

Chapter 1

Introduction

For languages where the alphabet consists of a large number of characters or the characters are of complex shapes, character input through keyboard is not convenient. Characters written in scripts of Chinese, Japanese and Indian Languages (Tamil, Telugu, Bengali, Devanagari, Gurmukhi, Assamese etc.) are complex and keyboard design for these scripts is not easy. Handwriting is an alternative to keyboard. Handwriting could be the most suitable medium of human interaction with computers.

1.1 Online Handwriting Recognition

Online handwriting recognition is a major research topic because of emerging technologies such as Personal Digital Assistants (PDAs) and Tablet PCs. Online recognition refers to the methods and techniques dealing with the automatic processing of a character as it is written using a digitizer [1]. On-line Handwriting Recognition (OHR) involves automatic conversion of text as it is written on a special digitizer, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. This kind of data is known as digital ink and can be regarded as a dynamic representation of handwriting. The obtained signal is converted into letter codes which are usable within computer and text-processing applications. The elements of an OHR interface typically includes a pen or stylus for the user to write with, a touch sensitive surface, which may be integrated with, or adjacent to, an output display and a software application which interprets the movements of the stylus across the writing surface translating the resulting curves

into digital text. The online data contains the temporal information about the writing process, in addition to the spatial shape information of the characters. A recognition process can make use of both the above information for robust performance. OHR activities started in the late 50's and continued till 60's. However, emphasis on OHR diminished until it was renewed in the 80's [1]. Numerous methods and approaches have been proposed and tested for Chinese, Japanese and Western Scripts [2]. Research for online handwriting recognition has also emerged for Indian scripts. Examples of few Indian Scripts where online handwriting recognition activities are being carried out include Tamil [7,8,9], Telugu [9,10], Bengali [11,12], Devanagari [10,13,14] and Gurmukhi [15].

1.1.1 Issues and Challenges

India is a multilingual country with eighteen constitutional languages. Most of the Indian languages have their own scripts and most of the Indian languages have large alphabet of characters. The characters in Indian languages have more complex shapes compared to those of Western languages. This makes character recognition in Indian scenario more challenging than that of Western scripts. Two challenges of recognition are mainly the intra class variability and inter class similarity. The same character is usually written by different writers in different ways. This results in variability of characters within the same class (intra class variability). On the other hand, certain characters belonging to different classes appear to be almost similar to each other, resulting in inter class similarity among those characters. Therefore, recognition of these similar characters becomes difficult. Efficiency of a character recognition system depends on robust feature sets. Different types and categories of features for handwriting recognition have been introduced by many researches. Most of the existing features of online handwriting are script independent. This feature sets work well for Western scripts where character set are small and most of the characters are of simple shapes. For example English alphabet consists of only 52 characters which include 26 upper case characters and 26 lower case characters. However in Indian

scripts, the characters generally are of complex shapes. In most of the Indian languages, the alphabet generally consists of a large number of characters which besides vowels and consonants contain a large number of vowel modifiers and conjunct consonants having very complex shapes. Therefore obtaining robust features for handwritten character recognition in Indian scripts is a challenging task. Availability of standard datasets for character recognition experiment is equally important. A large number of standard datasets of online handwritten characters is available in Western scripts. Some examples of datasets in Western scripts are the English dataset *Pen-Based Recognition of Handwritten Digits* [3], the French dataset IRONOFF [4], the Spanish dataset UJLpenchars [5], and the UNIPEN dataset [6]. Few Indian scripts have also reported standard datasets of online handwritten characters. But not all datasets of Indian scripts are known to be publicly available. Some examples of few datasets of online handwritings in the context of Indian scripts are the HP Lab Tamil Dataset [21], HP Lab Telegu Dataset [22], Bangla Numeral Dataset [23] and the Devanagari Character Dataset [24]. In Assamese no online handwritten Assamese characters dataset is reported.

1.2 *Context and Motivation*

Assamese is an Eastern Indo-Aryan language used mainly in the state of Assam in North-East of India. It is written with the Assamese script. It is the official language of Assam. Assamese is spoken by over 13 million native speakers. It is also spoken in parts of Arunachal Pradesh and other northeast Indian states. Assamese is a major script in the North Eastern part of India. A detailed study on the recognition of online handwritten Assamese characters is important and relevant in this context. Characteristics of Assamese characters (including some distinct properties) are listed below:

- a) Assamese writing has evolved from the ancient Indian script, Brahmi. Brahmi had various classes, and Assamese originated from Gupta Brahmi Script [33].
- b) In Assamese there are 10 numeric characters, 52 basic alphabetic characters consisting of 11 vowels and 41 consonants.

1.2.1 OHR of Assamese Characters

Research in online handwriting recognition for Assamese script has just started in recent years. Not much work is found on OHR for Assamese scripts. A few works on recognition of online handwriting recognition for Assamese scripts are reported recently. GS Reddy et al. [17] describes a HMM based online handwritten numeral recognition system with 96.6% accuracy rate. After having combined this system with an offline numeral recognition system, a recognition rate 99.3% of this combined (online and offline) system is reported. GS Reddy et al. [16] describes a HMM based online handwritten numeral recognition system with 97.14% accuracy rate. B. Sharma et al. [18] reported a Hybrid HMM and SVM Assamese Handwritten Numeral Recognizer and reported a recognition rate of 98.3%. The reported recognition rates are only for online handwritten Assamese numerals. Prasanna et al. [19] describes the development of online Assamese stroke & akshara recognizer based on a set of language rules. A Hidden Markov Model is used to train 181 different stroke patterns which generates a model used during stroke level testing. They have reported a stroke level performance of 94.14% and akshara level performance of 84.2%.

1.3 Objectives

With a motivation to explore OHR for Assamese characters, the thesis focuses on the following objectives:

- a.** Development of a dataset of Online Handwritten Assamese Characters
- b.** Explore standard feature sets from literature for Recognition of Online Handwritten Assamese Characters
- c.** Analysis of Online Handwritten Assamese Characters and put forward a Qualitative Feature Set for Online Handwriting Recognition of Assamese characters.

1.4 Thesis Outlines

The thesis is organized as follows:

- Chapter 2 presents a detailed literature review on online handwriting Recognition. This chapter includes discussions on back grounds of online handwriting recognition, issues related to online handwritten data acquisition including data acquisition tools and techniques and standard datasets of handwritten characters in online domain, various preprocessing techniques applied to online handwritten data and features of online handwritten characters. This chapter also presents a discussion on Qualitative representation of planar outlines. The chapter concludes with a discussion on various classification methods used for online handwriting recognition.
- Chapter 3 presents a description on the development of the dataset of online handwritten Assamese characters (TU-OHAC Dataset)
- Chapter 4 presents initial recognition results of online handwritten Assamese characters. This chapter introduces three initial feature vectors and presents character recognition results based on samples of TU-OHAC Dataset.
- Chapter 5 introduces Qualitative representation of online handwritten Assamese characters.
- Chapter 6 introduces two qualitative feature vectors based on the above qualitative representation for the recognition of online handwritten Assamese characters. Experimental results on character recognition based on samples of TU-OHAC Dataset are presented.
- Chapter 7 summarizes the work and suggests possible future directions.