Creation of a Complete Hindi Handwritten Database for Researchers

Rama Gaur[1,*] and Dr. V.S. Chouhan[2]
[1,*]Ph.D. Scholar (ECE), Jodhpur National University, Jodhpur, Rajasthan, India.
[2]Professor and Head ECE department, MBM Engineering college, Jodhpur, Rajasthan, India.

The standard database plays a vital role in handwritten character recognition. Performance of various algorithms and results obtained by researchers can be evaluated only by a benchmark database. In the field of Hindi handwritten research such database is not available. My paper is focuses on creation complete database of Hindi character and numerals. I have generated almost 100-200 samples of each character. The images are stored in JPEG format. The dimension of each character is 105X125 pixels.The size of one character is in between 12 to 25 KB. I have taken fix dimension of character to reduce the complexity for the beginners in testing their algorithms. I am sure that this database will be helpful to the future researchers.

Keywords: Character recognition, Database, Algorithms.

1.INTRODUCTION

To automatically recognize the handwritten Hindi character is a very tricky task.The major difficulty with handwritten characters is the unevenness of writing styles, both between different writers and between separate examples from the same writer overtime. The handwritten text in unlike sizes, dimensions, orientations, thicknesses and formats. Sothere is a need of benchmark handwritten database. Advance development of computational power hasincreased the interest of researchers towards machine simulation of human reading. This can be possible through optical character reader (OCR) which recognizes the characters from digitized image of optically scanned documents and is converted into ASCII code or in some other code. A lot of research has been done in developed countries for English, European, and Chinese languages. There is big problem database in front of the researchers of Indian languages[1].

Some work is also done for Indic scripts such as Bangla [2], Kannada [3], and Devanagari [4–7]. India is a multilingual and multiscript country having more than 1.2 billion population with 22 constitutional languages and 10 different scripts. Devanagari is the most popular script in India. Hindi, the national language of India which is spoken by more than 500 million populations worldwide, is written in the Devanagari script. Moreover, Hindi is the third most prevalent language in the world [8].The Department of Information Technology, Government of India, started a program on technology development for Indian languages [9] where language aspects are studied and developed. Another government undertaking Centre for Development of Advanced Computing [10] is actively involved in development of Indian languages fonts,

translators.This paper describes an attempt for generation of a comprehensive database for handwritten Devanagari numerals and characters. This databasedevelopedfor the researchers.And make it accessibleeasily as a benchmark database for handwriting recognition research.The present paper is organized as follows: Section 2 describes the details of offline database generation. Section 3 discusses statistical analysis of this work. Conclusion and further work direction are discussed in Section 4.

2.HINDI HANDWRITTEN OFFLINE DATABASE GENERATION DETAILS

Following steps were followed to generate database:-

2.1.Data Collection

A sheet of A4 size having blank boxes of equal size was designed. And a plane paper of equal size is placed over it and these two papers lying exactly over each other placed over glass table so that squares should be visible and writer has instructed to write within the square.Persons of various ages, sex, education, and occupation were requested to write Devanagari numbers and characters.One sheet contains approximately 150 samples of single character e.g. 02 Numeral database sheets are shown in Figure 1 and 02 Character database sheet are shown in Figure 2.Such sheets are prepared for every character.The writers were carefully chosen to make the database representative. Persons of various languages and educational background have been chosen.The sheet of every character was scanned through Canon CanoscanLide 100 flatbed scanner at horizontal resolution 200dpi and vertical resolution is also 200dpi.

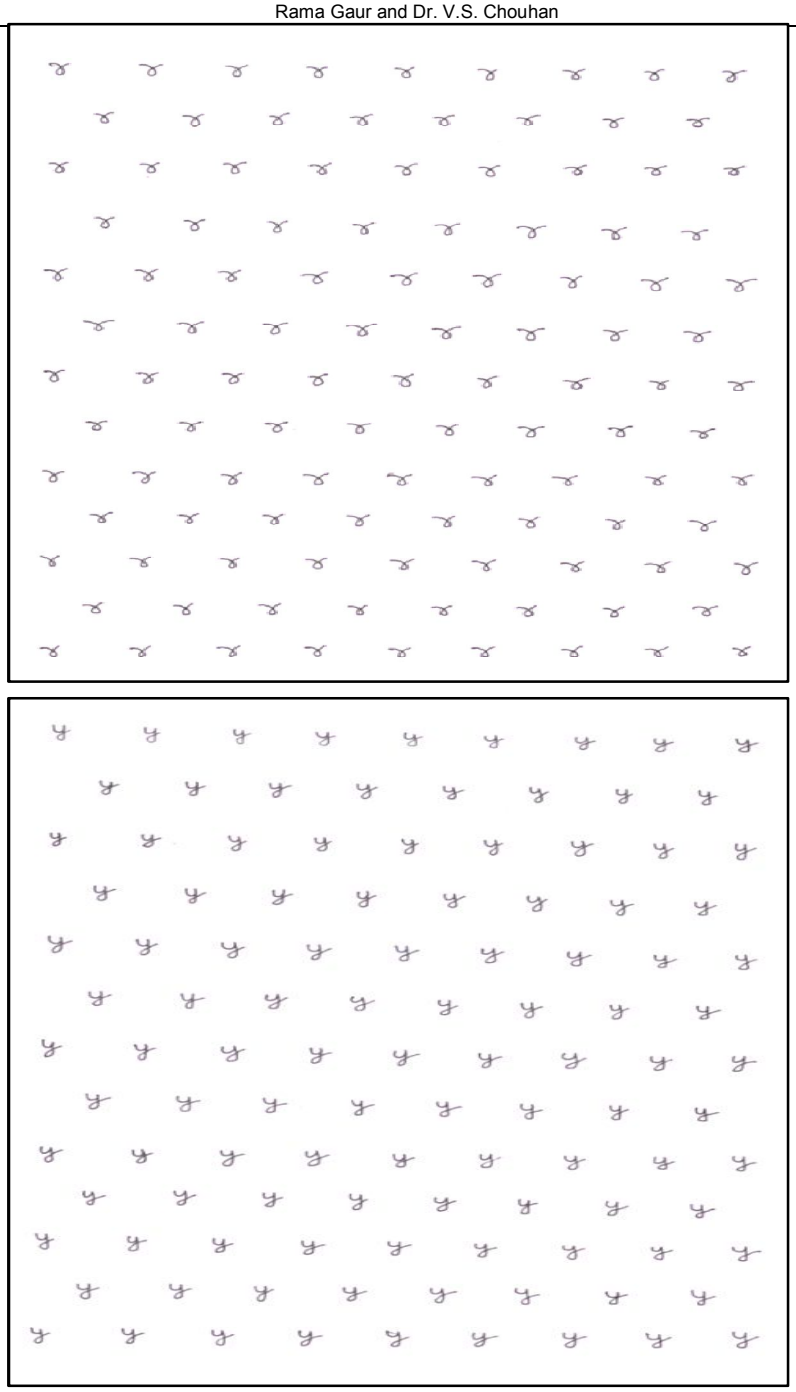

Fig. 1: Valid numeral database sheet.

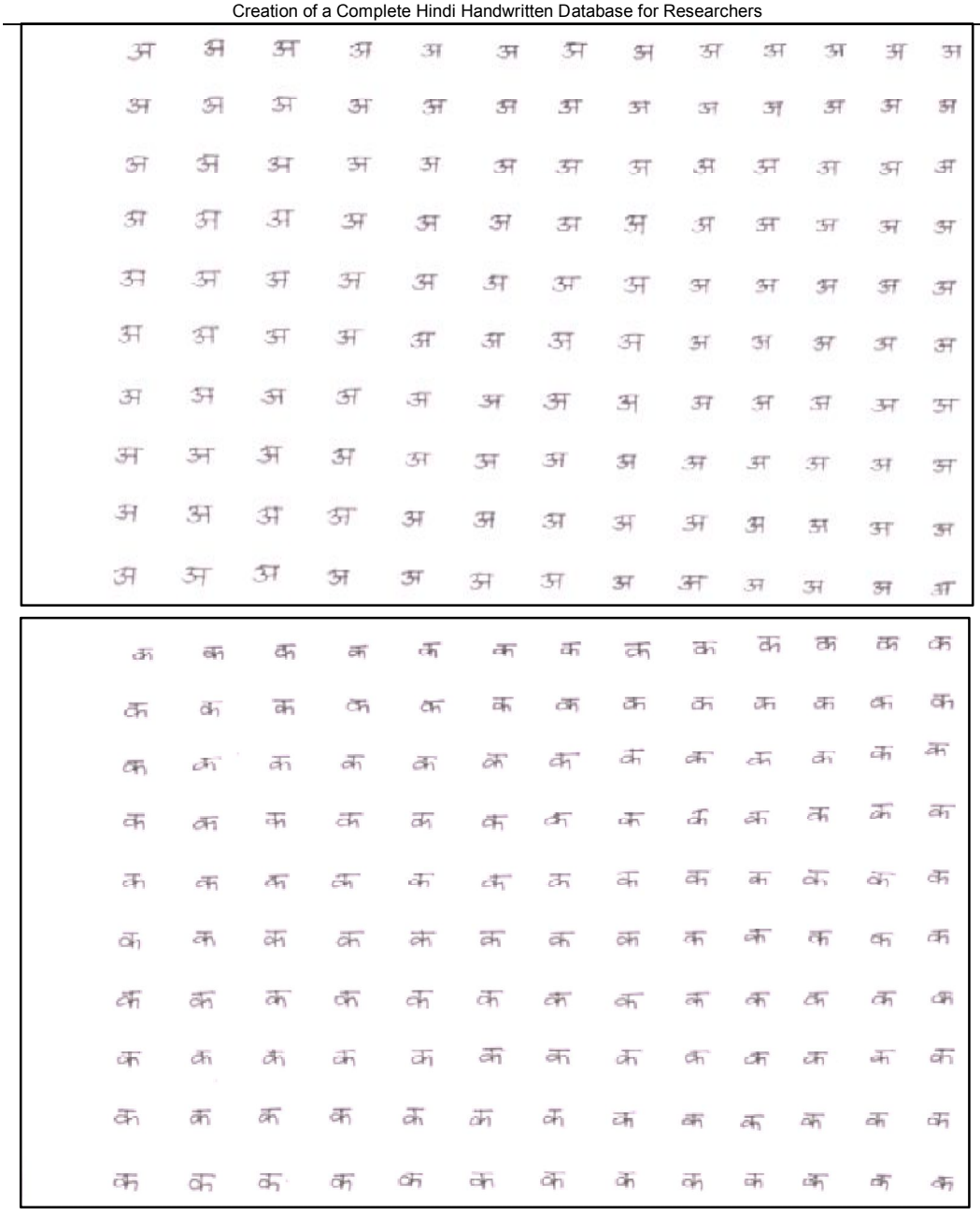

Fig.2: Valid character database sheets.

2.2. Data Preparation

The A4 size paper sheet having the data written by various writers shown in Figure 1 and Figure 2, is digitized using Canon Canoscan Lide 100 flatbed scanner at 300 dpi. The images were stored in JPG format. Now every character is cropped and stored as an image of jpeg format the size of every character image was 105X125 pixcels and the size of jpeg file varies from 12kB to25kB.The images were colored that can be converted to binary form as per requirement by simple matlab code of format conversion for saving the memory. In my case recognition algorithm code is doing this task. All the sample images of one character are stored in same folder. In such a way we create seprate folder for every numeral and character. A few samples of isolated numerals and characters from the present database are shown in Figure 3 and Figure 4. Various image symbol files are serially numbered for further convenient use. Table 1 shows size of numeral database for each numeral, and Table 2 shows size of database for each character.Useful characters segmented are stored in individual jpeg format files. The separated symbols are visually checked for proper shapes before sorting and storing in proper folders. 58 folders are formed for storing numeral and characters. A few samples of isolated numerals and characters from the present database are shown in Figure 3 and 4.

Fig.3: Cropped numerals from database sheet.

Fig. 4: Cropped character from database sheets.

3. STATISTICS OF DATA GENERATED

Some Devanagari compound characters are not widely used in modern writing. Some characters are written in more than one way. The database mostly contains first form of the numeral as it is written by most of the writers. The ideal Devanagari script consists of curves and connected lines. Lines are not isolated from main symbol. But in practice, the handwritten character and the strokes are involuntarily isolated due to incorrect writing of writers. This imposes serious problems in document segmentation and further recognition. Isolated strokes of modifiers are mistakenly considered as individual symbol in the character segmentation stage and thus stored separately. Correctly segmented numerals and characters are shown in Figure 3 and Figure 4. The incorrectly captured characters are rejected after visual inspection and removed from database. Some characters which are improperly written by writers are also rejected.In the entiresample sheets, all the symbols were processed. Due to the reasons mentioned in the previous paragraph, various databases differ in frequency as shown in Table 1 and Table 2.

The Figure5 shows the printed Hindi numerals and Figure 6 shows the Hindi Vowels whereas Figure 7 shows Hindi Consonants. These three Figuresshow complete character set of Hindi Devanagari Script.

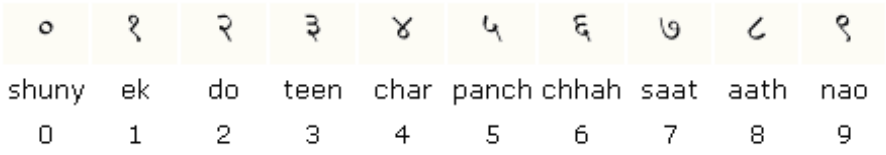

Fig.5: Hindi Numerals.

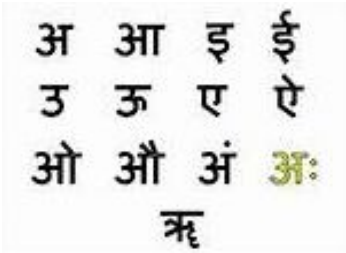

Fig.6: Hindi Vowels.

Fig.7: Hindi Consonants.

The statistics of constructed database is shown in the following Table 1 and Table 2. The size of the database is adequate to use by researchers.

Table1: Statistics of Hindi handwritten numeral database.

Database No.	Symbol	Frequency	Database No.	Symbol	Frequency
00	०	110	05	५	111
01	१	113	06	६	115
02	२	113	07	७	200
03	३	204	08	८	212
04	४	113	09	९	212
Total					**1503**

Table2: Statistics of Hindi handwritten character database.

Database No.	Symbol	Frequency	Database No.	Symbol	Frequency
10	अ	130	35	ड	130
11	आ	130	36	ढ	130
12	इ	130	37	ण	130
13	ई	130	38	त	130
14	उ	130	39	थ	130
15	ऊ	129	40	द	130
16	ए	130	41	ध	130
17	ऐ	130	42	न	120
18	ऑ	130	43	प	130
19	औ	130	44	फ	130
20	अं	130	45	ब	130
21	अः	130	46	भ	116
22	ऋ	130	47	म	130
23	क	130	48	य	130

24	ख	131	49	र	129
25	ग	130	50	ल	130
26	घ	130	51	व	130
27	ङ	130	52	श	140
28	च	130	53	ष	130
29	छ	130	54	स	130
30	ज	130	55	ह	130
31	झ	130	56	क्ष	116
32	ञ	126	57	ट्र	130
33	ट	130	58	ज्ञ	130
34	ठ	128			
TOTAL					**7385**

4. CONCLUSION AND FUTURE WORK

In this paper, we have generated a comprehensive database for Devanagari numerals and characters. Database of 1503 symbols is generated for numerals and database of 7385 symbols is generated for characters.

It is found that some symbols obtained need to be rejected as the writings of many persons are not recognizable by visual inspection. It will be impossible for computer software to recognize such symbols. The data images are stored in JPEG format for versatile needs. This database will be further grown with more samples from variety of writers. This database is more useful as training set, as I have reduced some complexity. The samples are almost of same size and have proper shape so that researchers have not to fight with poor database at the testing stage of algorithm. Once the algorithm tested, it can be examined by other database. This will surely help the research community for benchmarking their research results.

REFERENCES

[1] T. Saito, H. Yamada, and K. Yamamoto; “On the database ELT9 of hand printed characters in JIS Chinese characters and its analysis”, Transactions of the Institute of Electronics and Communication Engineers of Japan, Vol. J.68-D(4), pp. 757-764, 1985 (Japanese).

[2] B.B. Chaudhuri; “A complete handwritten numeral database of Bangla-a major Indic script”, CVPR Unit, Indian Statistical Institute, Kolkata, India.

[3] U. Bhattacharya and B.B. Chaudhuri; “Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(3), pp. 444-457, 2009.

[4] U. Bhattacharya and B.B. Chaudhuri; “Databases for research on recognition of handwritten characters of Indian scripts”, in Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 05), pp.789-793, September 2005.

[5] “Handwritten character databases of Indic scripts”, 2012. http://www.isical.ac.in/~ujjwal/download/database.html

[6] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D.K. Basu; “CMATERdb1: a database of unconstrained handwritten Bangla and Bangla-English mixed script document image”, International Journal on Document Analysis and Recognition, Vol. 15(1), pp. 71-83, 2012.

[7] M.P. Kumar, S.R. Kiran, A. Nayani, C.V. Jawahar, and P.J. Narayanan; “Tools for developing OCRs for Indian scripts”,Proceedings of the Computer Vision and Pattern Recognition Workshop (CVPRW ’03), pp. 33-38, and 2003.

[8] U. Pal and B.B. Chaudhuri; “Indian script character recognition: a survey”, Pattern Recognition, Vol. 37(9), pp. 1887-1899, 2004.

[9] Technology Development for Indian Languages, 2012.http://www.tdil.mit.gov.in/

[10] Center for Development of Advanced Computing, 2012.http://www.cdac.in/