Chapter 3

Characteristics of Devanagari Script and Database

This chapter deals with the characteristics of Devanagari script especially Hindi language written in Devanagari script. The database used for experimental purposes has also been described.

3.1 Characteristics of Devanagari Script

India is a multi-lingual country with twenty-two official languages written in twelve scripts. Different languages are used in different parts of the country. In India, more than 300 million people use Devanagari script for documentation. Devanagari was originally developed to write Sanskrit language but later on its use was extended to develop other Indian languages. There are many languages in India like Marathi, Sanskrit, Nepali and Hindi, written in Devanagari script. Devanagari script originated from ancient Brahmi script. Devanagari is derived from Sanskrit words- Deva means ‘god’ and Nagari means ‘city’, so Devanagari means ‘city of gods’.

Character set of Devanagari script consists of large number of basic characters called aksharas. The words in Devanagari not only consist of aksharas but also consist of vowels, consonants and conjunct characters called samyuktaksharas. It also includes compound characters which consist of two or more basic characters. Devanagari script consists of 34 consonants and 13 vowels. Besides consonants and vowels, it also contains
vowel modifiers called matra’s which are placed at left or right part of the character. The complexity of the script increases with the presence of half characters. Another type of modifiers present in Hindi text are upper and lower modifiers which makes the script even more complex. It has no concept of upper or lowercase characters. Writing style in Devanagari script is from left to write. It is a phonetic script. Devanagari script is phonetic as words are written exactly as they are pronounced. It is also syllabic script, which means that text is written using consonants and vowels that together form syllables.

In 1950, Indian constitution, declared Hindi language written in Devanagari script as one of the official language of India. Hindi is the official language of the following states of India: Bihar, Jharkhand, Uttarakhand, Madhya Pradesh, Rajasthan, Uttar Pradesh, Chhattisgarh, Himachal Pradesh, Haryana and Delhi.

The basic aksharas in Hindi language are shown in Figure 3.1.

![Figure 3.1: Basic aksharas in Hindi language](image)

xlii
Apart from basic characters there are compound characters in Hindi language. Combination of two or more basic characters forms a compound character which is very difficult to recognize.

Some of the compound characters which consist of two main aksharas as shown in Figure 3.2.

![Figure 3.2: Compound Characters](image)

The word can be divided into three zones-

i) Upper zone

ii) Middle zone

iii) Lower zone.

The three zones are shown in Figure 3.3.

![Figure 3.3: Different Zones in a word](image)

Upper modifiers are present in upper zone and lower modifiers are present in lower zone.

Half characters may be present in middle zone or lower zone of the word. Left part of the conjunct character is half character and right part is full character. Sample words with half characters are shown in Figure 3.4.
Figure 3.4: Half characters in a word

A horizontal line is drawn over the words in top of the middle region called header line or ‘shirekha’. All the consonants and vowels in the middle region are connected through header line. Header line is the key for line segmentation of text. Due to large character set and presence of conjuncts and compound characters, the Hindi language is difficult to recognize.

Half character is to be separated from full character to recognize the text in segmentation based recognition approach. Occurrence of most frequently used characters in Devanagari is given Table 1 (Jayadevan et al.).

Table 1: Most frequently occurring consonants in Devanagari

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Occurrence (%)</th>
<th>Symbol</th>
<th>Occurrence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>र</td>
<td>10.126</td>
<td>म</td>
<td>3.754</td>
</tr>
<tr>
<td>क</td>
<td>7.949</td>
<td>न</td>
<td>3.360</td>
</tr>
<tr>
<td>ः</td>
<td>7.409</td>
<td>त</td>
<td>2.911</td>
</tr>
<tr>
<td>ः</td>
<td>6.532</td>
<td>य</td>
<td>2.771</td>
</tr>
<tr>
<td>ः</td>
<td>5.064</td>
<td>प</td>
<td>2.473</td>
</tr>
<tr>
<td>ः</td>
<td>4.478</td>
<td>ल</td>
<td>2.456</td>
</tr>
<tr>
<td>ह</td>
<td>4.396</td>
<td>ब</td>
<td>2.268</td>
</tr>
<tr>
<td>ः</td>
<td>4.368</td>
<td>द</td>
<td>1.720</td>
</tr>
<tr>
<td>ः</td>
<td>4.324</td>
<td>ज</td>
<td>1.462</td>
</tr>
<tr>
<td>ः</td>
<td>4.224</td>
<td>ग</td>
<td>1.334</td>
</tr>
</tbody>
</table>
Keeping in view the frequency of occurrence of these characters, efforts are made to extract the features of these characters more accurately.

3.2 Database

No standard database of handwritten Hindi text is available till today (Bag and Harit, 2013). So, we have developed our own database. All experiments were conducted on database constructed by taking handwritten data from 100 writers. Each writer writes one document. Size of documents varies from four to 12 lines of text. Some of the handwritten documents were reduced in size to 35% using MS paint in order to increase the speed of execution.

The handwritten documents were scanned. Some changes creep in the document during various stages of processing. After scanning, some degree of skew may occur in a document. Skew angle is the angle that text lines make with the horizontal direction. So, in some documents, up to 2 degree skew correction was done manually in MS paint. This was done on whole document and not on particular lines. The scanned documents are stored as bitmap files with .bmp extension. No other preprocessing was performed on the data.

All the handwritten documents are scanned at 300 dpi. There are many types of problems in handwritten text. The badly written text can lead to decrease in segmentation rate and hence recognition rate. The irregularities in handwritten text can be divided into two categories:

1) The irregularities that can be avoided

2) The irregularities which can not be avoided.

Some of the irregularities in the text can not be avoided due to writer’s natural way of writing the text. The problems related with writer’s natural handwriting i.e. the
way of writing different characters creates irregularities in data which are difficult to overcome. This leads to decrease in recognition rate.

The irregularities that can be avoided occur due to bad quality of material, bad scanning and most important factor is speed of writing. If a writer uses the gel pen for writing the text then chances are more for touching characters as compared to thin tip pen.

The bad quality of material like paper and pen create fewer irregularities as compared to irregularities created by speed of writing the text. The major number irregularities in same text written by a single writer in different situations occur due to his natural handwriting and speed of writing. The irregularities due to speed of writing the text can be avoided by writing the text in normal speed.

Figure 3.5, Figure 3.6 and Figure 3.7 contain samples of handwritten Hindi database.

Figure 3.5: Part of database
गुठी रोजा से पीड़ित बीमा सहीलों की पहले हिस्सी के लिए अस्पताल से बाहर डायरीशन करना पड़ सही। उन सब्बी के लिए आशा बढ़ती है कि मंगलवार को ठिंडी है। सेवा अस्पताल में आई गई डायरीशन सहीलों के मंगलवार से काम शुरू कर हो रहा है। मंगलवार को डॉक्टर के हैं। आयुक्त की ठीकता से पहले सहील का डायरीशन किया गया है। सौहादी विभाग ने यह सुविधा 750 रुपए में शुरू की है। डायरीशन के दौरान साहसिक विभाग अन्य सामान अपने पास से माहिता करता रहा है। इस योजना में सहीलों को सबसे अधिक प्रायः न्याय देने वाले हैं।

Figure 3.6: Part of database

शिक्षा विभाग की ओर से गत समारंभ लैंडचर व श्रीसीपल बनाने की पहली विवादों में रिहाई गई हैं। श्रीनिवास लैंडचर को नजर आंदोल करते हुए नौंच लैंडचर को प्रभाव करने के साथ आप मामलों के रिश्ता विभाग की पारदर्शिता पर सवालिया विश्वास बना हिचा है।

Figure 3.7: Part of database
An analysis of the database shows that there are many irregularities in the writing of different writers. For example one writer writes the text with no header line. The text with major irregularities has not been considered for experiments.

The review of literature shows that till now no standard database for handwritten Hindi text is available. Considering the application and requirement of the present time there is a need to develop the standard database for handwritten Hindi text so as to be able to compare the results.

3.3 Chapter Summary

In this chapter, characteristics of Devanagari script, especially Hindi language has been explained. The character set and unique features of Hindi language have been discussed. The database prepared for experimental proposes along with sample databases has been described. No standard database is available so all the experiments are performed on our own database.