Databases for Research on Recognition of Handwritten Characters of Indian Scripts

U. Bhattacharya and B. B. Chaudhuri
CVPR Unit, Indian Statistical Institute, Kolkata-108, India
{ujjwal,bbc}@isical.ac.in

Abstract

Three image databases of handwritten isolated numerals of three different Indian scripts namely Devnagari, Bangla and Oriya are described in this paper. Grayscale images of 22556 Devnagari numerals written by 1049 persons, 12938 Bangla numerals written by 556 persons and 5970 Oriya numerals written by 356 persons form the respective databases. These images were scanned from three different kinds of handwritten documents – postal mails, job application form and another set of forms specially designed by the collectors for the purpose. The only restriction imposed on the writers is to write each numeral within a rectangular box. These databases are free from the limitations that they are neither developed in laboratory environments nor they are non-uniformly distributed over different classes. Also, for comparison purposes, each database has been properly divided into respective training and test sets.

1. Introduction

Extensive experiments on recognition of off-line handwritten characters, in particular numerals, have been carried out during the last few decades because it has enormous application potentials [1]. Generally, existing techniques for recognition of handwritten characters are script dependent. A few reports include [2] for English, [3] for Chinese, [4] for Arabic, [5] for Korean and [6] for Kanji script and so on. However, enough research on recognition of handwritten characters/numerals of Indian scripts did not take place.

India is a multilingual country of more than 1 billion population with 18 constitutional languages and 10 different scripts. Although since 1965 English had been officially recognised as an "associated language" in India, according to the latest census report, less than 5 percent of the Indian population can either read or write English.

In a developing country like India there is an urgent need for the research and development of its own language technologies. The Department of Information Technology initiated the TDIL (Technology Development for Indian Languages) programme with the objective of developing information processing tools and techniques to facilitate human-machine interaction without language barrier. The two most popular languages in the Indian subcontinent are Hindi and Bangla and these are the fourth and fifth most popular languages of the world and used by about 500 million people. Bangla is also the official language of Bangladesh. The script of Hindi is Devnagari which is also used by languages like Nepali, Marathi etc. On the other hand, the script of Bangla is almost the same as two other scripts, *viz.*, Assamese and Manipuri. Oriya is another Indian language and script mainly used in its state of Orissa. Each of these three Indian scripts has their own native sets of numerals.

Recent but notable works on printed Indian scripts include [7] for Devnagari texts and [8] for Bangla OCR system. However, there exits only a few studies on handwritten characters of some Indian scripts which includes [9] and [10] for Bangla, [11] and [12] for Devnagari characters. These studies were reported on the basis of different databases collected either in laboratory environment or from smaller groups of the concerned population.

In fact, work on handwritten numerals/characters in Indian scripts lacks appropriate standard/benchmark databases. For Latin numerals there exist a number of such standard databases *viz.* NIST, MNIST [13], CEDAR [14], CENPARMI etc. Similar databases exist for a few other scripts also [15, 16, 17].

This paper describes an attempt for generation of moderately large and representative sample databases for numerals of three Indian scripts Devnagari, Bangla and Oriya. These may be obtained free of cost by a research group on request.

2. Data collection

Before collection of data, the following points were decided to make the databases as much representative as possible. Common factors responsible for variations in handwriting styles include age, sex, education, profession, writing instrument, writing surface, mood of the writer etc. Enough care was paid to include samples from at least ma-

Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)
1520-5263/05 $20.00 © 2005 IEEE

IEEE COMPUTER SOCIETY

jor categories under each of the above issues. 776 mail pieces had been collected and samples of isolated handwritten numerals were extracted from the postal code part of their addresses. These samples provide true real-life data. However, such data collected from real-world has its own drawbacks. These samples cannot be evenly distributed among the ten possible classes and more seriously it cannot include all the sections of the population. So, a job application form (Figure 1) is considered and its three fields, viz. age, date-of-birth and the pin-code are used for extraction of required numeral data. This second choice also cannot completely reduce the difficulty of uneven distribution of samples among possible 10 classes. So, as the third and last option we designed a form (Figure 2) consisting of horizontally arranged strings of adjacent rectangular boxes. In this form numerals are written sequentially along each horizontal string of boxes. A subject was requested to write one character per box. No other restriction was imposed on the writers. The purpose of data collection had not been disclosed to them so that they should produce samples reflecting their natural handwriting styles. In approximately 75% cases, the same subject was asked to write on both forms on two different occasions using his/her own writing instrument. In case writing instrument was not available with the subject, it was supplied at random from a set of different types of such instruments. Both forms were printed on papers of different quality and the samples have been collected over a span of more than two years through the students of Degree Engineering Colleges as a part of their training programme.

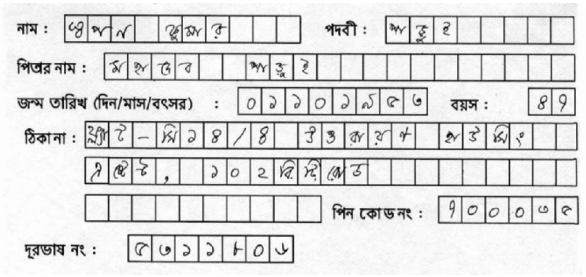

Figure 1. Job application form for data collection

3. Data preparation

Manual extraction of isolated numerals from the scanned images of the filled-in forms involve huge amount of man-hours. To save time a software has been developed for the identification of numerals. This software can automatically identify each horizontal strings of adjacent rectangular boxes and also each box within such a string by detecting horizontal and vertical thick lines. The numeral image component within such a box is obtained and it is

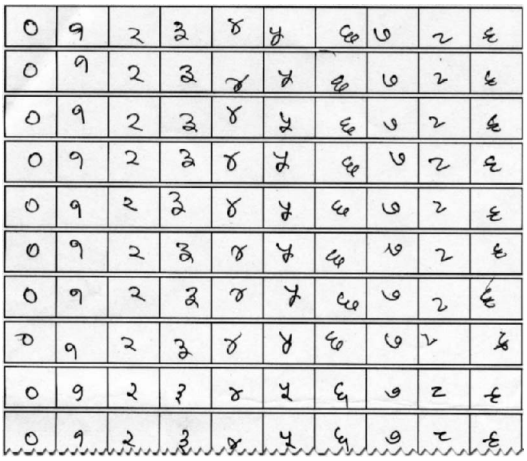

Figure 2. Specially designed form for data collection

extracted after allowing extra margins of a few rows and columns on each of its four sides. Difficulty arises only if a numeral touches/crosses the horizontal or vertical line of the box. The situation is tackled by analyzing the grey values in the neighbourhood of the pixels in the overlapping region. Since such a software is bound to produce some erroneous results, all the TIF files of individual numeral images have been checked manually and necessary manual corrections were made using an image editor.

4. Data definition

The forms are scanned at 300 d.p.i. resolution using a state-of-the-art HP flatbed scanner. These are stored as grayscale images using 1 byte per pixel. This may help the researchers to experiment with various preprocessing techniques including thresholding or recognition in the grayscale domain. Since comparison of approaches by different research groups on such a pattern recognition problem is very important, we have precisely split the whole sets of available numeral data on each of the three scripts into respective training and test sets. In certain cases, the researchers require the use of a validation set of samples in addition to the above two sets. Since requirement of this validation set depends on the training strategy, we do not exclusively provide such a validation set but the researchers may partition the training set to obtain this set. Since a larger training set of handwritten data often found yielding better recognition accuracy, we randomly divide the whole available data approximately in the ratio 5:1 for partitioning it into training and test sets.

Often one or more digits are repeated in the pin code field of a mail piece or the numeric fields of a job applica-

IEEE COMPUTER SOCIETY

tion form. In such cases, the concerned samples have been verified manually and if they are found almost similar, only one was included into the respective databases.

4.1. Devnagari numerals

Devnagari descended from the Brahmi script sometime around the 11th century AD. Its original form was developed to write Sanksrit but was later adapted to write many other languages such as Hindi, Marathi and Nepali. The ideal (printed) Devnagari numerals are shown in Figure 3. From this figure it is seen that there are variations in the shapes of numerals 5, 8 and 9 in their printed forms. However, from the samples in Figure 4 it can be observed that there exist wide variations in the handwritten forms of Devnagari numerals.

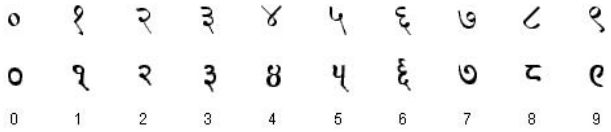

Figure 3. Devnagari numeral shapes

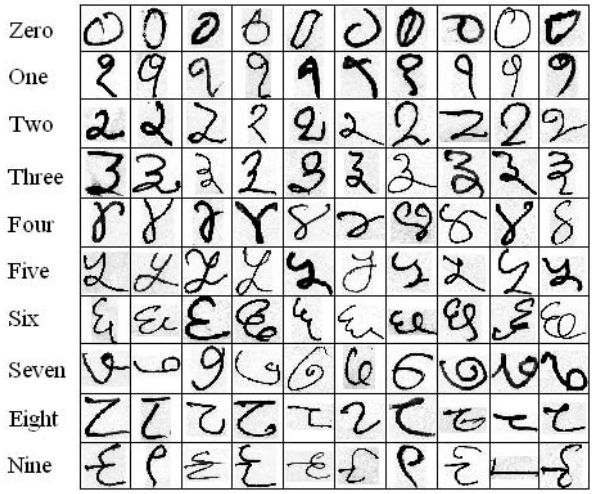

Figure 4. Handwritten Devnagari numeral samples

Our database of isolated handwritten Devnagari numerals consists of 22556 samples from 1049 persons. This database was formed from 368 mail pieces, 274 job application forms and for the rest we used the form in Figure 2. The whole set of available data have been split into a training set consisting of 18794 samples and a test set consisting of 3762 samples. The distribution of samples of these training and test sets into 10 classes are given in Table 1.

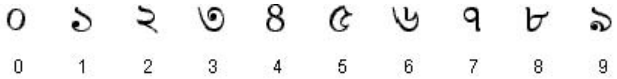

Figure 5. Bangla numeral shapes

4.2. Bangla numerals

Bangla is also derived from the Brahmi script and it started to diverge from the Devnagari script during the 11th Century AD. This script is also used to write a few other languages such as Assamese, Manipuri etc. The ideal (printed) forms of Bangla numerals are shown in Figure 5. However, the samples shown in Figure 6 give an idea about the variations in the handwritten forms of Bangla numerals.

Table 1: Distribution of numerals in Devnagari database

Digits	Training Set	Test Set	Total
0	1842	369	2211
1	1890	378	2268
2	1890	378	2268
3	1881	377	2258
4	1875	375	2250
5	1888	378	2266
6	1868	374	2242
7	1868	378	2264
8	1886	377	2262
9	1885	378	2267
Total	18794	3762	22556

Our database of isolated handwritten Bangla numerals consists of 12938 samples from 556 persons. This database was formed from 303 mail pieces, 116 job application forms and for the rest we used the third type of form described above. The whole set of available data have been split into a training set consisting of 7938 samples and a test set consisting of 5000 samples. The distribution of samples of these training and test sets into 10 classes are given in Table 2. This database has been used in a recent work on recognition of Bangla numerals [18].

4.3. Oriya numerals

The Oriya script developed from the Kalinga script, one of the many descendants of the Brahmi script of ancient India. The ideal (printed) forms of Oriya numerals are shown in Figure 7. A set of 100 samples of handwritten isolated Oriya numerals are shown in Figure 8.

IEEE COMPUTER SOCIETY

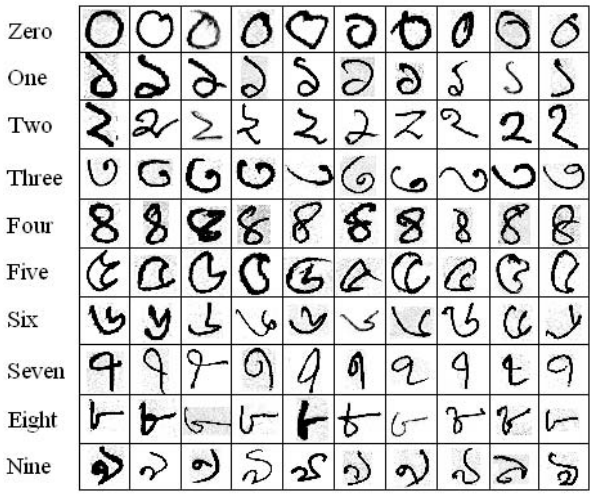

Figure 6. Handwritten isolated Bangla numeral samples

Table 2: Distribution of numerals in Bangla database

Digits	Training Set	Test Set	Total
0	928	500	1428
1	851	500	1351
2	871	500	1371
3	779	500	1279
4	815	500	1315
5	685	500	1185
6	803	500	1303
7	773	500	1273
8	753	500	1253
9	680	500	1180
Total	7938	5000	12938

Our database of isolated handwritten Oriya numerals consists of 5970 samples from 356 persons. This database was formed from 105 mail pieces, 166 job application forms and for the rest we used the third type of form described above. The whole set of available data have been split into a training set consisting of 4970 samples and a test set consisting of 1000 samples. The distribution of samples of these training and test sets into 10 classes are given in Table 3.

5. Summary

In this article, a detailed description of newly developed databases of handwritten isolated numerals of three different Indian scripts has been provided. These three databases form a very important infrastructure to develop and compare various recognition schemes for handwritten Devnagari, Bangla and Oriya numerals. A few unique characteristics of these databases are (i) they include real-life data collected from mail pieces and job application forms (ii) they maintain the balance of representations of different classes; numerals have also been collected using special simple forms for this purpose (iii) all the data are stored in gray-scale using TIF format providing maximum possible information. Since the data in these databases are not pre-processed ones, one has the freedom to play in this stage also.

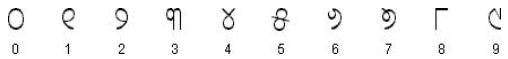

Figure 7. Oriya numeral shapes

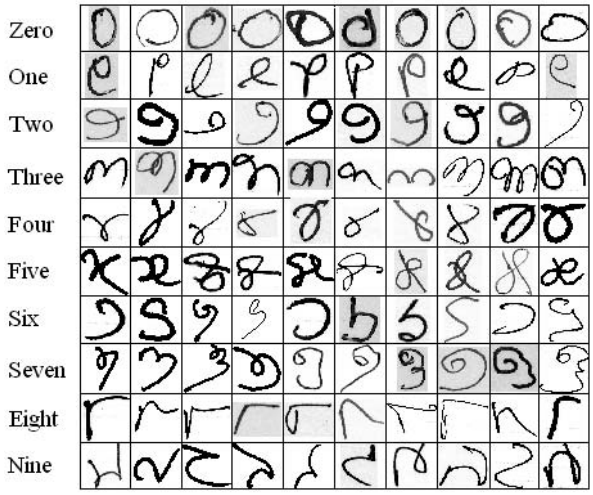

Figure 8. Handwritten isolated Oriya numeral samples

Table 3: Distribution of numerals in Oriya database

Digits	Training Set	Test Set	Total
0	502	100	602
1	498	100	598
2	496	100	596
3	498	100	598
4	498	100	598
5	490	100	590
6	500	100	600
7	498	100	598
8	498	100	598
9	492	100	592
Total	4970	1000	5970

IEEE COMPUTER SOCIETY

In the mean time, we have tested several recognition schemes on the above described databases and obtained significant recognition accuracies. Interested readers may consult articles [18], [19] and [20].

Presently, we are also generating a large database of isolated handwritten characters of the complete set of alphabets of Bangla, the second major Indian script. This database includes Bangla Basic, Vowel modifiers and Compound characters. This will definitely initiate research activities in handwritten form processing of Indian scripts. More and latest details of these databases are available at the http://www.isical.ac.in/~ujjwal/. Initial recognition works on the database of Bangla basic characters include [21] and [22].

Acknowledgements: The authors wish to acknowledge the supports of a number of students of different Universities and Institutions of W. Bengal, India who worked hard in the various stages of generation of these databases.

References

[1] C. Y. Suen, M. Berthod and S. Mori. Automatic recognition of handprinted characters – the state the art. *Proceedings of the IEEE*, Vol.**68(4)**, pp. 469-487, 1980.

[2] S. N. Srihari, E. Cohen, J. J. Hull and L. Kuan. A system to locate and recognize ZIP codes in handwritten addresses. *IJRE*, Vol. **1**, pp. 37-45, 1989.

[3] J. Tsukumo and H. Tanaka. Classification of handprinted Chinese characters using nonlinear normalization methods. *9th. Int. Conf. Pattern Recognition*,pp. 168-171, 1988.

[4] A. Amin and H. B. Al-Sadoun. Hand printed Arabic character recognition system. In *Proc. of the 12th. ICPR*, pp. 536-539, 1994.

[5] S. W. Lee and J. S. Park. Nonlinear shape normalization methods for the recognition of large-set handwritten characters. *Patt. Recog.*, Vol. **27**, pp. 895-902, 1994.

[6] H. Yamada, K. Yamamoto and T. Saito. A non-linear normalization method for handprinted Kanji character recognition – line density equalization. *Patt. Recog.*, Vol. **23**, pp. 1023-1029, 1990.

[7] V. Bansal and R. M. K. Sinha. Integrating knowledge sources in Devnagari text recognition system. *IEEE Trans. Syst. Man & Cybern.* Vol. **SMC-A 30**, pp. 500-505, 2000.

[8] B. B. Chaudhuri and U. Pal. A Complete Printed Bangla OCR System. *Pattern Recognition*. Vol. **31**, pp. 531-549, 1998.

[9] A. Dutta and S. Chaudhuri. Bengali alpha-numeric character recognition using curvature features. *Pattern Recognition*, Vol. **26**, pp. 1757-1770, 1993.

[10] A. F. R. Rahman, R. Rahman and M. C. Fairhurst. Recognition of handwritten Bengali characters: A novel multistage approach. *Pattern Recognition*, Vol. **35**, pp. 997-1006, 2002.

[11] S. D. Connell, R. M. K. Sinha and A. K. Jain. Recognition of Unconstrained On-line Devnagari Characters. *Proc. of Int. Conf. on Patt. Recog.*, Vol. **II**, pp. 368-371, 2000.

[12] R. Bajaj, L. Dey and S. Chaudhuri. Devnagari numeral recognition by combining decision of multiple connectionist classifiers. *Sadhana* Vol. **27**, Part 1, pp. 59 - 72, 2002.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86(11), pp.2278-2324, 1998.

[14] J. J. Hull. A Database for Handwritten Text Recognition Research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **16**, pp. 550-554, 1994.

[15] T. Saito, H. Yamada and K. Yamamoto. On the database ELT9 of handprinted characters in JIS Chinese characters and its analysis (in Japanese). *Transactions of IECEJ*, Vol. **J.68-D(4)**, pp. 757-764, 1985.

[16] Y. Al-Ohali, M. Cheriet, C. Suen, Databases for recognition of handwritten Arabic cheques. *Pattern Recognition*, Vol. **36**, pp. 111-121, 2003.

[17] T. Noumi, T. Matsui, I. Yamashita, T. Wakahara and T. Tsutsumida. Tegaki Suji Database 'IPTP CD-ROM1' no Ichi Bunseki (in Japanese). *1994 Autumn Meeting of IEICE, D-309*, September 1994.

[18] U. Bhattacharya, T. K. Das, A. Datta, S. K. Parui and B. B. Chaudhuri. A Hybrid Scheme for Handprinted Numeral Recognition Based On a Self-Organizing Network and MLP Classifiers. *IJPRAI*, Vol. **16(7)**, pp. 845-864, 2002.

[19] U. Bhattacharya, T. K. Das and B. B. Chaudhuri, A cascaded scheme for recognition of handprinted numerals, Proceedings of the 3^{rd} ICVGIP, Ahmedabad, India, 2002, pp. 137 - 142.

[20] U. Bhattacharya, B.B. Chaudhuri, "A Majority Voting Scheme for Multiresolution Recognition of Handprinted Numerals", Proc. of the 7th ICDAR, Edinburgh, Scotland, vol. I, page: 16-20, 2003.

[21] S. K. Parui, T. K. Bhowmik and U. Bhattacharya, "A novel scheme for extraction of shape descriptions from handwritten Bangla characters", Proc. WCVGIP-2004, Gwalior (M.P.), India, pp. 1-4, 21-22 Feb., 2004.

[22] T. K. Bhowmick, U. Bhattacharya and S. K. Parui, "Recognition of Bangla Handwritten Characters Using an MLP Classifier Based on Stroke Features", ICONIP 2004, LNCS 3316, (Eds. N. R. Pal et al.), pp. 814-819, 2004.

IEEE COMPUTER SOCIETY