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

Chapter · January 2016

DOI: 10.1007/978-81-322-2638-3_27

CITATIONS
2

READS
2,995

3 authors:

Maninder Singh Nehra
Malaviya National Institute of Technology Jaipur
14 PUBLICATIONS **86** CITATIONS
SEE PROFILE

Neeta Nain
Malaviya National Institute of Technology Jaipur
113 PUBLICATIONS **569** CITATIONS
SEE PROFILE

Mushtaq Ahmed
Malaviya National Institute of Technology Jaipur
51 PUBLICATIONS **254** CITATIONS
SEE PROFILE

Handwritten Devnagari Script Database Development for off-line Hindi Character with Matra (modifiers)

Maninder Singh Nehra
Malaviya National Institute of Technology, Jaipur
Email:maninder4unehra@yahoo.com

Neeta Nain
Malaviya National Institute of Technology, Jaipur
Email: neetanain@yahoo.com

Mushtaq Ahmed
Malaviya National Institute of Technology, Jaipur
Email: mahmed.cse@mnit.ac.in

Abstract:
Due to advancement in digital technology, handwritten Character recognition plays an important 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. Data set are collected from persons of different age, gender, profession and educational qualification. Data are also collected from person of different geographical location of India.

Keywords: Handwritten Text, Database, Modifiers.

1. 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 Research in the handwritten character recognition plays an important role in historical document recognition and their proper arrangement.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.

Devnagari script is an imitative of ancient Brahmi script which is 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 B.B. 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 Devnagari script has 36 consonants and 12 vowels as shown in Figure 1 and Figure 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.

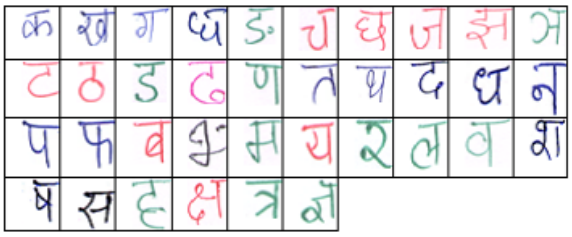


Figure 1 Hindi handwritten consonants.

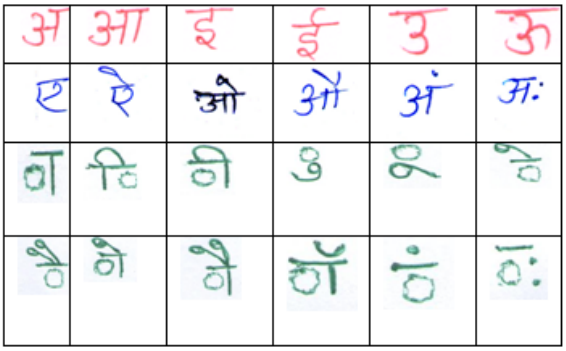


Figure 2. Hindi handwritten vowels and modifiers (Matras)

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 [19]. 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, H. Bunke[20], Urdu database by Malik, Chun Lei [15], Chinese database by Da-Han Wang, Cheng-Lin Liu [16], Spanish database by D. Llorens, F. Prat, A. Marzal, J. M. Vilar [17], Japanese database by Masaki Nakagawal and Kaoru Matsumotol [18] etc. There are some datasets for Devnagari handwritten character recognition also but no standard Devnagari benchmarked data set with modifiers is available till date.

Ujjwal Bhattacharya and B.B. Chaudhuri [4] [6] developed handwritten numeral database for Devnagari and Bangla Indic script, in which they scanned digitized data at 300 dpi. Suen, C.Y. and Nadal [7] developed database for numerals at CENPARMI(Centre for Pattern Recognition and Machine Intelligence) at Canada. Hull and J.J. [3] developed database for handwritten text at CEDAR (Center of Excellence for Document Analysis and Recognition) in state university of New York, U.-V. Marti, H. Bunke [1] developed IAM-DB database for English etc. are not available freely and are specific. Ram Sarkar and Dipak Kumar Basu[5] 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.

Ashutosh- Rajneesh [9] recognized handwritten Devnagari characters by gradient techniques. No standard data set was used, only a sample of 20 writers hand- writing is taken. Somaya and Dave [10] designed a data set for offline Arabic handwritten text with the help of around 100 different writers. Jonathan [11] prepared database for hand-written text of city names, states names and ZIP codes for post office. Vikash and Vijay [13] designed a data set for Devnagari handwritten character without matras (modifiers) with the help of 750 different writers. They scanned the handwritten sheet at 300dpi. Jabri-Ali [12] designed a data set for Arabic handwriting with help of five writers of different ages and educational qualification. Marti and Bunke [20] describes database for English sentences offline handwriting (IAM- database) recognition with the help of around 400 different writers. Marcus and Bunke [2] described a database for on-line English sentences of handwritten text. Bhattacharya and B. B. Chaudhuri [4] discussed a dataset that have 22,556 handwritten data. Sharma, U. Pal, and S. Pal [3] discussed quadratic classifier technique for recognition off-line handwritten Devnagari characters. In this technique the character data set used is around 11000 images, digitized at 300dpi.

II. HINDI HANDWRITTEN DATABASE CREATION

A. 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 figure 3(a). 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, and qualification. To cater to maximum syntactic variations the writers are chosen from different age groups, profession 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 Figure 3(b). Besides this the persons from different graphical 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.

B. 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. 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.

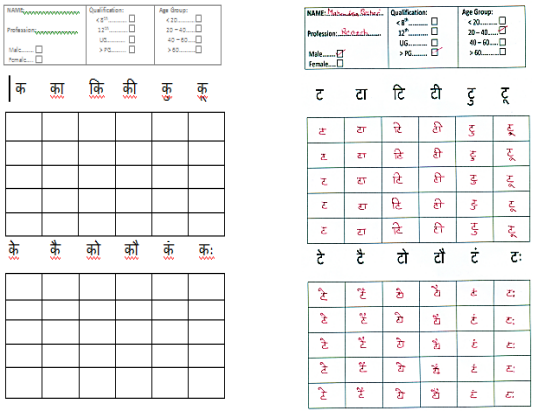
Step5: Check the segmented characters manually for proper shape.

Step6: Discard the erroneous characters.

Step6: Store the useful characters in separate folders (per character) in PNG format.

END

Total 48 folders are created for storage. A sample segmented characters output from is shown in Figure 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 Figure 5.



(a) (b)

Figure 3 (a). A sample blank form (b) A sample filled form

Figure 4: Sample segmented characters extracted from a form filled by a writer.

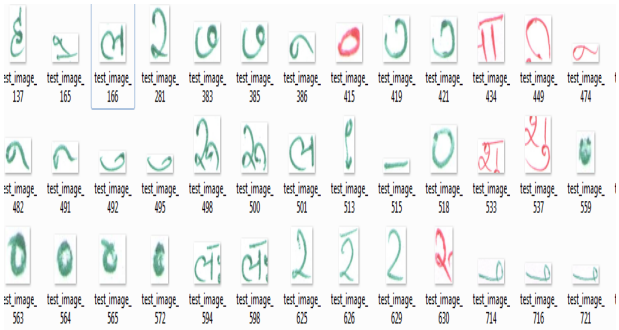


Figure 5. Discarded characters and invalid strokes of modifiers.

III. 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. Devnagari characters use about two thirds of the space between the lines. 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 finally 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 ञ, ड, 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

S. No.	Hindi char.	Freq.	S. No.	Hindi char.	Freq.
1.	क	390	25	म	390
2.	ख	304	26	य	290
3.	ग	250	27	र	260
4.	घ	290	28	ल	280
5.	ङ	230	29	व	270
6.	च	310	30	श	300
7.	छ	280	31	ष	280
8	ज	300	32	स	250
9	झ	320	33	ह	290
10	ञ	280	34	क्ष	230
11	ट	350	35	त्र	200
12	ठ	230	36	ज्ञ	190
13	ड	150	37	अ	300
14	ढ	230	38	आ	290
15	ण	200	39	इ	250
16	त	360	40	ई	230
17	थ	300	41	उ	180
18	द	380	42	ऊ	200
19	ध	320	43	ए	190
20	न	390	44	ऐ	160
21	प	340	45	ओ	170
22	फ	360	46	औ	150
23	ब	390	47	अं	190
24	भ	370	48	अः	160

IV. 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 data base 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

[1] U. Pal and B. B. Chaudhuri, "Automatic Separation of Machine-Printed and Hand-Written Text Lines", In Proc. of ICDAR IEEE, pp. 645-648, 1999.

[2] Marcus Liwicki and Horst Bunke, " IAM-OnDB - an On-Line English Sentence Database Acquired from Handwritten Text on a Whiteboard", at http://www.iam.unibe.ch/~fki/iamondb/.

[3] N. Sharma, U. Pal, F. Kimura and S. Pal, " Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier", In Proc. ICVGIP , pp. 805 – 816, 2006.

[4] Ujjwal Bhattacharya and B.B. Chaudhuri, "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals", In IEEE transaction on Pattern analysis and Machine Intelligence, pp. 444-457, 2009.

[5] Ram Sarkar , Nibaran Das, Subhadip Basu , Kundu , Mita Nasipuri and Dipak Kumar Basu, "MATERdb1: a database of unconstrained handwritten Bangla and Bangla–English mixed script document image", In IJDAR Springer, pp. 1-5, 2012.

[6] B.B. Chaudhuri, "A Complete Handwritten Numeral Database of Bangla – A Major Indic Script", CVPR Unit, Indian Statistical Institute, Kolkata-108, India.

[7] Suen, C.Y. and Nadal, " Computer recognition of unconstrained handwritten numerals", In Proc. IEEE 80(7), pp. 1-6, 1992.

[8] Peb Ruswono Aryan, Iping Supriana, Ayu Purwarianti, "Development of Indonesian Handwritten Text Database offline Character Recognition", In International Conference Electrical Engineering and Informatics, pp. 1-5, 2011.

[9] Ashutosh Aggarwal, Rajneesh Rani, RenuDhir, "Handwritten Devanagari Character Recognition Using Gradient Features", In International Journal of Advanced Research in Computer Science and Software Engineering, pp. 85- 90, 2012.

[10] Somaya Al-Ma'adeed, Dave Elliman, and Colin Higgins, "A Data Base for Arabic Handwritten Text Recognition Research", In The International Arab Journal of Information Technology, pp. 117-121, 2004.

[11] Jonathan J. Hull, "A Database for Handwritten Text Recognition Research", In IEEE Transaction on Pattern Analysis and Machine Intelligence Vol. 16, pp. 550-554, 1994.

[12] Jabril Ramdan, Khairuddin Omar, Mohammad Faidzul, Ali Mady, "Arabic Handwriting Database for Text Recognition", In pro. The 4th International Conference on Electrical Engineering and Informatics, pp. 580-584, 2013.

[13] Vikas J. Dongre and Vijay H.Mankar, "Development of Comprehensive Devnagari Numeral and Character Database for Offline Handwritten Character Recognition", In Proc. Applied Computational Intelligence and Soft Computing, pp. 1-5, 2012.

[14] U. Bhattacharya and B. B. Chaudhuri, "Databases for research on recognition of handwritten characters of Indian scripts" In Proc. 8th ICDAR, pp.789-793, 2005.

[15] Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, and Ching Y. Suen, "A New Large Urdu Database for Off-Line Handwriting Recognition", In Proc. of ICIAP Springer, pp. 538-546, 2009.

[16] Da-Han Wang, Cheng-Lin Liu, Jin-Lun Yu, Xiang-Dong Zhou, "CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters", In Proc. of ICDAC IEEE, pp. 1206-1210, 2009.

[17] D. Llorens, F. Prat, A. Marzal, J. M. Vilar, "The UJIpenchars Database: A Pen-Based Database of Isolated Handwritten Characters", pp. 2647-2651, at http://unipen.org.

[18] Masaki Nakagawa1, Kaoru Matsumoto1, "Collection of on-line handwritten Japanese character pattern databases and their analyses",In IJDAR,pp. 69–81, 2004.

[19] Rejean Plamondon and Sargur N. Srihari, "On-Line. and Off-Line Handwriting Recognition: A Comprehensive Survey'', In IEEE Transaction on Pattern Analysis and Machine Intelligence, pp. 63-84, 2000.

[20] U.V. Marti, H. Bunke, "The IAM- database: and English sentence database for offline handwriting recognition", In International Journal on Document Analysis and Recognition, pp. 39-46, 2002.