J Inf Process Syst, Vol.11, No.3, pp.325~341, September 2015
http://dx.doi.org/10.3745/JIPS.02.0008

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Dataset of Online Handwritten Assamese Characters

Udayan Baruah* and Shyamanta M. Hazarika**

Abstract

This paper describes the Tezpur University dataset of online handwritten Assamese characters. The online data acquisition process involves the capturing of data as the text is written on a digitizer with an electronic pen. A sensor picks up the pen-tip movements, as well as pen-up/pen-down switching. The dataset contains 8,235 isolated online handwritten Assamese characters. Preliminary results on the classification of online handwritten Assamese characters using the above dataset are presented in this paper. The use of the support vector machine classifier and the classification accuracy for three different feature vectors are explored in our research.

Keywords

Assamese, Character Recognition, Dataset Collection, Data Verification, Online Handwriting, Support Vector Machine

1. Introduction

Online handwriting recognition is a major research topic because of emerging technologies, such as PDAs and tablet PCs. Online recognition refers to the methods and techniques dealing with the automatic processing of a character as it is written using a digitizer [1]. Development of any recognition system requires a standard dataset, which is used for the training and testing of the system. The acquisition and distribution of standard datasets have increasingly gained importance. Some examples of datasets in the online domain are the English dataset Pen-Based Recognition of Handwritten Digits [2], the French dataset IRONOFF [3], the Spanish dataset UJIpenchars [4], and the UNIPEN dataset [5]. All of which contain online handwriting data from various alphabets (including Latin and Chinese), signatures, and pen gestures.

Research on online handwriting recognition has also emerged for Indian scripts. Examples of a few Indian scripts where online handwriting recognition activities are being carried out include Tamil [6], Telugu [7], Bengali [8], Devanagari [9], and Gurmukhi [10]. Few Indian scripts have also reported standard datasets of online handwritten characters. But not all datasets of Indian scripts are known to be publicly available. Some examples of few datasets of online handwritings in the context of Indian scripts are, the HP Lab Tamil Dataset [11], the Bangla Numeral Dataset [12] and the Devanagari Character Dataset [13]. Assamese is a major script in the northeastern part of India. No online

Manuscript received July 08, 2013; first revision December 18, 2013; second revision March 13, 2014; accepted April 16, 2014; onlinefirst October 20, 2014.

Corresponding Author: Udayan Baruah (udayan.baruah10@gmail.com)
* Dept. of Information Technology, Sikkim Manipal Institute of Technology, Sikkim 737136, India (udayan.baruah10@gmail.com)
** Dept. of Computer Science and Engineering, Tezpur University, Assam 784028, India (shyamanta@ieee.org)

handwritten Assamese characters dataset is available in any standard online repository of datasets. Therefore, the major objective is to build a dataset of online handwritten Assamese characters. The characteristics of Assamese characters (including some distinct properties unique to it) are listed below.

a. Assamese writing has evolved from the ancient Indian script, Brahmi. Brahmi had various classes, and Assamese originated from the Gupta Brahmi script [14].
b. In Assamese there are 10 numeric characters and 52 basic alphabetic characters that consist of 11 vowels and 41 consonants.
c. An Assamese character set is unique in that is has a large number of conjunct consonants (Juktakkhor) of about 164-201 [14].
d. Certain vowels, consonants, and conjunct consonants have headlines called matra in Assamese.
e. There is no upper and lower case writing in Assamese. This makes Assamese characters distinct from English characters.

In this paper, we report on the development of a dataset of online handwritten Assamese characters. The online handwritten Assamese characters dataset reported in this paper contains a total of 8,235 online handwritten Assamese characters from 183 classes, which consist of Assamese numerals, basic alphabetic characters, and conjunct consonants. We also present the preliminary results on the classification of online handwritten Assamese characters.

The paper is organized as follows: Section 2 describes the acquisition of data, which includes discussions on an experimental set-up for data acquisition (including the data acquisition program and data acquisition protocol) and the inspection of collected data for errors. Section 3 provides a discussion on different attributes of the collected data. The experimental results of character recognition on the samples of collected data are presented in Section 4 and we present our final comments in Section 5.

2. Data Acquisition

Data acquisition is the collection of online handwritten samples from different writers. This involves the recording of samples on some handwriting input devices like PDAs, tablet PCs, and pen-tablets. These devices record the trajectory of the writing.

2.1 Experimental Setup for Data Acquisition

For the creation of an online handwritten Assamese characters dataset, online handwritten samples of 121 conjunct consonants, along with the 10 numeric characters and 52 basic alphabetic characters for Assamese [15], were collected. The total number of samples corresponding to each writer is 183 (= 52 basic alphabetic characters + 10 numerals + 121 conjunct consonants). Thus there are a total of 8,235 samples in the dataset written by 45 writers (= 45 × 183). Printed Assamese Numerals, Basic Alphabetic Characters (Vowels and Consonants), and a few instances of Conjunct Consonants (Juktakkhor) are shown in Table 1. Figs. 1–3 show samples of online handwritten Assamese characters (the images of all the Assamese characters given as reference shapes can also be downloaded along with the dataset from http://mlr.cs.umass.edu/ml/datasets/Online+Handwritten+Assamese+Characters+Dataset. The images of the Assamese characters in printed form are documented in the file 'Data_Table.pdf' in the dataset).

The online Assamese handwriting samples were collected on an i-ball 8060U pen-tablet that was connected to a laptop and its cordless digital stylus pen was used through a GUI. A screenshot of the GUI is shown in Fig. 4. The resolution of the pen-tablet was 2540 LPI. The process of data acquisition was guided by an experimental protocol. The experimental protocol describes the systematic steps for the collection of data, which minimize variations in terms of the place of writing of characters, writing environment, and writing postures. The samples of the dataset were collected from students, research scholars, and faculty members of Tezpur University. The writers were 45 volunteers belonging to different age groups ranging from 20–42 years old with the average age being that of 31.

Table 1. Printed assamese characters

Assamese numerals	**Assamese vowels**
০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯	অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ
Assamese consonants	
ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম য ৰ ল ৱ শ ষ স হ ক্ষ ড় ঢ় য় ৎ ঃ ঁ	
A few instances of assamese conjunct consonants (*Juktakkhor*)	
ক্ক ক্ল ক্ত ক্ম ক্ষ্ম ষ্ম প্ত স্ক স্ত ঙ্গ ন্ত ত্ত ল্ল ঞ্চ শু হ্ল স্ন ষ্ট ন্ধ স্ত ব্ ম্ভ ন্ধ	

2.1.1 Data acquisition program

The GUI in the data acquisition program shows a text input box on the screen along with some other controls. The size of the text input box is 4,392 × 4,868 points, where the coordinates are integer numbers ranging from 0 to 4,392 for X and 0 to 4,868 for Y, respectively. The value of X goes left-to-right and that of Y goes downwards, assuming that the origin (0, 0) lies on the left-top corner of the text input box. The acquisition program records the handwriting as a stream of (X, Y) coordinate points using the appropriate pen position sensor along with the pen-up/pen-down switching. The pressure level values are not recorded. A single click on an onscreen Save button saves the current content of the text input box. This also clears the current content of the text input box so that the writer can write the next character in the box. The unsaved content of the text input box can be cleared by using an onscreen Reset button. The data collection system is initialized to the recording mode by using an onscreen Start button. The recorded information of each online handwritten character is stored in a text file. Each file is named based on the pair (M, N). The text file 'M.N.txt' represents the character with the ID 'M' written by the writer with the ID 'N'. For instance, the file '132.10.txt' represents the character with the ID '132' written by the writer with ID '10'. The distribution of the dataset consists of 45 folders (one for each writer) and a 'Data_Table.pdf' file. This file contains information about the character ID (ID), character name (Label) and the printed shape of the character (Char). Each folder contains 183 text files corresponding to the 183 characters written by a single writer.

2.1.2 Experimental protocol for data acquisition

The online Assamese handwriting samples were collected in a laboratory environment. As writers may not be at ease with the writing interface and the electronic pen, they were instructed to practice by writing several characters in the text input box before the actual recording of characters started. After a writer became familiar with the online hand writing tools and the environment, his/her actual data recording process took place according to the stages M0 to M4. Each writer contributed with his/her handwriting samples in a single, uninterrupted session.

M0: The writer is made to sit in a relaxed position with the data collection hardware setup in from of him/her.

M1: The list of all 183 candidate characters is viewed by the writer.

M2: The writer initializes the data collection system. The corresponding writer's ID and the ID of the first character in the list are entered through the GUI, which is followed by a click on the onscreen Start button.

M3: (1) For each of the 183 characters (in the ascending order of the serial number of the characters in the list provided) starting at the first character, the following two consecutive steps are performed by the writer:

a: The character is written in the text input box of the GUI.

b: The content of the text input box is saved by using the onscreen Save button.

(2) If a mistake or unsatisfactory writing of any of the k (for k = 1, 2, 3,…,183) characters is noticed immediately after the character is written, but before it is saved, the following two consecutive steps are performed by the writer:

a: The content of the text input box is cleared by using the onscreen Reset button.

b: The steps of M3(1) are followed corresponding to the kth character.

M4: All of the collected 183 samples are stored in the corresponding data folder.

2.2 Variability in Writing Styles

The shape of a character varies from writer to writer. Different writers write the same character differently. Similarly, the writers tend to write at different locations in the text input box. Therefore, the characters of the same class are of variable shapes, variable sizes, and a variable range of (X, Y) coordinate values. Some collected samples written by different writers are shown in Fig. 1. The variability of writing patterns of the same character is shown in Fig. 2.

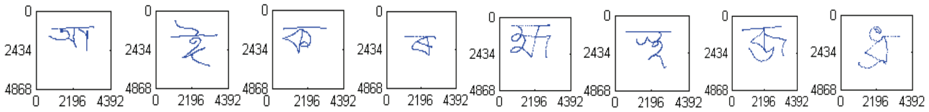


Fig. 1. Samples from the dataset of online handwritten Assamese characters written by different writers. In this figure we present samples from eight different writers.

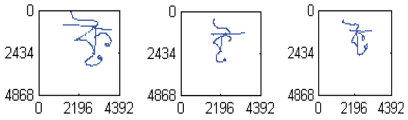


Fig. 2. Variability of patterns of the same character written by three different writers.

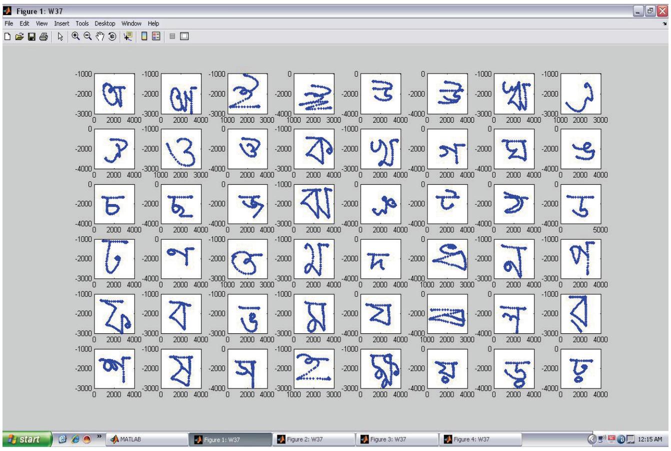
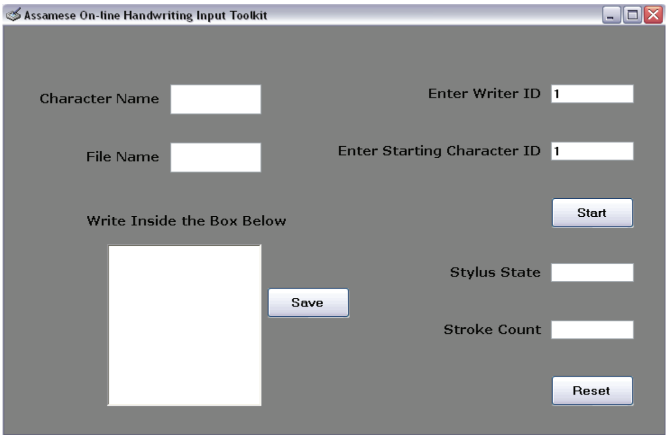


Fig. 3. A screenshot of a part of the Visual Data Inspection Program.


Assamese On-line Handwriting Input Toolkit
Character Name
File Name
Write Inside the Box Below
Save
Start
Stylus State
Stroke Count
Reset

Enter Writer ID 1

Enter Starting Character ID 1

Fig. 4. Data acquisition GUI.

2.3 Visual Data Inspection

The visual inspection of data aims at detecting human as well as system errors. Examples of these types of errors are skipping some character, wrongly recording some character, overwriting a character, etc. A separate program was developed for viewing the samples collected immediately after the completion of a data acquisition session (refer to Fig. 3). With the help of this program, it is possible to verify data visually in the presence of the writer immediately after his/her session of data acquisition and to overwrite the erroneous character files by rewriting those characters following the experimental protocol.

3. Data Attributes

Each character in the dataset contains attribute information. These attributes are the Name, Number of Strokes, and Sequence of Strokes associated with the character.

3.1 Character Name

The first line of each sample is 'CHARACTER_NAME: Character' (as shown for a sample character in Fig. 9 in Appendix A). The 'Character' is the Name of any one of the 183 characters.

3.2 Number of Strokes in a Sample

A stroke is a sequence of points from the pen-down to the pen-up events. The total number of strokes used to write a character is represented by the line 'STROKE_COUNT: Number' refer to Fig. 9 in Appendix A), where 'Number' is the total count of the strokes in the character.

3.3 Sequence of Strokes

Each stroke begins with the 'PEN_DOWN' information and the 'PEN_UP' information is followed by the 'PEN_DOWN' information between the two consecutive strokes. The end of a sample is represented by the 'PEN_UP' information, which is followed by the 'END_CHARACTER: Character' information. Each stroke consists of a sequence of X and Y coordinates values, which are given in the first and the second columns of the text file, respectively. Corresponding to each pair of values of X and Y coordinates, the 'STYLUS_STATE' and 'STROKE' information is given in the third and the fourth columns, of the text file, respectively (refer to Fig. 9 in Appendix A). The 'STYLUS_STATE' is either 1 or 0. Corresponding to each recorded (X, Y) point, the 'STYLUS_STATE' is 1 and corresponding to the 'PEN_UP' information the 'STYLUS_STATE' is 0. 'STYLUS_STATE' is kept blank corresponding to each piece of "PEN_DOWN" information. The 'STROKE' information represents the serial number of the constituent stroke of a sample.

4. Character Recognition

Character recognition experiments were performed on the 10 numeral classes (a total of 45×10=450 numerals) and 52 classes of basic alphabetic characters (a total of 45×52=2340 basic alphabetic characters) available in the dataset. The architecture for the classification of characters is shown in Fig. 5. In order to conduct the classification stage, we explored the support vector machine (SVM) [16] with three different kernels—linear, polynomial, and radial basis function (RBF).

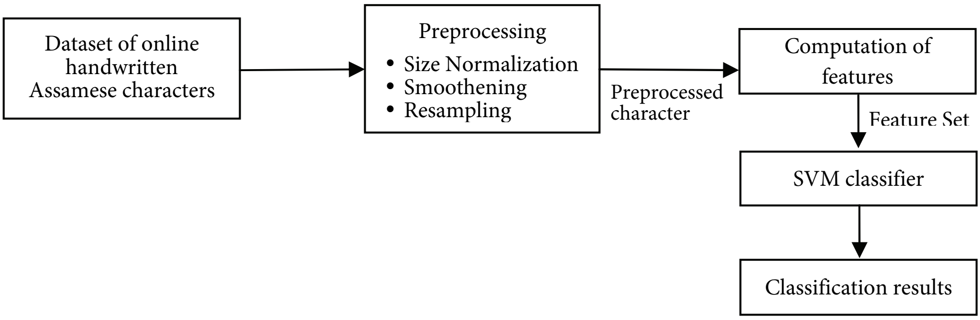
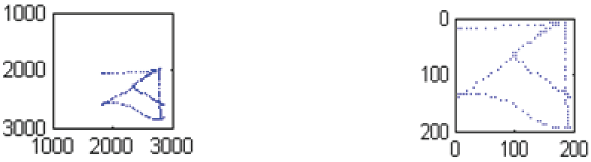


Fig. 5. Architecture for the classification of characters.

4.1 Data Preprocessing

4.1.1 Size normalization

Online handwritten characters are sequences of two-dimensional coordinate points (refer to Fig. 10 in Appendix A). Since the characters are written in different sizes at different locations of the text input box, all of the characters have different ranges of values for the X and Y coordinates. Therefore, the different ranges of values for the coordinate points need to be normalized to a uniform range of values. This step is performed by normalizing the values of the X and Y coordinates of the characters to the range (0, 200). Fig. 6(a) shows the original character and the character after size normalization is shown in Fig. 6(b).

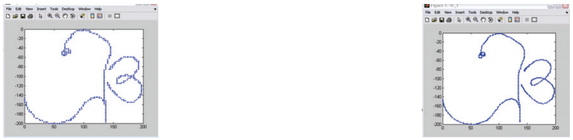


(a) (b)

Fig. 6. (a) Original character and (b) character after size normalization.

4.1.2 Smoothing

The smoothing of characters is required to reduce the sharp changes or roughness (jitters) from the stroke. It removes jitters from the data while preserving underlying patterns. Here, the smoothing of the characters was performed using a 3-point moving average filter in order to remove jitters from the characters. Fig. 7(b) represents the smoothing result of the character shown in the Fig. 7(a).

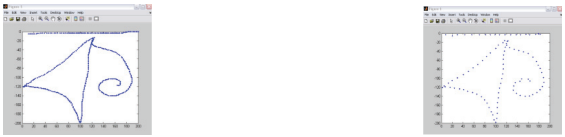


(a) (b)

Fig. 7. (a) Original character and (b) character after smoothing.

4.1.3 Resampling

An online handwritten character is a string of coordinate points. Different characters have a different number of points along the trajectory from the start to the end. Therefore, different characters need to be resampled to fixed number points. Here, all of the characters have been resampled to a fixed number of 100 points. Fig. 8(b) represents the resampling of the character shown in the Fig. 8(a) to 100 points.



(a) (b)

Fig. 8. (a) Original character and (b) character after resampling.

4.2 Feature Sets

We will now present the classification results, which are based on the following feature vectors that were computed from the characters taken from the dataset above. The feature set FS:2 is based on [17] and the feature set FS:3 is an enhancement of the feature set FS:2. The feature set FS:3 includes all the features of FS:2 along with five other features.

4.2.1 Feature set FS:1

- X and Y coordinate values of the resampled characters.
- The values of the X and Y coordinates of the characters are divided into sixty-four zones or sub-intervals of the normalized range of values (0, 200) of X and Y. The number of points in a character falling into each sub-interval of (0, 200) is considered as a feature of the character (zone feature).
- Coordinates of the start and end points of each character.
- Distance between the start and the end points of the character.
- Number of strokes in the character.
- Mean of the X and Y coordinate values.
- Standard deviation of X and Y coordinates values.
- Headline feature.

4.2.2 Feature set FS:2

- Normalized horizontal coordinates.
- Normalized vertical coordinates.
- Direction angle with reference to the horizontal axis (cosine of the angle).
- Direction angle with reference to the vertical axis (sine of the angle).
- Curvature with reference to the horizontal axis.
- Curvature with reference to the vertical axis.
- Position of the stylus (up or down).

4.2.3 Feature set FS:3

- Normalized horizontal coordinates.
- Normalized vertical coordinates.
- Direction angle with reference to the horizontal axis (cosine of the angle).
- Direction angle with reference to the vertical axis (sine of the angle).
- Curvature with reference to the horizontal axis.
- Curvature with reference to the vertical axis.
- Position of the stylus (up or down).
- The number of points contained in each of the sixty-four subintervals of the character (zone feature).
- Distance between the start and the end points of the character.
- Number of strokes in the character.
- Mean of the X and Y coordinate values.
- Standard deviation of the X and Y coordinate values.

4.3 SVM Based Classification

Experiments using polynomial kernel and Gaussian radial basis functions, which are the common choices for kernel functions in SVM, were performed for the classification of both numerals and basic alphabetic characters. A *k*-fold cross validation procedure was used for the training and testing of the SVM classifier with various parametric settings. In *k*-fold cross-validation, the original sample was randomly partitioned into *k* equal size subsamples. Of the *k* subsamples, a single subsample was retained as the validation data for testing the model, and the remaining *k* – 1 subsamples was used as training data. The cross-validation process was then repeated k times, in which each of the k subsamples were used exactly once as the validation data. The *k* results from the folds were then combined to produce a single estimation. All of the observations were used for both training and validation, and each observation was used for validation exactly one time. 10-fold cross-validation was done in this work.

4.3.1 Support vector machines

SVMs (Support Vector Machines) are a useful technique for data classification [16]. The goal of SVM is to produce a model (based on the training data), which predicts the target values of the test data given only the test data attributes.

Given a training set of instance-label pairs (x_i, y_i), $i = 1,\ldots\ldots\ldots\ldots, l$ where $x_i \in R^n$ and $y \in \{1,-1\}^l$, the SVMs require that a solution is found for the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0 \end{aligned}$$

Here, the training vectors x_i are mapped into a higher dimensional space by the function ϕ. SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. The three types of kernels that are used the most are as follows:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0$

Here, γ, r and d are the kernel parameters. In this experiment, we have used the SVM with the linear, polynomial, and RBF kernels.

4.4 Experimental Results

The overall recognition rates achieved for the online handwritten assamese basic alphabetic characters were 71.54% (using linear kernel), 75.39% (using polynomial kernel), and 78.38% (using RBF kernel) with a 10 fold cross validation process based on the feature set FS:1 (refer to Table 2). A total of 2,340 characters were used as samples in the basic alphabetic character recognition experiment. The kernel

Table 2. Statistics for the recognition rates

Type of SVM kernel	Total no. of instances	Correctly classified instances	Across classes		
			Average recognition rate (%)	SD of the recognition rate (%)	Average±SD
Online handwritten Assamese basic alphabetic characters					
Feature set FS:1					
Linear (C=1, E=1)	2340	1674	71.54	15.77	71.54±15.77
Polynomial (C=3, E=4)	2340	1764	75.39	14.39	75.39±14.39
RBF (C=9, Gamma=0.07)	2340	1834	78.38	13.31	78.38±13.31
Feature set FS:2					
Linear (C=1, E=1)	2340	1795	76.71	14.29	76.71±14.29
Polynomial (C=3, E=4)	2340	1862	79.57	12.23	79.57±12.23
RBF (C=8, Gamma=0.03125)	2340	1879	80.29	11.50	80.29±11.50
Feature set FS:3					
Linear (C=1, E=1)	2340	1834	78.38	13.45	78.38±13.45
Polynomial (C=3, E=4)	2340	1896	81.03	12.24	81.03±12.24
RBF (C=8, Gamma=0.03125)	2340	1899	81.15	11.08	81.15±11.08
Online handwritten Assamese numerals					
Feature set FS:1					
Linear (C=1, E=1)	450	441	98.00	2.21	98.00±2.21
Polynomial (C=1, E=4)	450	446	99.11	1.54	99.11±1.54
RBF(C=3, Gamma=0.05)	450	445	98.89	1.16	98.89±1.16
Feature set FS:2					
Linear (C=1, E=1)	450	435	96.67	5.37	96.67±5.37
Polynomial (C=1, E=4)	450	441	98.00	2.21	98.00±2.21
RBF (C=8, Gamma=0.03125)	450	440	97.78	1.79	97.78±1.79
Feature set FS:3					
Linear (C=1, E=1)	450	437	97.11	3.32	97.11±3.32
Polynomial (C=1, E=4)	450	440	97.78	2.56	97.78±2.56
RBF (C=8, Gamma=0.03125)	450	440	97.78	1.79	97.78±1.79

Table 3. Recognition rates of online handwritten Assamese basic alphabetic characters based on FS:1 for different combinations of the polynomial kernel parameters C and E

C (complexity parameter)	E (exponent)	Recognition rate (%)
1	1	71.54
1	2	73.89
1	3	74.66
1	4	75.39
1	5	74.79
2	1	71.30
2	2	73.96
2	3	74.76
2	4	75.39
2	5	74.79
3	1	71.26
3	2	73.91
3	3	74.95
3	4	75.39
3	5	73.79
3	6	73.46

parameter setting C = 3 and E = 4, which are associated with the polynomial kernel, was obtained by varying different values of C and E within a range and getting the recognition rate at each combination of parameters. Here C is the complexity (or penalty) parameter and E is the exponent of the polynomial kernel. The recognition rates of the basic alphabetic characters based on FS:1 for different combinations of the parameters C and E associated with the polynomial kernel are shown in Table 3. Similarly, the kernel parameter setting C = 9 and gamma = 0.07 associated with RBF kernel was obtained by varying different values of C and gamma within a range and getting the recognition rate at each combination of parameters. Similarly, based on FS:2 [17], the overall recognition rate achieved was 80.29% and based on FS:3, the overall recognition rate achieved was 81.15% (refer to Table 2). The individual recognition rates (in percentage) of the basic alphabetic characters class are represented in Table 4.

Table 4. Individual recognition rates of the online handwritten Assamese basic alphabetic characters

Sl. no.	Class	Recognition rate (%)			Sl. no.	Class	Recognition rate (%)		
		Feature set FS:1	Feature set FS:2	Feature set FS:3			Feature set FS:1	Feature set FS:2	Feature set FS:3
		RBF kernel with C=9 & Gamma= 0.07	RBF kernel with C=8 & Gamma= 0.03125	RBF kernel with C=8 & Gamma= 0.03125			RBF kernel with C=9 & Gamma= 0.07	RBF kernel with C=8 & Gamma= 0.03125	RBF kernel with C=8 & Gamma= 0.03125
1	A	68.9	82.2	82.2	27	TA	73.3	88.9	88.9
2	AA	80.0	77.8	80.0	28	THA	88.9	93.3	95.6
3	E	66.7	77.8	75.6	29	DA	86.7	88.9	88.9
4	EE	66.7	66.7	71.1	30	DHA	68.9	68.9	71.1
5	U	80.0	77.8	80.0	31	NA	55.6	66.7	64.4
6	UU	80.0	82.2	80.0	32	PA	77.8	82.2	82.2
7	REE	82.2	84.4	86.7	33	PHA	68.9	64.4	66.7
8	AE	100	100	100	34	BA	77.8	68.9	71.1
9	OI	91.1	93.3	93.3	35	BHA	71.1	88.9	88.9
10	O	91.1	93.3	93.3	36	MA	48.9	64.4	64.4
11	OU	84.4	91.1	91.1	37	AJA	60.0	62.2	64.4
12	KA	77.8	80.0	80.0	38	RA	62.2	60.0	60.0
13	KHA	80.0	82.2	84.4	39	LA	64.4	68.9	68.9
14	GA	84.4	77.8	77.8	40	WA	71.1	75.6	75.6
15	GHA	66.7	66.7	68.9	41	TXA	88.9	86.7	88.9
16	NG	93.3	95.6	95.6	42	MXA	57.8	64.4	64.4
17	CA	86.7	91.1	91.1	43	DXA	44.4	48.9	51.1
18	CCA	86.7	86.7	86.7	44	HA	84.4	88.9	88.9
19	JA	91.1	84.4	84.4	45	KHYA	77.8	91.1	91.1
20	JHA	91.1	88.9	91.1	46	AYA	60.0	64.4	66.7
21	NIYA	100	100	100	47	DRA	91.1	88.9	88.9
22	MTA	88.9	73.3	82.2	48	DHRA	71.1	73.3	77.8
23	MTHA	80.0	75.6	77.8	49	KTA	100	91.1	91.1
24	MDA	68.9	80.0	80.0	50	ANSR	97.8	91.1	91.1
25	MDHA	77.8	77.8	77.8	51	BXG	97.8	93.3	93.3
26	MNA	75.6	77.8	77.8	52	CBN	88.9	86.7	86.7

The overall recognition rate achieved for the online handwritten Assamese numerals was 99.11% (refer to Table 2), which was obtained by using the polynomial kernel with the kernel parameter settings C = 1 and E = 4 and a 10 fold cross validation process based on FS:1. 450 Assamese numeric characters were used as samples in the experiment of numeral recognition. The parameter setting C = 1 and E = 4 was obtained by varying different values of C and E within a range and getting the recognition rate at each combination of parameters. Similarly, based on the feature set FS:2 [17], the overall recognition rate achieved was 98.00% (refer to Table 2) and based on FS:3, the overall recognition achieved was 97.78% (refer to Table 2). The individual recognition rates (in percentage) of the numeral classes are represented in Table 5.

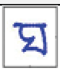
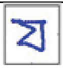
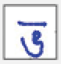
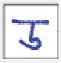
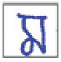
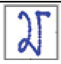
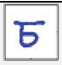
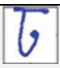
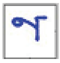
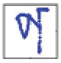
Table 5. Individual recognition rates of online handwritten Assamese numerals

Sl. No.	Class	Recognition rate (%)		
		Feature set FS:1	Feature set FS:2	Feature set FS:3
		Polynomial kernel with C=1 & E=4	Polynomial kernel with C=1 & E=4	RBF kernel with C=8 & Gamma=0.03125
1	SUNYA	97.8	100	100
2	EK	100	100	95.6
3	DUI	100	97.8	97.8
4	TINI	100	97.8	97.8
5	CARI	95.6	93.3	95.6
6	PAC	100	97.8	97.8
7	CAY	100	100	100
8	XAT	100	97.8	97.8
9	ATH	100	100	100
10	NAA	97.8	95.6	95.6

4.5 Shape Analysis of Similar Characters

The overall recognition rates achieved for the online handwritten Assamese basic alphabetic characters is not very encouraging. This may be because the major strokes of several characters are almost similar to each other. Accordingly, there are chances of these characters being misclassified, which would result in a low recognition rate. Table 6 illustrates some of these characters and each row in Table 6 represents similar characters from different classes.

Table 6. Similar characters from different classes

Similar characters	
 GHA	 AJA
 BHA	 MDA
 MA	 THA
 CA	 MDHA
 MNA	 PA

Improvement in the recognition rates (refer to Table 4) of some of these characters has been noticed (from feature set FS:1 through feature set FS:3) when the direction angle feature and curvature feature along with zone feature are used. Identifying the discriminating features for an improved classification of these types of similar characters is part of ongoing research.

5. Final Comments

In this work, we have reported on the development of a dataset of online handwritten Assamese characters. The samples were collected from a variety of writers belonging to various groups, in order to achieve a variety of writing patterns of characters. The samples of the dataset were not preprocessed. The recorded information was directly stored in raw format, which provided a scope for applying different preprocessing techniques to the dataset. The dataset of online handwritten Assamese characters can be downloaded for free from the UCI Machine Learning Repository [18]. This reported dataset is the only publicly available dataset of online handwritten Assamese characters. The dataset aims at providing samples for research in online handwriting recognition for Assamese scripts.

The preliminary results on the recognition of numerals and basic alphabetic characters taken from the dataset are reported in this work. From among the three SVM kernels used in this experiment on numerals, the polynomial kernel gives the best recognition rate of 99.11% based on the feature set FS:1. The recognition rate obtained in the case of online handwritten Assamese numerals is higher than the recognition rate of 96.6% obtained in an HMM based technique reported in [19]. From among the three SVM kernels used in this experiment for Assamese alphabetic characters, the RBF kernel gives the best recognition rate of 81.15% based on the feature set FS:3. No results are available in the literature on the topic of recognition of online handwritten Assamese alphabetic characters. Exploring a robust set of features for improving the recognition of characters and extending the same for recognition of conjunct consonants (*Juktakkhor*) is part of ongoing research.

Acknowledgement

The Online Handwritten Assamese Characters Dataset is the result of online handwriting samples contributed by students, research scholars, and faculty members of Tezpur University. Their support is gratefully acknowledged.

References

[1] R. Plamondon and S. Srihari, “Online and offline handwriting recognition: a comprehensive survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, 2000.

[2] E. Alpaydin and Fevzi. Alimoglu, “Pen-based recognition of handwritten digits dataset,” 1998; http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits.

[3] C. Vivard-Gaurdin, P. M. Lallican, S. Knerr, and P. Binter, “IRESTE On/Off (IRONOFF) dual handwriting database,” in *Proceeding of the 5th International Conference on Document Analysis and Recognition*, Bangalore, India, 1999, pp. 455-458.

[4] D. Llorens, F. Prat, A. Marzal, J. M. Vilar, M. J. Castro, J. C. Amengual, S. Barrachina, A. Castellanos, S. España, J. A. Gómez, J. Gorbe, A. Gordo, V. Palazón, G. Peris, R. Ramos-Garijo, and F. Zamora, "The UJIpenchars Database: a pen-based database of isolated handwritten characters," in *Proceeding of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008, pp. 2647-2651.

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

[6] A. Bharath and S. Madhvanath, "Hidden Markov model for online handwritten Tamil word recognition," in *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Curtiba, Brazil, 2007, pp. 506-510.

[7] L. Prasanth, J. Babu, R. Sharma, and P. Rao, "Elastic matching of online handwritten Tamil and Telegu scripts using local features," in *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Curtiba, Brazil, 2007, pp. 1028-1032.

[8] U. Bhattacharya, B. K. Gupta, and S. K. Parui, "Direction code based features for recognition of online handwritten characters of Bangla," in *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Curtiba, Brazil, 2007, pp. 58-62.

[9] N. Joshi, G. Sita, A. G. Ramakrishnan, V. Deepu, and S. Madhvanath, "Machine Recognition of Online Handwritten Devanagari Characters," in *Proceedings of the 8th International Conference on Document Analysis and Recognition*, Seoul, Korea, 2005, pp. 1156-1160.

[10] A. Sharma, R. Kumar, and R. K. Sharma, "Online handwritten Gurmukhi character recognition using elastic matching," in *Proceedings of the Congress on Image and Signal Processing*, Hainan, China, 2008, pp. 391-396.

[11] A. S. Bhaskarabhatla and S. Madhvanath, "Experiences in collection of handwriting data for online handwriting recognition in Indic scripts," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.

[12] B. B. Chaudhuri, "A complete handwritten numeral database of Bangla: a major Indic script," in *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*, France, 2006.

[13] K. C. Santosh, C. Nattee, and Bart Lamiroy, "Relative positioning of stroke based clustering: a new approach to on-line handwritten Devanagari character recognition," *International Journal of Image & Graphics (IJIG)*, vol. 12, no. 2, 2012.

[14] N. Saharia and K. M. Konwar, "LuitPad: a fully unicode compatible Assamese writing software," in, Proceedings of the 2nd Workshop on Advances in Text Input Methods (WTIM 2), Mumbai, India, 2012, pp. 79-88.

[15] N. S. Bhabendra, *Amar Akhar (Second Part)*. Guwahati: Assam Book Hive, 2008.

[16] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2003.

[17] A. R. Ahmed, C. V. Gaudin, M. Khalid, and R. Yusof, "Online handwriting recognition using support vector machine," in *Proceedings of the 2nd International Conference on Artificial Intelligence in Engineering and Technology*, 2004, Sabah, Malaysia, pp. 250-256.

[18] U. Baruah and S. M. Hazarika, "Online handwritten Assamese characters dataset," 2011; http://mlr.cs.umass.edu/ml/datasets/Online+Handwritten+Assamese+Characters+Dataset.

[19] G. S. Reddy, P Sharma, S. R. M. Prasanna, C. Mahanta and L. N. Sharma, "Combined online and offline Assamese handwritten numeral recognizer," in *Proceedings of the National Conference on Communications (NCC2012)*, Kharagpur, 2012, pp. 1-5.

Appendix A

An illustration on representation of handwritten characters in the dataset is given below. The text file corresponding to this data format is ‘77.37.txt’, which represents the sample with name TTT, Character ID 77 written by the writer with Writer ID 37. The .TXT file representation of the character TTT corresponding to Writer ID 37 is reproduced in Fig. 9. A plot of the (X, Y) points of the corresponding character is shown in the Fig. 10. From the character listing ‘Data_Table.pdf’, it can be verified that the name of the character is ‘TTT’.

```
CHARACTER_NAME: TTT
STROKE_COUNT: 2
 X         Y              STYLUS_STATE        STROKE
PEN_DOWN
1217      1217                  1                  1
1191      1217                  1                  1
1244      1217                  1                  1
1323      1217                  1                  1
1429      1217                  1                  1
1561      1217                  1                  1
1720      1217                  1                  1
1905      1217                  1                  1
2090      1217                  1                  1
2302      1217                  1                  1
2461      1217                  1                  1
2619      1217                  1                  1
2752      1217                  1                  1
2831      1217                  1                  1
2910      1217                  1                  1
2963      1217                  1                  1
2990      1217                  1                  1
2963      1244                  1                  1
2937      1244                  1                  1
PEN_UP                          0
PEN_DOWN
2143      1693                  1                  2
2170      1720                  1                  2
2170      1693                  1                  2
2196      1693                  1                  2
2196      1667                  1                  2
2223      1640                  1                  2
2249      1614                  1                  2
2302      1614                  1                  2
2355      1588                  1                  2
2434      1614                  1                  2
2514      1640                  1                  2
2593      1667                  1                  2
2646      1720                  1                  2
2725      1773                  1                  2
2778      1826                  1                  2
2805      1879                  1                  2
2805      1931                  1                  2
2805      1984                  1                  2
2778      2037                  1                  2
2725      2090                  1                  2
2646      2117                  1                  2
2593      2170                  1                  2
```

```
2514    2196            1               2
2461    2223            1               2
2408    2249            1               2
2381    2275            1               2
2408    2249            1               2
2434    2223            1               2
2461    2223            1               2
2514    2223            1               2
2566    2196            1               2
2646    2223            1               2
2725    2223            1               2
2805    2275            1               2
2858    2328            1               2
2910    2381            1               2
2937    2434            1               2
2963    2514            1               2
2990    2566            1               2
2963    2619            1               2
2937    2672            1               2
2884    2725            1               2
2805    2752            1               2
2725    2752            1               2
2619    2752            1               2
2514    2725            1               2
2381    2699            1               2
2275    2619            1               2
2170    2540            1               2
2064    2434            1               2
1958    2328            1               2
1879    2196            1               2
1826    2064            1               2
1773    1931            1               2
1746    1799            1               2
1720    1667            1               2
1693    1588            1               2
1720    1535            1               2
1720    1482            1               2
1746    1455            1               2
1746    1429            1               2
PEN_UP                  0
END_CHARACTER: TTT
```

Fig. 9. The .TXT file representation of the character TTT corresponding to Writer ID 37.

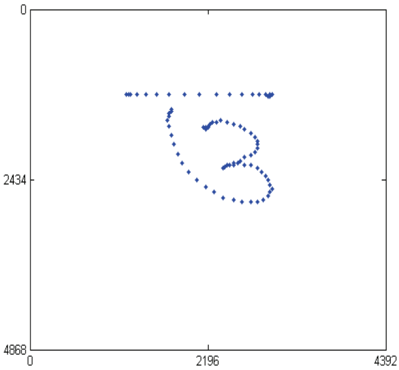


Fig. 10. Plot of the character TTT corresponding to Writer ID 37.

Udayan Baruah

He is an Associate Professor of Information Technology at Sikkim Manipal Institute of Technology, Sikkim, India. He received his M.Sc. in Mathematics from the Faculty of Mathematical Sciences, North Campus, University of Delhi, India and M.Tech. in Information Technology from the Department of Computer Science and Engineering, Tezpur University, Assam, India. Presently, he is working towards his Ph.D. degree in the Department of Computer Science and Engineering, Tezpur University. His field of research is Online Handwriting Recognition.

Shyamanta M. Hazarika

He is a Professor of Computer Science and Engineering at Tezpur University, Assam, India. He completed his B.E. from Assam Engineering College, Guwahati, India, M.Tech. in Robotics from IIT Kanpur, India and Ph.D. from University of Leeds, England. He has been a Full Professor for Cognitive Systems and Neuro Informatics at the Cognitive Systems Group, University of Bremen, Germany, for winter 2009-2010. His work has focused primarily in knowledge representation and reasoning and rehabilitation robotics with an aim towards development of intelligent assistive systems.