as a pre-processing step, the height of the text line is also normalized and the resulting image is fed to convolutional layers for feature extraction. These features were fed to recurrent layers for classification. We trained the model on a 1080x GPU workstation in 30 minutes (Due to the lesser number of training text lines). We also performed cross-validation using 500 Urdu text lines. The training loss of a model as a function of the number of epochs is illustrated in Figure 4.1. After training the model, we evaluated our

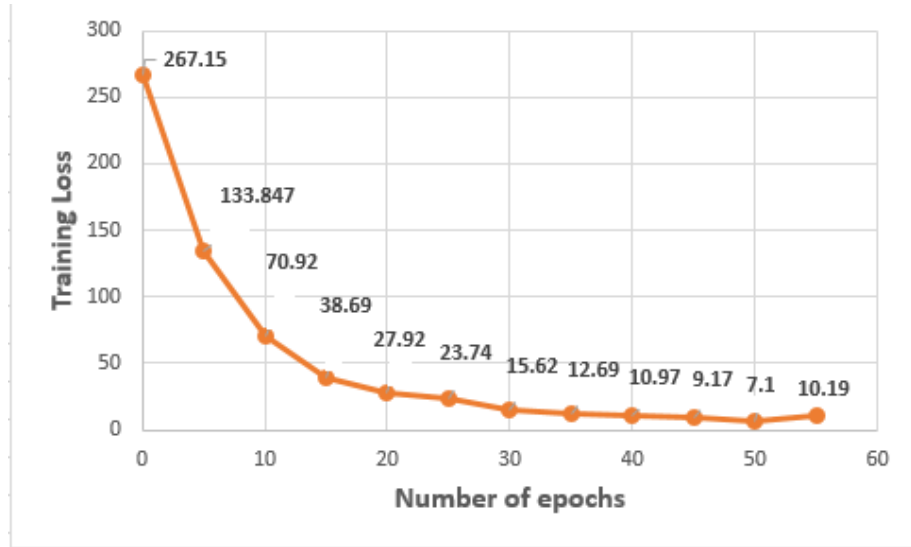


Figure 4.1: The training loss of a model on UNHD as a function of the number of epochs

trained model on test data (500 Urdu text lines). The (character) recognition rate is computed by calculating the Levenshtein distance between the ground truth and the predicted transcription. We computed a very promising character recognition rate which is **88.35%**. The reason for this high recognition rate is text-dependent behavior of UNHD data set because the training and test set contains the same context. Later, We also computed the character recognition rate with different size of training images. The computed recognition rate on different size of training images is presented in Figure 4.2.

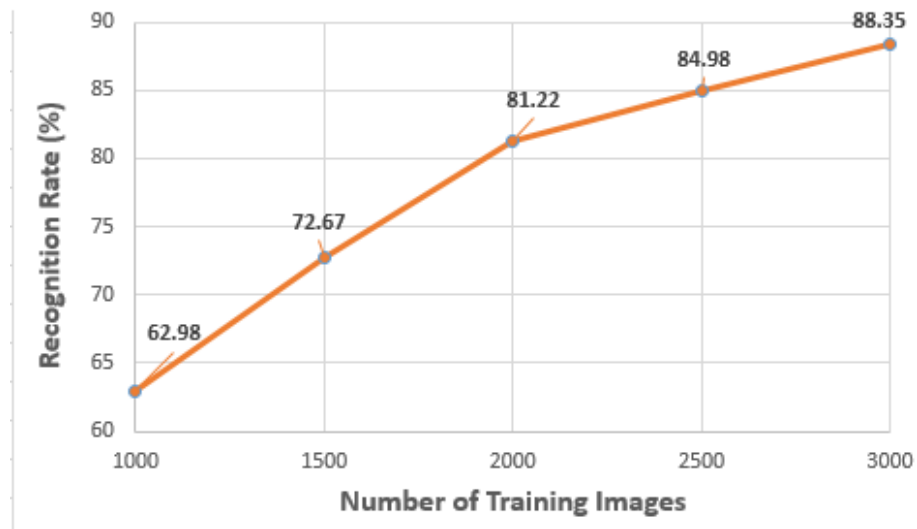


Figure 4.2: Recognition rate as a function of the size of training data

4.1.2 Character recognition rate on UHTI data set

In this analysis, we evaluated our recognition system on UHTI data set. The objective of this experiment is to analyze the character recognition rate on text-independent data where training and test sets contain different context as happens in real scenarios. The data set contains 6000 Urdu text lines where each line has unique information. The data set was acquired from 600 Urdu writers. We employed 4000 text lines for training and 1000 in each validation and test set. We adopted the same experimental setting to train the model on this dataset. We trained our model with a different number of epochs. The training loss of a model as a function of the number of epochs is presented in Figure 4.3

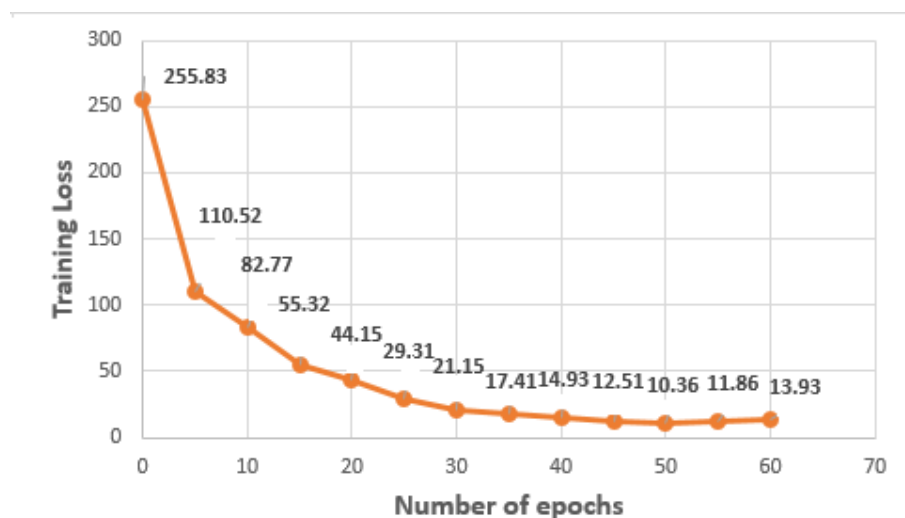


Figure 4.3: The training loss of a model on UHTI as a function of the number of epochs

After model training, we evaluated our trained model on 1000 Urdu text lines where we computed the edit distance between predicted text and ground truth information. We realized **83.69%** character recognition accuracy. The recognition rate on our data set is quite low as compared to the recognition rate computed on UNHD. Our data set contains independent text lines in the training set as well as in test set which is closer to the practical situation. We computed the recognition rate on different size of training images as presented in Figure 4.4.

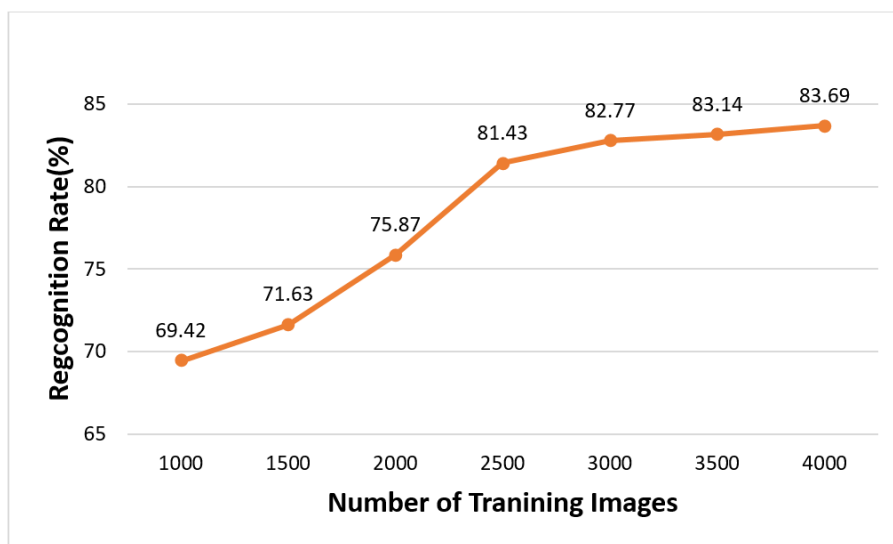


Figure 4.4: Recognition rate as a function of the size of training data

4.2 Performance Analysis and Discussion

We computed the recognition rate on UNHD and UHTI data set using a hybrid network of CNN and LSTM. Instead of this, we also conducted a different experiment with different protocols in order to highlight the effectiveness of our recognition system. We considered two experimental scenarios along with proposed recognition technique.

- **Scenario-I:** Character Recognition rate using LSTM with pixel values using UNHD as well as UHTI data set.
- **Scenario-II:** Character Recognition rate using transfer- learning where trained on printed Urdu data set (UPTI) and fine-tuned on Urdu handwritten data set (UNHD and UHTI).

4.2.1 Character recognition using BLSTM with pixel values

Recurrent Neural networks are good in recognizing patterns that appeared in time series. Traditional RNNs have few drawbacks i.e requirement of pre-segmented input and decay in predicting output when input patterns are too long (vanishing gradient). The LSTM model of RNN overcomes the vanishing gradient problem by introducing the memory cells in hidden layers. LSTMs are suitable

Table 4.3: BLSTM with pixel value based recognition rate

Dataset	Recognition rate
UNHD	73%
UHTI	70.35%

for text classification particularly when the input sequence is too long. As learned from literature many studies employed BLSTM model of RNN for text classification. In this context, we also employed BLSTM for handwritten character recognition where we fed pixel values as a features. The computed recognition rate on UNHD and UHTI is illustrated in Table 4.3.

4.2.2 Character recognition using transfer learning

A number of studies employed CNN for feature extraction and reported a high recognition rate. The lower layers of CNN is used to extract the more generic features while higher layers are used to classify these features. CNN is also used for transfer learning where CNN is trained on one type of problem and fine-tuned on another but similar problem. The transfer learning is employed in deep learning where the model is trained on large and complex data set and their general features are used to enhance the performance of the second task. As inspired from [30], in this analysis, we carried transfer learning using CNN. We employed very famously printed Urdu data set (UPTI) to train our model from scratch and UHTI and UNHD dataset were used to fine-tune the model respectively. The division of UPTI data set for model training is presented in Table 4.4. We trained

Table 4.4: UPTI Data set division

Training Set	8000 Urdu text lines
Validation Set	2000 Urdu text lines

the model using the above mentioned experimental setting. The training graph of a model as a function of a number of epochs is illustrated in Figure 4.5.

Later, we considered UNHD as well as UHTI handwritten data set respectively to fine-tuned the model. The division of UNHD and UHTI data set for fine tuning is illustrated in Table 4.5

Table 4.5: UNHD and UHTI Datasets division for fine-tuning

UNHD	Training: 3200 lines Validation: 500 lines Test: 500 lines
UHTI	Training: 4000 lines Validation: 1000 lines Test : 1000 lines

During fine tuning, we freeze lower 4 layers of CNN in order to use previously learned weight values and retrained top 3 layers to extract high features from handwritten text lines. The retraining loss of UNHD and UHTI dataset is illustrated in Figure 4.6 and 4.7

We realized our recognition accuracy is increased 4 times on both datasets. We computed **92.5%** recognition accuracy on **UNHD data set** and **87.35%** on **UHTI dataset**.

The summary of the conducted experiment using UNHD and UHTI data set is presented in Table 4.6.

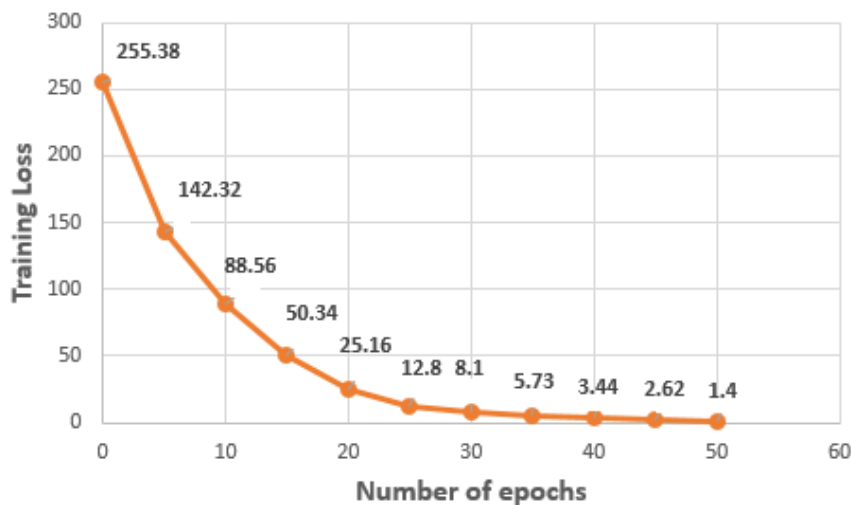


Figure 4.5: UPTI-Model Training

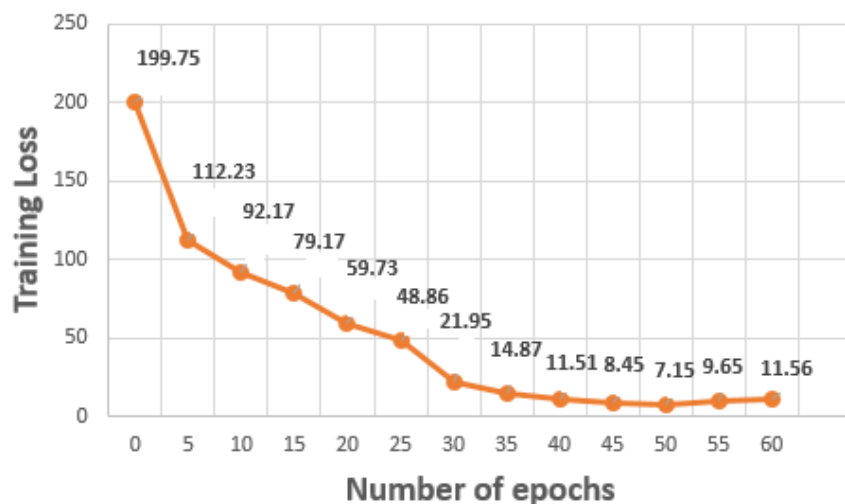


Figure 4.6: The training loss of a model during fine-tuning on UNHD data set as a function of the number of epochs

4.3 Comparison

For comparison purposes, we summarize a few of the recent studies on recognition of printed and handwritten Urdu text (along with the realized results) in Table 4.7. The recognition rates on the printed text are naturally high and are provided as a baseline only. Among techniques targeting recognition of handwritten text, Sagheer et al. [48] report a recognition rate of 97%. The technique, however, is evaluated on isolated characters and a very small subset of frequently used (isolated) Urdu words and hence cannot be employed in real-world recognition scenarios. The recognition rate of 94% is reported in [42] on a small set of 50 text lines in training and 20 in the test. Recognition with LSTMs on raw pixels reports a recognition rate of 92% in [43] on 1840

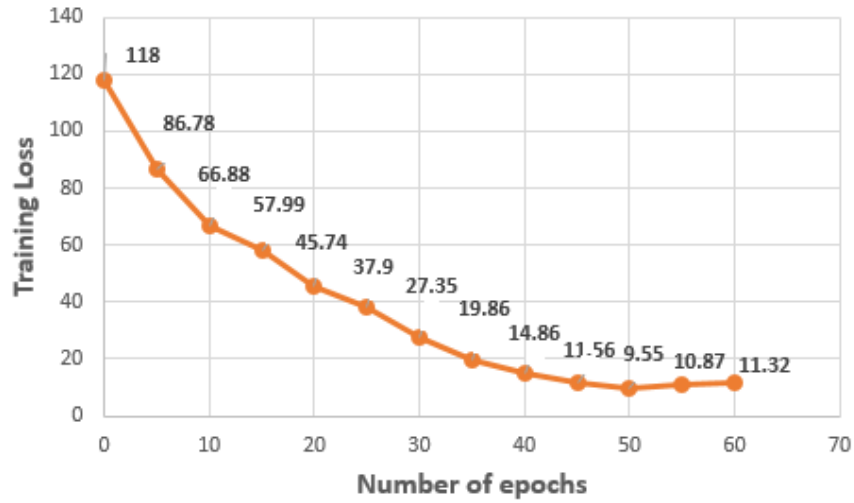


Figure 4.7: The training loss of a model during fine-tuning on UHTI data set as a function of number of epochs

Table 4.6: Summary of Results

Technique	Data-Base	Experiment	Result
CNN + BLSTM	UNHD	Training Set: 3200 Text lines Validation Set: 500 Text lines Test Set: 500 Text lines	88.35 %
Transfer learning using UPTI	UNHD	Training Set: 3200 Text lines, Validation Set: 500 Text lines, Test Set: 500 Text lines	92.5 %
Raw pixel with BLSTM	UNHD	Training Set: 3200 Text lines, Validation Set: 500 Text lines, Test Set: 500 Text lines	73 %
CNN + BLSTM	UHTI	Training Set: 4000 Text lines Validation Set: 1000 Text lines Test Set: 1000 Text lines	83.69 %
Transfer learning using UPTI	UHTI	Training Set: 4000 Text lines Validation Set: 1000 Text lines Test Set: 1000 Text lines	87.35 %
Raw pixel with BLSTM	UHTI	Training Set: 4000 Text lines Validation Set: 1000 Text lines Test Set: 1000 Text lines	70.35 %

test lines. It is, however, important to note that though the dataset is claimed to have 10,000 text lines, the number of unique lines is only 700 implying that text lines in train and test sets have same semantic content. Though our system reports a recognition rate of 83%, all 6000 text lines in our dataset are unique and no text line is common in training and test sets to match the challenging real-world scenarios. Considering the complexity of the script and the challenging experimental setup, the reported recognition rate is indeed very promising.

To provide an insight into recognition errors we also illustrate few of the errors in Figure 4.8 where it can be seen that in most cases, the ground truth and the predicted characters have very

Table 4.7: Recognition rates of notable studies on (printed and handwritten) Urdu text

Type	Study	Technique	Database	Experiments	Results
Printed	Hussain et al. [14]	DCT with HMMs	CLE	5249 primary ligatures	87.44%
	Ahmed et al. [15]	Raw pixels with BLSTM	UPTI	Training: 12,415 lines, Test: 2,836 lines	96%
	Naz et al. [11]	Statistical features with MD-RNN	UPTI	Training: 6800 lines, Test: 1600 lines	96.40%
	Din et al. [58]	CNN	CLE & UPTI	2782 Ligature classes	88% /95%
Handwritten	Sagheer et al. [48]	Gradient and structural features with SVM	CENPARMI	1817 Handwritten words	97%
	Ahmed et al. [42]	Raw pixels with BLSTM	UCOM	Training:50 text lines Test:20 text lines	94%
	Ahmed et al. [43]	Raw pixels with BLSTM	UNHD	Training:6400 text lines Test:1840 text lines	92.07%
	Proposed Study	CNN with BLSTM	Custom data set (UHTI)	Training:4000 text lines Test:1000 text lines	83.69%

similar shapes. In some cases, the main body of the character is correctly identified but the wrong number of dots lead to recognition errors.

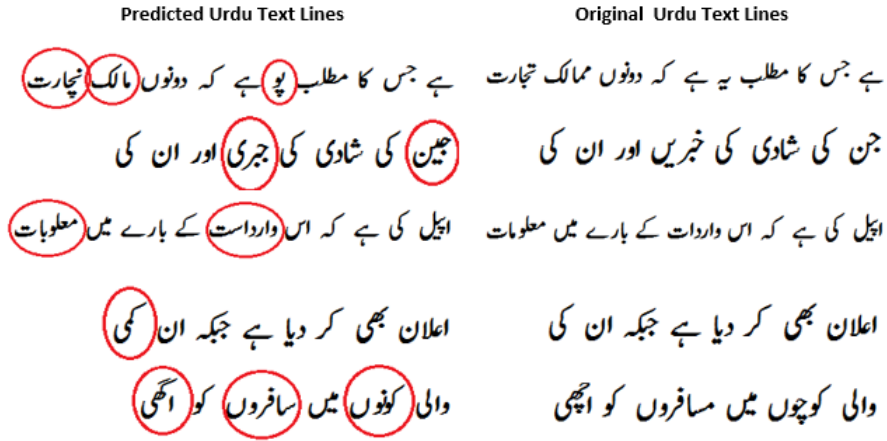


Figure 4.8: Examples of recognition errors

Chapter 5

Conclusion and Perspectives

Handwriting recognition has remained one of the most investigated pattern classification problems. The problem has witnessed many decades of extensive research and has progressively matured from the recognition of isolated characters to complex cursive scripts. Handwriting recognition systems convert handwritten text into the machine-readable form and work either on offline images (scanned or camera based) or on writing captured directly on a digitizing device (online recognition). Recognition of handwritten text is considered more challenging as opposed to the printed text primarily due to writer-specific preferences in drawing character shapes (allographs) and joining various characters. In addition to these writer-dependent variations, another important factor is the complexity of the writing script. This study investigated the problem of recognition of Urdu handwritten characters in cursive Nastalique style. The concluding remarks on the results of this study as well as future research direction are presented in subsequent sections.

5.1 Conclusion

This study presented an effective recognition technique for cursive handwritten Urdu text. The technique implies on implicit segmentation of characters where ground truth transcription and text line images are fed to the learning algorithm to learn character shapes and segmentation points. We employed a combination of convolutional and recurrent neural networks where features extracted by convolutional layers are fed to a BLSTM network for classification. Experimental validation of the proposed recognition system is carried out on two standard benchmarks of Urdu handwritten datasets i.e UNHD and UHTI (custom developed dataset). We evaluated the performance of the proposed system on both datasets separately.

The key findings of this research study are illustrated below.

- We realized a high character recognition rate on our proposed recognition system using UNHD and UHTI datasets.
- We also evaluated two standard benchmarks on BLSTM model of RNN for character recognition in order to highlight the effectiveness of machine-learned based features as compared to raw pixel values.

- We also investigated the effect of transfer learning using the pre-trained model on character recognition problem where we trained our model using UPTI data set and fine-tuned on UNHD as well as on UHTI dataset respectively. We observed the character recognition rate improves 4 times on both datasets.

For comparison purposes, we summarized a few of the recent studies on recognition of printed and handwritten Urdu text (along with the realized results). The recognition rates on the printed text are naturally high and are provided as a baseline only. Though our system reported a recognition rate of 83%, all 6000 text lines in our dataset are unique and no text line is common in training and test sets to match the challenging real-world scenarios.

5.2 Perspectives

This study primarily focused on the recognition of Urdu handwritten text, the proposed technique can also be implemented for other cursive languages like Arabic, Pashto, and Persian. The process of data collection and labeling continues and we intend to collect a dataset of around 25000 labeled text lines. In our further investigations on this problem, we aim to compare the performance of implicit segmentation based recognition with ligature (partial word) level recognition. Furthermore, separate recognition of main body ligatures and dots can also be explored to reduce the number of character classes.

Chapter 6

Research Publication

6.1 Accepted for Publication

1. **Shahbaz Hassan**, Ayesha Irfan, Ali Mirza, and Imran Siddiqi. Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A case study on Urdu Handwriting. In 1st International Conference on Deep Learning and Machine Learning in Emerging Applications (DEEP-ML 2019).

Bibliography

- [1] Chun Lei He, Ping Zhang, Jian Xiong Dong, Ching Y Suen, and Tien D Bui. The role of size normalization on the recognition rate of handwritten numerals. In *The 1st IAPR TC3 NNLPAR workshop*, pages 8–12. Citeseer, 2005. No Citations.
- [2] Matthias Zimmermann and Horst Bunke. Automatic segmentation of the iam off-line database for handwritten english text. In *Object recognition supported by user interaction for service robots*, volume 4, pages 35–39. IEEE, 2002. Cited on pp. 1 and 4.
- [3] Paul W Handel. Statistical machine, June 27 1933. US Patent 1,915,993. Cited on p. 2.
- [4] EE Fournier d’Albe. On a type-reading optophone. In *Proc. R. Soc. Lond. A*, pages 373–375, 1914. Cited on pp. ix and 2.
- [5] HF Schantz. The history of ocr, optical character recognition. <https://catalog.hathitrust.org/Record/000102872>, 1982. Cited on p. 2.
- [6] A Chaudhuri. Some experiments on optical character recognition systems for different languages using soft computing techniques. Technical report, Birla Institute of Technology Mesra, Patna Campus, India, 2010. Cited on p. 2.
- [7] John G Linvill. *Research and development of tactile facsimile reading aid for the blind: the optacon*. US Dept. of Health, Education and Welfare, Office of Education, Bureau of Education for the Handicapped, 1973. Cited on pp. ix and 3.
- [8] Sherif Abdelazeem and Hesham M Eraqi. On-line arabic handwritten personal names recognition system based on hmm. In *2011 International Conference on Document Analysis and Recognition*, pages 1304–1308. IEEE, 2011. Cited on p. 3.
- [9] Jin Chen, Bing Zhang, Huaigu Cao, Rohit Prasad, and Prem Natarajan. Applying discriminatively optimized feature transform for hmm-based off-line handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 219–224, 2012. Cited on p. 3.
- [10] Ibrar Ahmad, Xiaojie Wang, Ruifan Li, and Shahid Rasheed. Offline urdu nastaleeq optical character recognition based on stacked denoising autoencoder. *China Communications*, 14(1):146–157, 2017. Cited on pp. 3 and 19.
- [11] Saeeda Naz, Arif I Umar, Riaz Ahmad, Saad B Ahmed, Syed H Shirazi, Imran Siddiqi, and Muhammad I Razzak. Offline cursive urdu-nastaliq script recognition using multidimensional recurrent neural networks. *Neurocomputing*, 177(1):228–241, 2016. Cited on pp. 3, 4, 7, 18, 23, and 40.

- [12] Mohammad Tanvir Parvez and Sabri A Mahmoud. Offline arabic handwritten text recognition: a survey. *ACM Computing Surveys (CSUR)*, 45(2):23, 2013. Cited on p. 3.
- [13] Imran Siddiqi and Nicole Vincent. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. *Pattern Recognition*, 43(11):3853–3865, 2010. Cited on p. 3.
- [14] Sarmad Hussain, Salman Ali, et al. Nastalique segmentation-based approach for urdu ocr. *International Journal on Document Analysis and Recognition (IJ DAR)*, 18(4):357–374, 2015. Cited on pp. 4, 7, 17, and 40.
- [15] Saad Bin Ahmed, Saeeda Naz, Muhammad Imran Razzak, Shiekh Faisal Rashid, Muhammad Zeeshan Afzal, and Thomas M Breuel. Evaluation of cursive and non-cursive scripts using recurrent neural networks. *Neural Computing and Applications*, 27(3):603–613, 2016. Cited on pp. 4, 18, and 40.
- [16] Irfan Ahmad, Sabri A Mahmoud, and Gernot A Fink. Open-vocabulary recognition of machine-printed arabic text using hidden markov models. *Pattern Recognition*, 51:97–111, 2016. Cited on p. 4.
- [17] Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsheer, and Awais Adnan. Urdu nastaleeq optical character recognition. In *Proc. of world academy of science, engineering and technology*, pages 249–252, 2007. Cited on pp. 4, 16, and 17.
- [18] U Pal and Anirban Sarkar. Recognition of printed urdu script. In *Proc. 7th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1183–1187, 2003. Cited on pp. 4, 16, and 17.
- [19] Inam Shamsheer, Zaheer Ahmad, Jehanzeb Khan Orakzai, and Awais Adnan. Ocr for printed urdu script using feed forward neural network. In *Proc. of World Academy of Science, Engineering and Technology*, pages 172–175, 2007. No Citations.
- [20] Junaid Tariq, Umar Nauman, and Muhammad Umair Naru. Softconverter: A novel approach to construct ocr for printed urdu isolated characters. In *Proc. of 2nd International Conference on Computer Engineering and Technology (ICCET)*, pages V3–495, 2010. Cited on p. 4.
- [21] Ronaldo Messina and Jérôme Louradour. Segmentation-free handwritten chinese text recognition with lstm-rnn. In *Proc. of 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 171–175, 2015. Cited on p. 4.
- [22] Anupama Ray, Sai Rajeswar, and Santanu Chaudhury. Text recognition using deep blstm networks. In *Proc. of International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–6, 2015. No Citations.
- [23] Adnan Ul-Hasan, Syed Saqib Bukhari, and Andreas Dengel. Ocroract: A sequence learning ocr system trained on isolated characters. In *Proc. of 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 174–179, 2016. No Citations.
- [24] Mohammad Reza Yousefi, Mohammad Reza Soheili, Thomas M Breuel, Ehsanollah Kabir, and Didier Stricker. Binarization-free ocr for historical documents using lstm networks. In *Proc. of 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1121–1125, 2015. Cited on p. 4.

- [25] Nazly Sabbour and Faisal Shafait. A segmentation-free approach to arabic and urdu ocr. In *Proc. of Conference on Document Recognition and Retrieval*, page 86580N, 2013. Cited on pp. 7, 13, and 21.
- [26] Sobia Tariq Javed and Sarmad Hussain. Segmentation based urdu nastalique ocr. In *Proc. of Iberoamerican Congress on Pattern Recognition CIARP*, pages 41–49, 2013. Cited on p. 7.
- [27] Israr Uddin Khattak, Imran Siddiqi, Shehzad Khalid, and Chawki Djeddi. Recognition of urdu ligatures-a holistic approach. In *Proc. of 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 71–75, 2015. Cited on p. 7.
- [28] Ankur Rana and Gurpreet Singh Lehal. Offline urdu ocr using ligature based segmentation for nastaliq script. *Indian Journal of Science and Technology*, 8(35):1–9, 2015. Cited on p. 7.
- [29] Ibrar Ahmad, Xiaojie Wang, Yuz hao Mao, Guang Liu, Haseeb Ahmad, and Rahat Ullah. Ligature based urdu nastaleeq sentence recognition using gated bidirectional long short term memory. *Cluster Computing*, (1):1–12, 2017. Cited on pp. 7 and 19.
- [30] Saeeda Naz, Arif I Umar, Riaz Ahmad, Saad B Ahmed, Syed H Shirazi, and Muhammad I Razzak. Urdu nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features. *Neural Computing and Applications*, 28(2):219–231, 2017. Cited on pp. 7, 18, 23, and 37.
- [31] Herbert Jaeger. *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach*, volume 5. GMD-Forschungszentrum Informationstechnik Bonn, 2002. Cited on p. 7.
- [32] Mohamed Attia, Mohamed S El-Mahallawy, Mohsen AA Rashwan, Waleed Nazih, and Mohamed ASAA Al-Badrashiny. Omnifont text recognition of printed cursive scripts via hmms, compact lossless features, and soft data clustering. *Pattern Analysis and Applications*, 18(3):507–521, 2015. Cited on p. 9.
- [33] Abdulwahab Krayem, Nasser Sherkat, Lindsay Evett, and Taha Osman. Holistic arabic whole word recognition using hmm and block-based dct. In *Proc. of 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1120–1124, 2013. No Citations.
- [34] Raid Saabni. Efficient recognition of machine printed arabic text using partial segmentation and hausdorff distance. In *Proc. of 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 284–289, 2014. Cited on p. 9.
- [35] A Durrani. Pakistani: lingual aspect of national integration of pakistan. Technical report, Ministry of Education Curriculum Review, Pakistan, 2009. Cited on p. 9.
- [36] Sobia Tariq Javed. Investigation into a segmentation based ocr for the nastaleeq writing system, 2007. Cited on pp. ix and 10.
- [37] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. Cited on p. 13.
- [38] Mario Pechwitz, S Snoussi Maddouri, Volker Märgner, Nouredine Ellouze, Hamid Amiri, et al. Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer, 2002. Cited on p. 13.

- [39] Somaya Al Maadeed, Wael Ayouby, Abdelâali Hassaïne, and Jihad Mohamad Aljaam. Quwi: An arabic and english handwriting dataset for offline writer identification. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 746–751. IEEE, 2012. Cited on pp. 13 and 14.
- [40] Sabri A Mahmoud, Irfan Ahmad, Mohammad Alshayeb, Wasfi G Al-Khatib, Mohammad Tanvir Parvez, Gernot A Fink, Volker Märgner, and Haikal El Abed. Khatt: Arabic offline handwritten text database. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 449–454. IEEE, 2012. Cited on p. 13.
- [41] Text Corpora and, Image Corpora. <http://www.cle.org.pk/clestore/index.htm>. accessed: 2018-02-07. Cited on pp. 14 and 21.
- [42] Saad Bin Ahmed, Saeeda Naz, Salahuddin Swati, Imran Razzak, Arif Iqbal Umar, and Akbar Ali Khan. Ucom offline dataset-an urdu handwritten dataset generation. *International Arab Journal of Information Technology (IAJIT)*, 14(2), 2017. Cited on pp. 15, 19, 20, 21, 38, and 40.
- [43] Saad Bin Ahmed, Saeeda Naz, Salahuddin Swati, and Muhammad Imran Razzak. Handwritten urdu character recognition using one-dimensional blstm classifier. *Neural Computing and Applications*, pages 1–9, 2017. Cited on pp. 15, 19, 20, 23, 38, and 40.
- [44] Ihtesham Haider and Kamran Ullah Khan. Online recognition of single stroke handwritten urdu characters. In *2009 IEEE 13th International Multitopic Conference*, pages 1–6. IEEE, 2009. Cited on p. 16.
- [45] Syed Afaq Hussain, Safdar Zaman, and Muhammad Ayub. A self organizing map based urdu nasakh character recognition. In *2009 International Conference on Emerging Technologies*, pages 267–273. IEEE, 2009. Cited on pp. 16 and 17.
- [46] Hamna Malik and Muhammad Abuzar Fahiem. Segmentation of printed urdu scripts using structural features. In *2009 Second International Conference in Visualisation*, pages 191–195. IEEE, 2009. Cited on p. 16.
- [47] Q Akram, S Hussain, F Adeeba, S Rehman, and M Saeed. Framework of urdu nastalique optical character recognition system. In *the Proceedings of Conference on Language and Technology.(CLT 14)*, 2014. Cited on p. 17.
- [48] Adnan Ul-Hasan, Saad Bin Ahmed, Faisal Rashid, Faisal Shafait, and Thomas M Breuel. Offline printed urdu nastaleeq script recognition with bidirectional lstm networks. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1061–1065. IEEE, 2013. Cited on pp. 18, 38, and 40.
- [49] Shuwair Sardar and Abdul Wahab. Optical character recognition system for urdu. In *Proc. of International Conference on Information and Emerging Technologies (ICIET)*, pages 1–5, 2010. Cited on pp. 18 and 19.
- [50] Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, and Ching Y Suen. Holistic urdu handwritten word recognition using support vector machine. In *2010 20th International Conference on Pattern Recognition*, pages 1900–1903. IEEE, 2010. Cited on pp. 19 and 20.

- [51] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2017. Cited on p. 23.
- [52] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. Cited on p. 23.
- [53] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Proc. of Advances in neural information processing systems*, pages 396–404, 1990. Cited on p. 25.
- [54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. Cited on p. 25.
- [55] David Bouchain. Character recognition using convolutional neural networks. *Institute for Neural Information Processing*, 2007(1), 2006. Cited on p. 25.
- [56] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. Cited on p. 25.
- [57] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997. Cited on p. 25.
- [58] Israr Uddin, Nizwa Javed, Imran Siddiqi, Shehzad Khalid, and Khurram Khurshid. Recognition of printed urdu ligatures using convolutional neural networks. *Journal of Electronic Imaging*, 28(3):033004, 2019. Cited on p. 40.

test report

ORIGINALITY REPORT

17 %	11 %	13 %	11 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Israr Uddin, Imran Siddiqi, Shehzad Khalid. "A Holistic Approach for Recognition of Complete Urdu Ligatures Using Hidden Markov Models", 2017 International Conference on Frontiers of Information Technology (FIT), 2017 Publication	1 %
2	export.arxiv.org Internet Source	1 %
3	Submitted to Higher Education Commission Pakistan Student Paper	1 %
4	Naila Habib Khan, Awais Adnan. "Urdu Optical Character Recognition Systems: Present Contributions and Future Directions", IEEE Access, 2018 Publication	1 %
5	jivp-urasipjournals.springeropen.com Internet Source	1 %
6	www.mitpressjournals.org Internet Source	<1 %
