A Complete Handwritten Numeral Database of Bangla – A Major Indic Script

B.B. Chaudhuri

A Complete Handwritten Numeral Database of Bangla – A Major Indic Script

B. B. Chaudhuri
CVPR Unit, Indian Statistical Institute, Kolkata-108, India
bbc@isical.ac.in

Abstract

This paper describes the ISI database of handwritten Bangla numerals. Bangla is the second most popular language and script of the Indian subcontinent and it is used by more than 200 million people all over the globe. The present database has several components which include both on-line and off-line handwritten numerals. Samples of numeral strings and isolated numerals have been collected under both modes of writing. This database has been developed at the Computer Vision and Pattern Recognition Unit laboratory of Indian Statistical Institute, Kolkata. Samples of the present database are properly ground thruthed and subdivided into respective training and test sets. The off-line sample images are stored in TIFF image format and the on-line samples are stored along with various information as header in ASCII file format. This database will facilitate fruitful research on handwriting recognition of Bangla through free access to the researchers.

Keywords: Handwritten character database, on-line database, off-line database, numerals of Indian scripts

1. Introduction

There is a special importance of numerals in automatic handwritten character recognition research. This is so because numerals carry more important information in application areas like bank check or postal address reading systems as well as automatic tabular form processing. So, the document processing community puts emphasis on recognizing numerals with a high degree of accuracy. To facilitate results on uniform data set, several document processing research groups have collected large numeral databases to make them available to the fellow researchers. However, such existing databases are limited to only a few scripts of developed nations such as English, Japanese, Chinese. These standard databases include NIST, MNIST [1], CEDAR [2], CENPARMI etc. for Latin numerals and [3, 4, 5] for a few other scripts.

India is a multilingual country of more than 1 billion population with 22 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 in English. Bangla is a popular language and script of the Indian subcontinent which is second to Devanagari. Bangla is also the official language of Bangladesh. It is used by more than 200 million people all over the world. The script of Bangla is almost the same as two other Indic scripts, *viz*., Assamese and Manipuri.

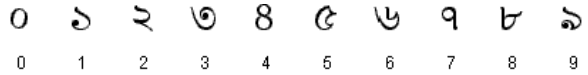

Figure 1. Bangla numeral shapes

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, Govt. of India 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. As a result of such initiatives, recent thrusts have been observed on research attempts for automatic recognition of printed / handwritten characters of various Indic scripts.

Some recent but notable works on printed Indian scripts include [6] for Devnagari texts and [7] for Bangla OCR system. However, there exits only a few studies on handwritten characters of some Indian scripts which includes [8] and [9] for Bangla, [10] and [11] 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, effective research work on handwriting recognition for Indic scripts is seriously hampered because of the unavailability of standard/benchmark databases those may be used for testing of algorithms and for comparison of results. This paper describes a pioneering attempt for generation of a complete database for handwritten Bangla numerals. This new handwritten database is complete in the sense that it consists of a large number of samples of on-line/off-line and also segmented / connected numerals. This database has been developed to make it available freely to the fellow researchers for the furtherance of handwriting recognition research on Indic scripts. The ideal (printed) forms of Bangla numerals are shown in Figure 1.

In Section 2 of this article, we describe details of on-line annotated database for handwritten Bangla numerals

and Section 3 provides description of off-line handwritten Bangla numeral database. Section 4 summarizes the present work.

2. Bangla on-line numeral data

To the best of our knowledge, the only existing on-line handwritten database in an Indic script is the Tamil data in UNIPEN format [12] and a major competition during IWFHR-10 on on-line character recognition using this database has been already announced [13]. On the other hand, pen based computing devices are gaining popularity even in India and the present paper describes the first ever attempt towards the development of an annotated on-line handwritten data set in Bangla, the second most popular Indian script. This database and the tool used for this purpose have been recently developed at our laboratory. The same tool will be used in near future for the development of similar databases of a few other Indian scripts. Such attempts will cause further research activities on recognition of on-line handwritten characters.

2.1. Collection of On-line numeral samples

The on-line numeral samples of the present database have been collected using Genius NewSketch 1212 tablet. The sampling rate is 150 points per second. No restriction was imposed on the writers except for specifying rectangular regions for writing either isolated numeral or its string of different sizes. Since such rectangular regions are large enough, the restriction may not be considered as a serious one.

A software has recently been developed at our laboratory for speedy collection, automatic ground truthing and annotation of on-line handwritten data. A screen shot of this software is shown in Figure 2. This software, written using MSVC++ under Widows XP environment, provides a screen with an on-line form as the header and large rectangular regions for writing isolated or numeral strings.

2.2. Format of the annotation

Now-a-days XML is a popular choice for representation of annotated on-line data [14]. Recently, an *XML* representation for annotation of online handwriting data using *InkML* standard had been described in [15]. Although the present database has not been developed in XML representation, however, it stores in ASCII most of the significant information stored in UNIPEN format or XML representation of raw handwriting data called Digital Ink Markup Language (InkML), a standard being developed by the W3C for the description of digital ink [16].

The format used for the present on-line database encodes information about the version number, date and time, writer's name, age, sex, education, occupation, city, script, mother tongue, writing habit (right-handed or left-handed), number of samples generated, coordinates of two extreme and opposite corner points of both the frame provided to write and the minimum bounding rectangle for pen-down positions.

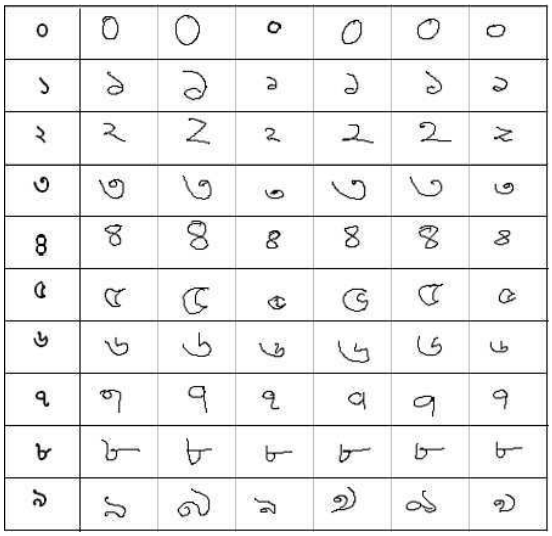

Figure 3. Samples from the database of On-line handwritten Bangla isolated numerals

At the start of each data collection session, the software provides option for isolated or numeral string. If numeral string is selected, it further provides option for selection of the string length. Depending on these parameters, the screen is divided into enough large rectangular regions inside each of which, the writer should provide samples. Such a tool helps automatic ground truthing. Handwritten sample within each rectangular region is stored in distinct files and filenames are generated from the name of the writer. During generation of filename, the software also look for possible existence of samples from the same writer under the target directory and accordingly concatenate a counter value with the filename. All the samples of the present database collected through the said software have been manually checked to ensure reliable ground truth.

Table 1: Distribution of samples in on-line Bangla numeral database

Digits	Training Set	Test Set	Total
0	700	236	936
1	700	230	930
2	700	235	935
3	700	235	935
4	700	233	933
5	700	236	936
6	700	236	936
7	700	235	935
8	700	236	936
9	700	236	936
Total	7000	1348	8348

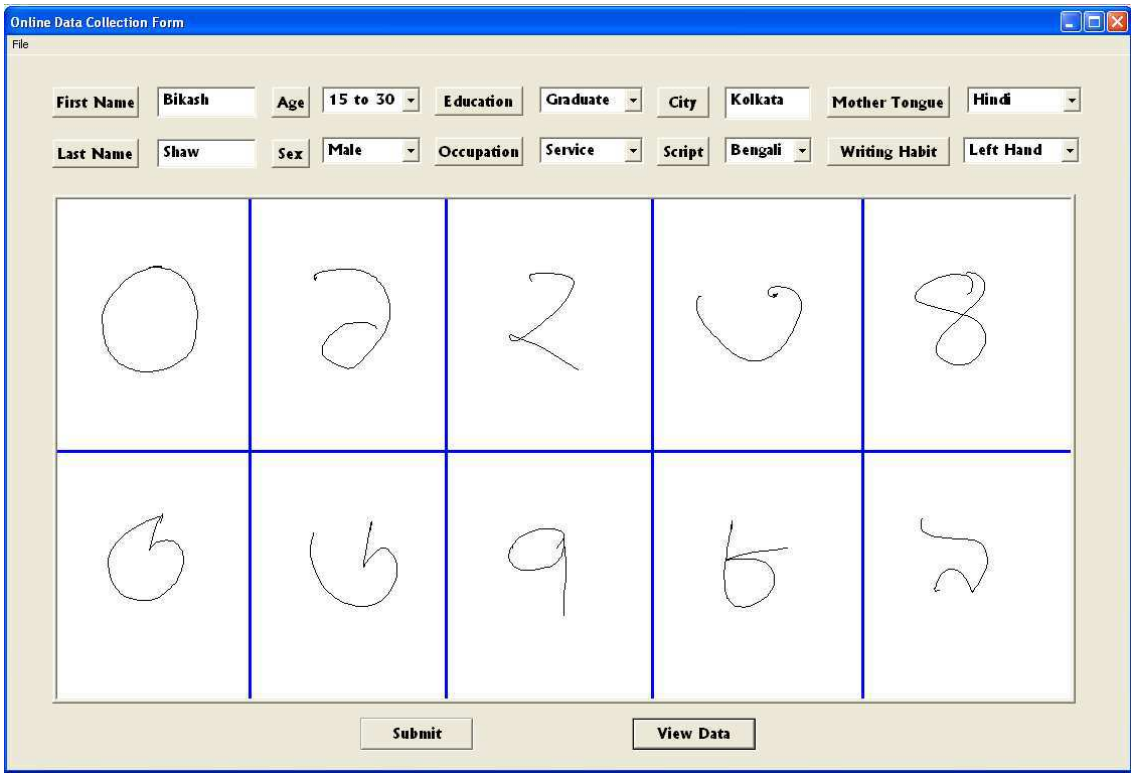

Figure 2. On-line data collection form

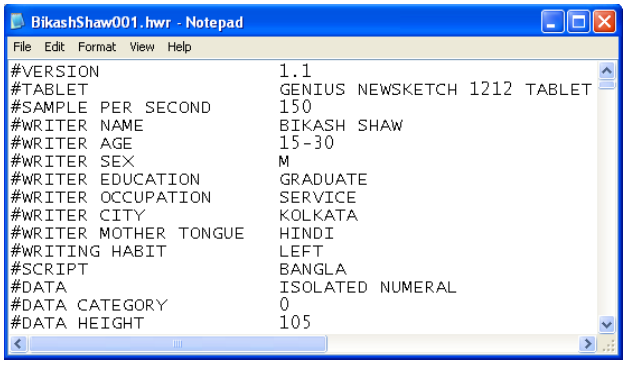

Figure 4. Part of the header of a sample data file

2.3. Online Bangla numeral database details

A sample set of online isolated Bangla numerals are shown in Figure 3. The following listing in Figure 4 shows an example of a part of the header information of a sample file from the present ISI database of on-line isolated numeral samples. Also, a listing of the data portion from a sample of our database is shown in Figure 5. Distribution statistics of on-line isolated numeral sample data is provided in Table 1.

The on-line part of the present database also consists of numeral strings of various lengths. Most of these strings consists of 2 numerals and there are 1265 such numeral strings. The maximum number of numerals in a on-

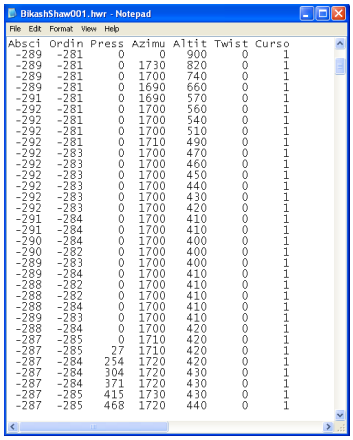

Figure 5. A portion of the data in a sample data file

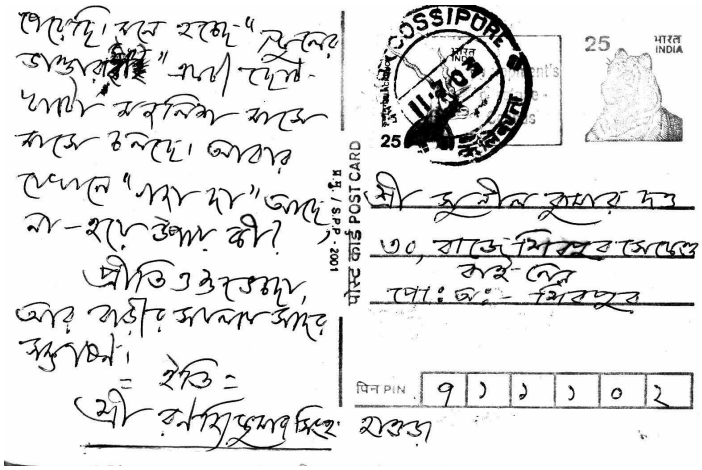

Figure 6. A sample postal mail piece used for collection of off-line numeral samples

line numeral string of our database is 10 and there are 157 such numeral strings. Total number of numeral strings in the on-line part of the present database is 2876.

3. Bangla off-line numeral data

The present handwritten numeral database for Bangla includes off-line samples of both isolated numerals and numeral strings. Descriptions of both of these components of the present database are given below.

Collection of off-line data is easier compared to the same for on-line data. This is because off-line data may easily be collected by scanning huge quantity of available paper forms including postal documents. However, preparation of the data including its ground truthing is labor intensive and time consuming.

3.1. Collection of off-line numeral samples

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 major categories under each of the above issues. Handwritten Bangla isolated numeral samples were extracted from the postal code part of the address written on a large number of mail pieces. Bangla numeral strings were also extracted from these mail pieces whenever these are found on the real-life mail pieces in our collection. One such mail piece in which both numeral string and isolated numerals are found is shown in Figure 6. 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 numeral classes and more seriously it cannot include all the sections of the population. So, a job application form (Figure 7) 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 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.

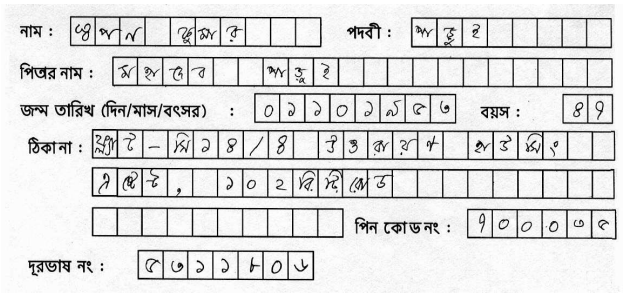

Figure 7. Job application form for data collection

3.2. Data preparation

Manual extraction of isolated numerals from the scanned images of the filled-in forms or postal mails involve huge amount of man-hours. So, we developed a software for automatic extraction of numerals from the scanned forms or from the string of rectangular boxes of pincode part of scanned postal mails. The Postcard, Inland Letter and Envelopes sold by the Indian Post contain an array of six boxes printed in the address region. The senders normally write the six digit pincode number inside the boxes of this array.

The software developed by us for automatic extraction of handwritten numerals from images of forms or postal mails, locate strings of rectangular boxes in such images by identifying a pair of parallel lines which are cut by several vertical lines. For the above reasons, our software use digital horizontal/vertical line detection algorithms. The extracted data by the said software is manually checked and only the properly extracted numeral shapes are identified to add them to our database. Those mail pieces or form images which fail the box detection, are subjected

for manual extraction of numeral samples.

Here, it may be noted that the numeral data arising from mail pieces have wide variety of background arising from different paper quality and variations in color or gray shades. However, we did not maintain any color information in our database.

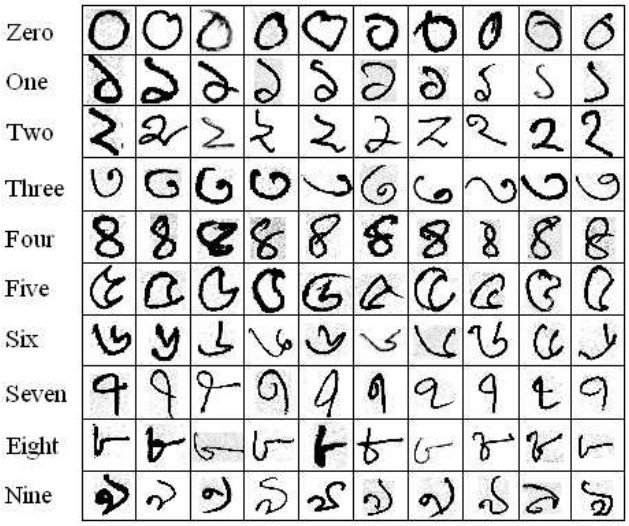

Figure 8. Off-line isolated Bangla numeral samples

3.3. Off-line Bangla numeral database details

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. A few samples of isolated numerals from the present database are shown in Figure 8. Also, several samples of off-line Bangla numeral strings are given in Figure 9.

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 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 obtaining 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 application 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.

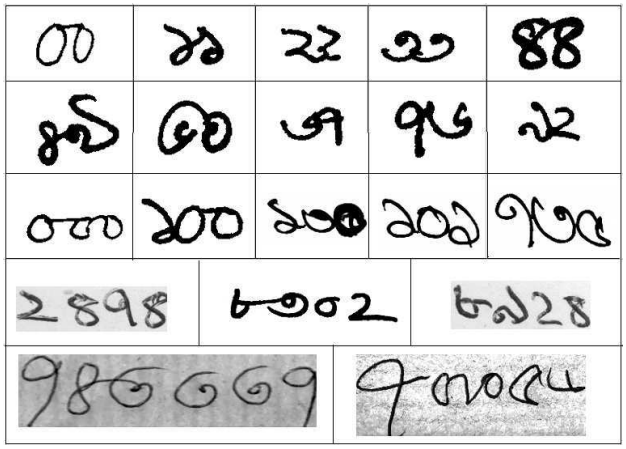

Figure 9. Samples of off-line Bangla numeral strings

3.4. Statistics of off-line handwritten Bangla numeral samples

Our database of isolated handwritten Bangla numerals consists of 23392 samples. These database were collected from 1202 mail pieces, 116 job application forms and for the rest we used the third type of form described above. The whole set of available isolated numeral data have been split into a training set consisting of 19392 samples and a test set consisting of 4000 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 [17].

Lengths of off-line samples of numeral strings in the present database varies between 2 to 10. However, most of these strings are of 2 numerals and there are 3156 such numeral strings. The second maximum number of numeral string samples are of length 10 and there are 954 such numeral strings. Total number of numeral strings in the off-line part of the present database is 6543.

Table 2: Distribution of numerals in Bangla database

Digits	Training Set	Test Set	Total
0	1933	400	2333
1	1945	400	2345
2	1945	400	2345
3	1956	400	2356
4	1945	400	2345
5	1933	400	2333
6	1930	400	2330
7	1928	400	2328
8	1932	400	2332
9	1945	400	2345
Total	19392	4000	23392

4. Summary

In this article, a detailed description of a newly developed complete database of Bangla numeral samples written in both on-line and off-line modes has been provided. This database should provide a very important in-

frastructure towards development and comparison of various schemes for recognition of handwritten Bangla numerals. A few unique characteristics of these database are (i) they include off-line real-life data collected from mail pieces and job application forms (ii) the on-line data are properly annotated and can be easily converted to the UNIPEN or XML format, (iii) they maintain the balance of representations of different classes (iii) off-line samples are stored in gray-scale using TIF format providing maximum possible information. Since the data in these databases are not preprocessed ones, one has the freedom to play in this stage also.

In the mean time, we have tested several recognition schemes during the initial stages of the development of off-line isolated Bangla numeral and obtained significant recognition accuracies. Interested readers may consult articles [17], [18] and [19].

Acknowledgements: I like to thankfully acknowledge U. Bhattacharya, my colleague at the Indian Statistical Institute, Kolkata for his all-round help towards the present work. Partial Support in the form of Jawaharlal Nehru Fellowship from JNFF is also gratefully acknowledged.

References

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

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

[3] 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), 1985, pp. 757-764.

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

[5] 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.

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

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

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

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

[10] 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, 2000, pp. 368-371.

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

[12] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman and S. Janet, "UNIPEN project of on-line data exchange and recognizer benchmarks", In *Proc. of the 12th International Conference on Pattern recognition*, Jerusalem, 1994, pp. 29-33.

[13] http://algoval.essex.ac.uk:8080/iwfhr2006/index.jsp, IWFHR 2006 Online Tamil Handwritten Character Recognition Competition.

[14] A. P. Lenaghan, R. R. Malyan, "XPEN: An XML Based Format for Distributed Online Handwriting Recognition", *Proc. of the International Conference on Document Analysis and Recognition*, pp. 1270 - 1274, 2003.

[15] A. S. Bhaskarabhatla, S. Madhvanath, M.N.S.S.K. Pavan Kumar, A. Balasubramanian and C. V. Jawahar, "Representation and Annotation of Online Handwritten Data", *Proc. 9th IWFHR*, 2004, pp. 136–141.

[16] W3C Multi-modal Interaction Working Group. Ink markup language (inkml). http://www.w3.org/2002/mmi/ink, 2003.

[17] 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), 2002, pp. 845-864.

[18] 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.

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