Handwritten Devnagari Script Database Development for Off-Line Hindi Character with Matra (Modifiers)

Maninder Singh Nehra, Neeta Nain and Mushtaq Ahmed

Abstract Due to advancement in digital technology, handwritten Character recognition plays a significant role for interaction between human and computer. For recognition of handwritten character a standard database is required. There is no benchmark data base of Devnagari script in Hindi with matra’s (modifiers). A database for off-line Hindi handwritten character with modifier is developed. The database consist more than 23000 images of their original size with programmatically segmented consonant and vowels. The Data set is collected from persons of different gender, age, profession and educational qualification. Data are also collected from person of different geographical location of India.

Keywords Handwritten text · Database · Modifiers

Introduction

Recognition of handwritten character is challenging area in pattern recognition because of different writing style of writers. Handwritten character is change according to human age, sex, qualification, working culture and frame of mind. For research in this area required proper escalation in technology. Digitization of handwritten document increase and it is an integral part of optical character recognition. Research in the handwritten character recognition plays a significant role in historical document recognition and their proper arrangement. There are many application of handwritten character recognition like in banking, offices, postal services, form processing, exam evaluation etc.

M.S. Nehra (✉) · Neeta Nain · Mushtaq Ahmed
Malaviya National Institute of Technology, Jaipur, India
e-mail: maninder4unehra@yahoo.com

Neeta Nain
e-mail: neetanain@yahoo.com

Mushtaq Ahmed
e-mail: mahmed.cse@mnit.ac.in

© Springer India 2016
N. Afzalpulkar et al. (eds.), Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing, DOI 10.1007/978-81-322-2638-3_27
Devnagari script is anonomatopoeic of ancient Brahmi script which is the mother of almost all Indian scripts. Devnagari is script in which many languages are written, like the most popular language Hindi, Sanskrit, Konkani, Nepali, Marathi and Sindhi due to Chaudhary [1]. Hindi is the national language of India and the third most frequently used language in the world. So the research in Devnagari script (Hindi) is very useful. The alphabet set of Devnagariscript has 36 consonants and 12 vowels as shown in Figs. 1 and 2. Besides the consonants and vowels it has modifiers also called matra’s which when combined with vowels form compound characters. The modifiers can be placed at the left or right and above or bottom of a vowel. Handwritten character recognition for Devnagari script (Hindi) is of two types, off-line and on-line character recognition. Off-line handwritten characters are those in which handwritten character of writer is converted into digital form by scanning the handwritten paper. On-line handwritten characters are those in which, character are written on electronic surface such as digitizer with special pen. In on-line handwritten, two-dimensional coordinates of successive points of the characters take as function of time are stored which is spatial temporal demonstration of the input character. Whereas in off-line handwritten, the completed character is existing as an image [2]. A standard data base of hand-written Devnagari (Hindi) character is required to train a classifier for recognition. There are many handwritten text databases for languages like English database by Marti and Bunke [3], Urdu database by Sagheer and He [4], Chinese database by Wang

Fig. 1  Hindi handwritten consonants

![Fig. 1](image1.png)

Fig. 2  Hindi handwritten vowels and modifiers

![Fig. 2](image2.png)
and Cheng-Lin [5], Spanish database by Llorens et al. [6], Japanese database by Nakagawa and Matsumoto [7] etc. There are some datasets for Devnagari handwritten character recognition also but no standard Devnagari benchmarked data set with modifiers is available till date.

Bhattacharya and Chaudhuri [8, 9] developed handwritten numeral database for Devnagari and Bangla-Indic script, in which they scanned digitized data at 300 dpi. Suen and Nadal [10] developed database for numerals at CENPARMI (Centre for Pattern Recognition and Machine Intelligence) at Canada. Hull [11] developed database for handwritten text at CEDAR (Center of Excellence for Document Analysis and Recognition) in state university of New York, Marti and Bunke [1] developed IAM-DB database for English etc. are not available freely and are specific. Ram Sarkar and Basu [12] developed a database for handwritten Bangla and Bangla-English (CMATERdb). In which they collected data from 40 different writers and scanned the documented data at 300 dpi.


**Hindi Handwritten Database Creation**

**Data Collection**

For database creation of Hindi handwritten characters with matra, a blank paper of A4 size on which 60 rectangle blocks was designed. Consonants with modifiers and vowels are printed as the first line. A sample form is shown in the Fig. 3a. A total of 37 forms are designed (36 for consonant and one for vowels). The form also has boxes for writers to write their name, profession and tick on rectangles for gender, age–group, region and qualification. To cater to maximum syntactic variations the writers are chosen from different age groups, profession, region and educational
qualifications. The forms filling sessions were carried out at geographically distant locations like shopping malls, railway stations, bus stand and hospitals etc., as the handwriting of a person sometimes gets affected by the mood, situation and surroundings. Multicolored ink (red, blue, green and black) and different styles like gel and ink pens are used for form filling. A sample of the filled sheet is shown in Fig. 3b. Besides this the persons from different geographical location of India are also involved in data collection; those who are comfortable with Hindi and those whose mother tongue is not Hindi. Approximately 1500 hundred of persons have filled the forms.

**Segmentation and Digitization of Data**

For digitization of collected data written by different writers, filled papers were scanned with the help of HP scanner at 600 dpi and scanned images were save in PNG image format. Total 48 folders are created for storage. A sample segmented characters output from is shown in Fig. 4. We have done a manual check on the
segmented characters and discarded the erroneous images. A sample of some discarded characters is shown in Fig. 5.

The segmentation of different characters from the scanned image is a puzzling work. The algorithm for segmentation of characters is explained in the following steps:

**Algorithm:** Handwritten character recognition.

**Input:** \{Scanned form of handwritten characters\}.

**Output:** \{Isolated segmented characters\}.

BEGIN

**Step1:** Read the scanned PNG image.

**Step2:** Convert the scanned image into binary image with appropriate threshold value.

**Step3:** Noise removal.

**Step4:** Calculate eight connected components for segmentation.

**Step5:** Check the segmented characters manually for proper shape.

**Step6:** Discard the erroneous characters.

**Step7:** Store the useful characters in separate folders (per character) in PNG format.

END
Information About Data Base

Devnagari script characters are written in cursive style and character are associated with lines and written on lined paper. The characters are sling from a horizontal line called the header stroke and there is no lower and upper cases like English characters. In Devnagari script about two third of space between line is used by the characters. In general the first stroke, or strokes, in a character are written from the left to the right and are then followed by any down strokes and eventually the head stroke is added. In handwritten documents due to improper writing of persons, character is not associated with lines. By this during the segmentation characters and their modifiers are segmented in improper format. Because of some writers may merge the characters which could not be recognized and sometimes the modifiers are also not properly placed. Such characters are not included in the data base and are discarded. Some of the Devnagari character(Hindi) are not used often like अ, द, ङ, etc. Some of the characters are written in more than one way like झ. The configuration of the properly segmented character with their modifiers is shown in the Table 1. The frequency of all characters is not equal as we have discarded the characters written in error and which encountered segmentation errors. The database is developed, so that it useful for the research in the handwritten character recognition.

Table 1  Hindi Handwritten character data set with frequency

<table>
<thead>
<tr>
<th>S. No</th>
<th>Hindi char.</th>
<th>Freq.</th>
<th>S. No</th>
<th>Hindi char.</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>अ</td>
<td>390</td>
<td>14</td>
<td>आ</td>
<td>290</td>
</tr>
<tr>
<td>2</td>
<td>ख</td>
<td>304</td>
<td>15</td>
<td>ऊ</td>
<td>250</td>
</tr>
<tr>
<td>3</td>
<td>ग</td>
<td>250</td>
<td>16</td>
<td>न</td>
<td>260</td>
</tr>
<tr>
<td>4</td>
<td>घ</td>
<td>290</td>
<td>17</td>
<td>घ</td>
<td>280</td>
</tr>
<tr>
<td>5</td>
<td>ङ</td>
<td>230</td>
<td>18</td>
<td>घ</td>
<td>270</td>
</tr>
<tr>
<td>6</td>
<td>च</td>
<td>310</td>
<td>19</td>
<td>च</td>
<td>300</td>
</tr>
<tr>
<td>7</td>
<td>छ</td>
<td>280</td>
<td>20</td>
<td>छ</td>
<td>300</td>
</tr>
<tr>
<td>8</td>
<td>ज</td>
<td>300</td>
<td>21</td>
<td>ज</td>
<td>320</td>
</tr>
<tr>
<td>9</td>
<td>झ</td>
<td>320</td>
<td>22</td>
<td>झ</td>
<td>250</td>
</tr>
<tr>
<td>10</td>
<td>ञ</td>
<td>280</td>
<td>23</td>
<td>ञ</td>
<td>250</td>
</tr>
<tr>
<td>11</td>
<td>ट</td>
<td>350</td>
<td>24</td>
<td>ट</td>
<td>260</td>
</tr>
<tr>
<td>12</td>
<td>ठ</td>
<td>230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>ड</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S. No</th>
<th>Hindi char.</th>
<th>Freq.</th>
<th>S. No</th>
<th>Hindi char.</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>आ</td>
<td>230</td>
<td>15</td>
<td>ँ</td>
<td>200</td>
</tr>
<tr>
<td>15</td>
<td>ऊ</td>
<td>200</td>
<td>16</td>
<td>ऊ</td>
<td>260</td>
</tr>
<tr>
<td>17</td>
<td>ऊ</td>
<td>300</td>
<td>18</td>
<td>उ</td>
<td>300</td>
</tr>
<tr>
<td>19</td>
<td>उ</td>
<td>380</td>
<td>20</td>
<td>उ</td>
<td>300</td>
</tr>
<tr>
<td>21</td>
<td>ँ</td>
<td>340</td>
<td>22</td>
<td>ँ</td>
<td>360</td>
</tr>
<tr>
<td>23</td>
<td>ं</td>
<td>390</td>
<td>24</td>
<td>ं</td>
<td>370</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

M.S. Nehra et al. deepa.modi22@gmail.com
Conclusion and Future Work

We have developed database for off-line Hindi hand written characters with matras (modifiers) written by around 1500 writers from diverse places and from different backgrounds. During the development procedure, in output, some of the character images are not visibly identifiable, because they are not written properly by the writer and they not recognized by the system. Such characters are discarded. In this database more than 23000 handwritten (alphabets) characters images of consonants and vowels scanned at 600 dpi are created and the character images are stored as images in PNG image format for efficient use. To the best of our knowledge there is no such type of data set available for handwritten characters of Devnagari script in Hindi with modifiers. Such a dataset is very useful for validation of handwritten text recognition algorithms. It could be used for cross validation by dividing it suitably as training and testing data set. The database will be made available publically for researchers.

In future the data set will be extended to develop a complete corpus of handwritten Hindi words and lines which is very useful for benchmarking of handwritten segmentation algorithms.

References


