Chapter 1

Introduction

Writing which has been the most natural method of collecting, storing and transmitting information through the centuries, now serves not only for the communication among humans, but also for the communication of humans and machines.

From the recorded history, only humans were able to recognize and interpret the handwriting of other human being. The term “Handwriting” is defined as meaning to a surface consisting of artificial graphic mark conveying some messages through the marks conventional relation to language [1]. Fueled by the curiosity to uncover the secret of human minds, many scientists began to focus their attention on attempting to mimic intelligent behavior. One of such attempt is to imitate the way the human read and recognize printed and handwritten matter.

Optical Character Recognition (OCR) is the most crucial part of Electronic Document Analysis Systems. The solution lies in the intersection of the fields of pattern recognition, image and natural language processing. Although there has been a tremendous research effort, the state of the art in the OCR has only reached the point of partial use in recent years. Nowadays, cleanly printed text in document with simple layouts can be recognized reliably by off-the-shelf OCR software. There is only limited success in handwriting recognition, practically for isolated and neatly hand-printed characters and word for limited vocabulary. However, in spite of the intensive effort of more than thirty years, the recognition of free style and cursive handwriting as in the example of figure 1 continues to remain in the research arena.
Figure 1: An Example of Handwriting in English and Devanagari script

The difficulty of the free style handwriting recognition is not only because of the great amount of variations involved in the shape of characters, but also because of the overlapping and the interconnection of the neighboring characters. In addition to the peculiarities of an author’s idio-script, which means one writer can be identified among thousands; there are the peculiarities of writing in different situations, with different media and for different purposes. Sometimes it may not be possible to extract the characters from the whole word. Furthermore, when observed in isolation, characters are often ambiguous and require context to minimize the recognition errors.

In order to reach the ultimate goal of fluent machine reading many sub problems have been investigated applying some constraints to the recognition system. One constraint may be imposed on data acquisition equipment, by using a digitizer as in the on-line recognition task is performed concurrently. This approach avoids complicated pre-processing operations and captures the temporal information. On the other hand, the input in the off-line recognition systems is a pre-written document image that is converted into a bit pattern data through an optical scanner. The data is contaminated with various sources of noise. Another
constraint is the restriction of the system to the writing of a specific user. In this case, the system is “writer dependent”. It is possible to make the system learn accurately the writing style of the user and have a satisfactory performance in recognizing a specific writing. Thirdly, the size of the vocabulary can be limited. The system looks for the whole word in lexicon. Which best fits the unknown word. Thus, it avoids the recognition errors due to imperfect segmentation and increases the recognition rates. Lastly, some constraints can be applied on the writing style using pre-printed guidelines or boxes on forms or envelopes.

Therefore, the methods and the success of the recognition rates depend on the level of constraints on handwriting. Considering the roman script, the difficulty is less for handwriting produced as a sequence of separated characters than for the cursive script. For other writing systems, character recognition is hard to achieve, as in the case of Chinese characters, which is characterized by complex shapes and a huge number of symbols. In the case of Devanagari script the situation gets more complicated due to number of vowels, consonants and the half consonants, i.e. similar characters with lot of modifiers and the extensive use of Header line called as Shirekha[2].

Nowadays, there is a renewed interest in the character recognition field, which involves the recognition of both printed and handwritten characters. It is also supported now by the more compact and powerful computers and more accurate electronic equipment such as scanner, camera, electronic tablet etc, which tries to utilize advanced methodologies such as Neural Networks, Hidden Markov Models or Fuzzy Set Reasoning for some better results. But still there is a long way to go in order to reach the ultimate goal of machine simulation of fluent human reading.

Researchers refer to the three important terms: Recognition, Interpretation and Identification [1]. The following figure 2 shows the different Categories of handwritten document processing. It is the act of transforming the graphical marks associated with human handwritten scripts into symbols that are stored on a computer system [1].
The research described in this thesis explores the Recognition of handwritten Devanagari word.
1.1 Types of Handwriting Input: Offline Versus Online

The type of handwriting recognition systems is evolved in two categories due the mode of data acquisition:

- Online recognition
- Offline Recognition

Off-line recognition systems captures the data from paper through optical scanners or cameras whereas the On-line recognition systems utilize the digitizer which directly captures writing with the order of the strokes, speed, pen-up and pen-down information.

It is generally accepted that the on-line recognition method has achieved better result than its offline counter part. This may be attributed to the fact that more information may be captured in the online case such as the direction, speed and the order of strokes of the handwriting. This information is not as easy to recover from handwritten words written on an analog medium such as paper [1,3].

The major advantage of the offline recognizers is to allow previously written and printed texts to be processed and recognized. Some of the applications of the offline recognition are large scale data processing such as postal address reading, cheque sorting, short hand transcription, automatic inspection and identification, reading aid for visually handicapped

1.2 Systems classified according to the Text Type

There are two main areas of interest in character recognition, namely:

- Printed recognition
- Handwritten recognition

The problem of printed character recognition is well understood and solved with little constraints. When the documents are typed on a high quality paper with modern printing technologies, the available systems yield as 99% recognition
accuracy. On the other hand, handwritten character recognition systems have still limited capabilities even for recognition of Latin characters because of the variability in writing style.

Handwriting is an art of drawing pictures for expressing the human thoughts in a condensed and systematic fashion. As a result there are infinite variations of handwriting. A compressive study, where the variability effects in handwritten documents are given in two major categories, is presented in [4]:

1. Perceptual Variability: it is a relatively simple concept, which looks for an answer to the question of “how human reads?” It deals with the subjectivity in perception of the same shape by different subjects or the same subject at different times.

2. Generative Variability: It deals with how characters are written. One important generative variability effect is various styles for connecting the strokes within a word. Figure 3 represents an example for generating three different sets of strokes for the same word.

![Figure 3: Variants for same word generated by connecting the strokes](image)

1.3 Motivation and Approach used

As has been mentioned previously, the recognition of handwritten words via computers is a problem that may be traced back to the first quarter of 20th century. The recognition accuracy for separated handwritten numerals and characters had improved significantly in recent years. Many commercial and accurate systems are available now [5]. Unfortunately the success obtained on machine printed OCR has not readily transferred to the handwriting recognition arena. This may
be attributed to the immense variation in human handwriting [3]. However the final frontier remains the accurate recognition of handwritten printed and cursive words. Again research in this area has also been phenomenal, many systems currently exist.[1][6]. However these are mostly worked on English, Latin, Chinese scripts. Even some research is going on Devanagari script, Telugu and some regional Indian languages, but the pursuit of more accurate recognition rates continues to spur in this field. It can also be seen that along with the challenging nature of handwritten word recognition problem, immense potential lies in the commercial sector to apply the system created.

Major research works in Indian languages are happening at CDAC GIST group, ISI Kolkata, IIIT Hyderabad and Department of Computer Science and Engineering, University of Buffalo. Motivation for this research is due to the challenges in Indian languages, tremendous number of applications and the development of algorithms for intelligence of the system by mimic the human intelligence.

Our work is focused on recognizing the words written in Devanagari script. For our purpose we have used Hindi as well as Marathi language which are based on Devanagari script. A small vocabulary of 10 words, numbers written in words from 1 to 10 each were taken from 31 people from different field and ages. The whole concentration was to get the good features as a vast variety of handwriting style were present in the database. Firstly the database was created as there is no standard database available. The samples from the database were preprocessed using morphological operations; the header line was removed to maintain the uniformity of the characters. The word was segmented to letters without header line and modifiers. Its sparse matrix was taken for dividing into subunits which are acting as a pseudo character or symbol for us. On that Invariant moments were applied for extracting the features. These features were statistically analyzed and classification was done. A novel approach of structured based feature extraction is done with the combination of rule based classification. This feature extraction takes process of scanning the character from to bottom and from left to right for fitting the box within the character. For word recognition this recognized symbols
were used to generate the codebook and finally a hardware display is implemented.

1.4 Organization of the thesis

The remainder of this thesis is organized into the following four chapters. Chapter 2 presents the in-depth overview of handwriting recognition research to date. Chapter 3 discusses the proposed techniques as well as methodology for developing a new handwriting recognition system. Chapter 4 details the result obtained and lastly in chapter 5 have conclusions and also plans for improvements and future research.