Chapter 3

A Dataset of Online Handwritten Assamese Characters

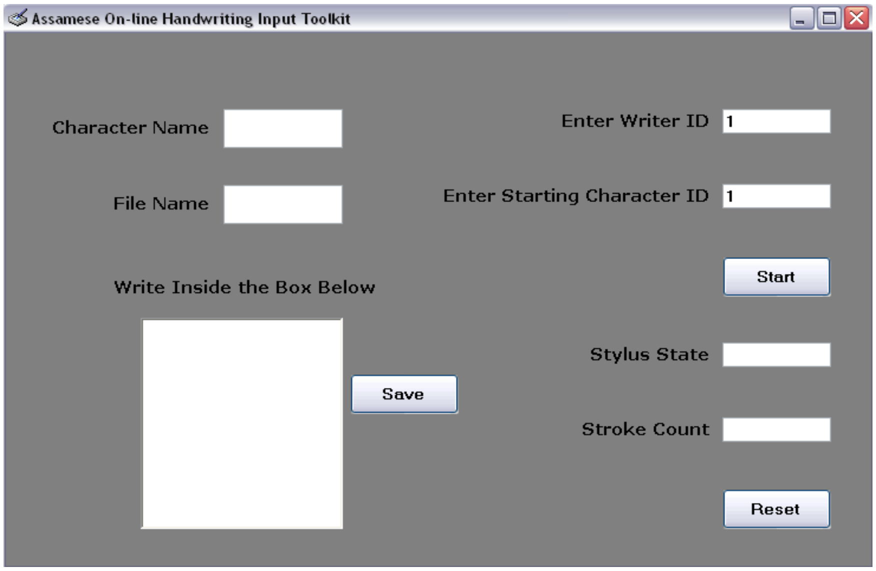
This chapter describes the development of a dataset of online handwritten Assamese characters. Assamese is a major script in the northeastern part of India. No online handwritten Assamese characters dataset is available. Therefore, the major objective is to build a dataset of online handwritten Assamese characters. The online handwritten Assamese characters dataset contains a total of 8,235 online handwritten Assamese characters of 183 classes, consisting of Assamese numerals, basic alphabetic characters, and conjunct consonants (Juktakhors). The dataset was developed in Tezpur University as a part of this research project. We name this dataset as Tezpur University-Online Handwritten Assamese Characters (TU-OHAC) Dataset.

3.1 Materials and Methods

This section describes briefly about subjects and experimental setup used to collect online handwriting samples of Assamese characters.

3.1.1 Subjects and Experimental Setup

For the creation of the 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 [98], 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) belonging to different age groups ranging from 20–42 years old with the average age being that of 31. Printed Assamese Numerals, Basic Alphabetic Characters (Vowels and Consonants), and a few instances of Conjunct Consonants (*Juktakhors*) are shown in Table 1.1 of Chapter 1. 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. The resolution of the pen-tablet was 2540 LPI. A screenshot of the GUI is shown in Figure 3.1.


Enter Writer ID: 1

Enter Starting Character ID: 1

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

Figure 3.1: Data Acquisition GUI

3.1.2 *Experimental Protocol*

The Experimental protocol for data acquisition is presented in Figure 3.2. 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. 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* of the experimental protocol. 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 k^{th} *character.*

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

Figure 3.2: Experimental Protocol for Data Acquisition

3.2 Data Acquisition and Verification

Data acquisition is the collection of online handwritten samples from different writers. Data collected once are verified for the detection of errors which may occur during data acquisition. The samples of the dataset were collected from students, research scholars, and faculty members of Tezpur University.

3.2.1 Data Acquisition

The online Assamese handwriting samples were collected in a laboratory environment. Screenshots of parts of data acquisition sessions are shown in the Figure 3.3. 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 saved in a text file when the onscreen *Save* button is clicked.

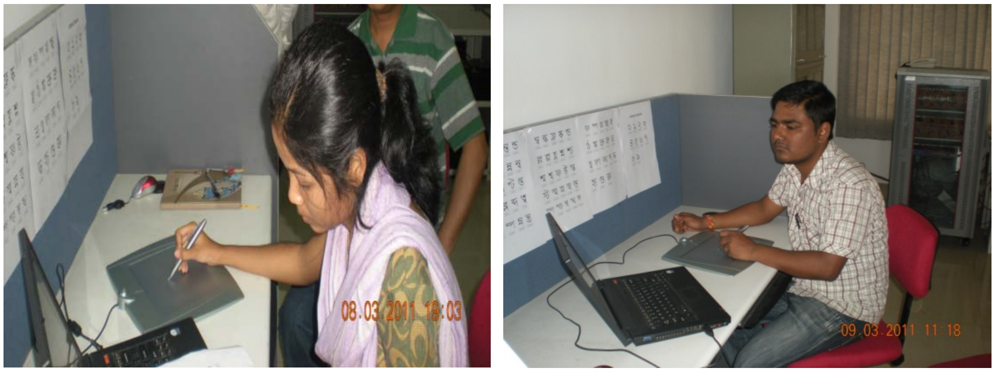


Figure 3.3: Data acquisition sessions

3.2.2 *Data Verification*

The visual verification 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. A screenshot of a part of the visual data verification program is shown in the Figure 3.4. 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.

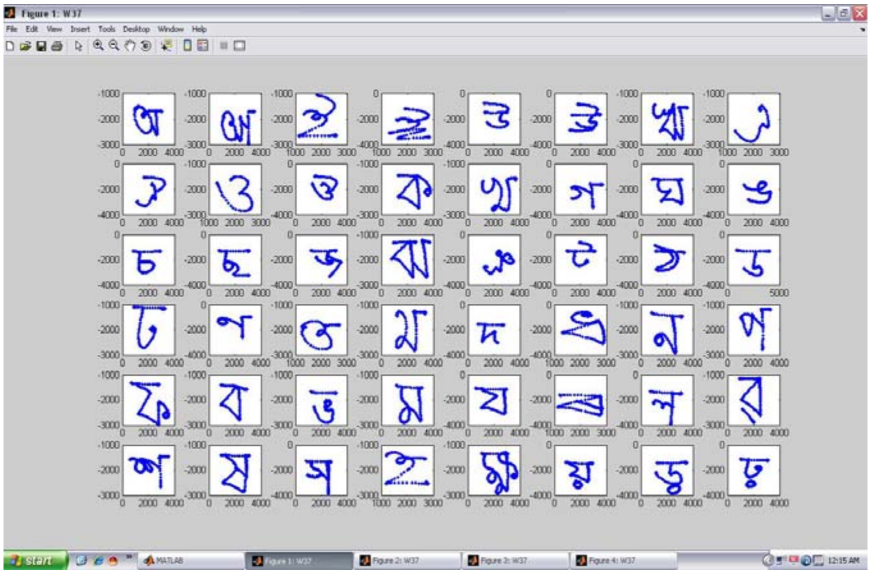


Figure 3.4: A screenshot of a part of the Visual Data Inspection Program

3.3 *TU-OHAC Dataset*

This section provides a description of TU-OHAC Dataset by highlighting some of its recent usage. The dataset of online handwritten Assamese characters can be downloaded for free from the UCI Machine Learning Repository [20, 25]. This reported dataset is the only publicly available dataset of online handwritten Assamese characters.

3.3.1 Dataset Description

The distribution of TU-OHAC dataset consists of 45 folders (one for each writer) and a data description file named 'Data_Table.pdf' (Appendix A). The images of all the Assamese characters given as reference shapes can also be downloaded along with the dataset from the UCI Machine Learning Repository (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. Along the printed shape of the character (Char), this file also contains information about the character ID (ID) and character name (Label). Each data folder contains 183 text files corresponding to the 183 characters written by a single writer. Each text file corresponding to each recorded character 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'. An illustration on representation of handwritten characters in the dataset is given in Appendix B. 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 Figure B.1 in Appendix B. A plot of the (X, Y) points of the corresponding character is shown in the Figure 3.5. From the character listing 'Data_Table.pdf', it can be verified that the name of the character is 'TTT'. Some collected samples written by different writers are shown in Figure 3.6. Moreover, 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. The variability of writing patterns of the same character is shown in Figure 3.7. Again, major strokes of several characters are almost similar to each other. Table

3.1 illustrates some of these similar characters where each row in Table 3.1 represents similar characters from different classes.

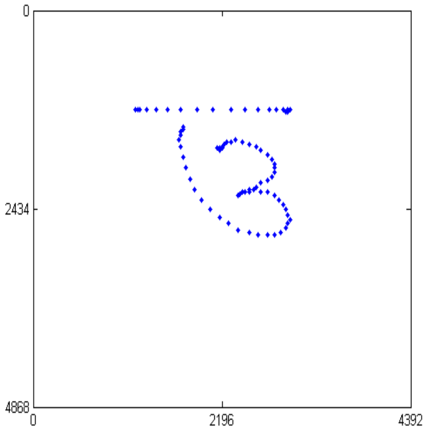


Figure 3.5: Plot of the character TTT corresponding to Writer ID 37

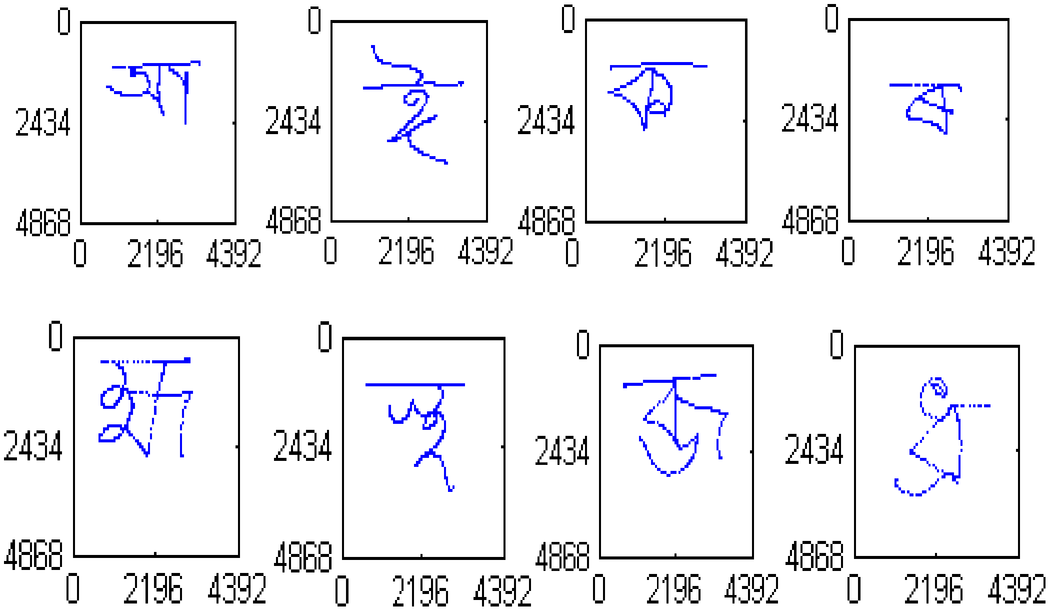


Figure 3.6: Samples from the dataset of online handwritten Assamese characters written by different writers. In this figure we present samples from eight different writers.

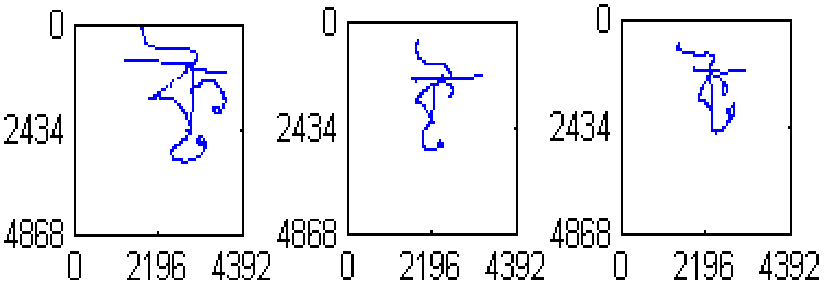
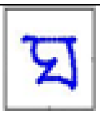
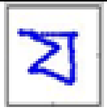
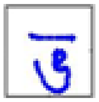
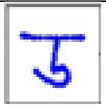
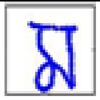
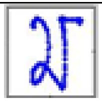
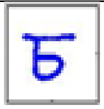
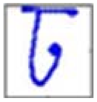
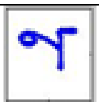
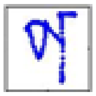


Figure 3.7: Variability of patterns of the same character written by three different writers

Table 3.1: Similar Characters from Different Classes

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

3.3.1.1 Data Attributes

Each character in the dataset contains attribute information. These attributes must be selected properly such that these depict the relevant information about the recorded online handwritten characters. The attributes selected for online handwritten Assamese characters are the *Name*, *Number of Strokes* and *Sequence of Strokes* associated with the character.

- *Character Name*

The first line of each sample is 'CHARACTER_NAME: Character' (as shown for a sample character in Figure B.1 in Appendix B). The 'Character' is the *Name* of any one of the 183 characters.

- *Number of Strokes*

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' (Figure B.1 in Appendix B), where '*Number*' is the total count of the strokes in the character.

- *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 (Figure B.1 in Appendix B). 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.

3.3.2 Usage of the Dataset

The following are the recent usage of TU-OHAC (online handwritten Assamese characters dataset) dataset in online handwriting recognition activities found in literature.

- Graham et al. [99] used TU-OHAC dataset to test the accuracy of Convolution Neural Network based online handwritten character recognition system. In this work, out of 45 samples of each of the 183 online handwritten Assamese characters first *k* handwriting samples are used as training set and the remaining 45-*k* samples are used for a test set considering *k*=15 or 36.
- Reizenstein et al. [100] investigated the use of iterated integral signatures as a representation of handwritten characters for improved machine recognition, using data from TU-OHAC dataset. They have compared methods for including information about pen-liftings in multi-stroke characters. It was showed that the ink dimension is most useful and much useful information is lost when moving to a rotationally-invariant representation. Out of 45 samples of online handwritten Assamese characters, all from different writers, of each of 183 characters the dataset was split so that the first 36 of each character are used for training and the last 9 of each character are used for testing.
- Aydin et al. [102] used TU-OHAC dataset to study a system of online handwriting recognition system based on Feed Forward Backpropagation Artificial Neural Network and Radial Basis Artificial Neural Network. In

this study (X,Y) coordinate values of online hand written Assamese characters are saved by the program. Features were found by getting maximum, minimum, average, variant, Standard deviation and range values after size of these values are decreased by Principle Component Analysis. Test results showed that Feed Forward Backpropagation Artificial Neural Network gave 96% of accuracy and Radial Basis Artificial Neural Network was 82% successful.

- Egho [103] empirically evaluated similarity measure in a clustering problem on the samples of TU-OHAC dataset. The results reported are a useful contribution with direct practical applications to different discriminative approaches. An extensive empirical study on qualitative experiments was reported with datasets consisting of trajectories of online handwritten Assamese characters.

3.4 Challenges Encountered

The challenges encountered in the creation of a dataset of online handwritten Assamese characters can be the choice of size of the alphabet, availability of a user friendly GUI for handwriting inputs and human factors.

3.4.1 Choice of Size of the Alphabet

The size of the alphabet or the character set is a major challenge in the creation of a dataset of handwritten characters in any language. In Assamese there are 10 numeric characters and 52 basic alphabetic characters that consist of 11 vowels and 41 consonants. An Assamese character set is unique in that is has a large number of conjunct consonants (*Juktakkhors*) of about 164-201 [33]. But not all characters are frequently used by the users. Specially, some of the Juktakkhors are so frequently used in writing. Therefore, choosing an optimal number of characters which have better usability is a challenge in the creation of a dataset of handwritten characters.

3.4.2 User Friendly Data Acquisition GUI

Designing a user friendly GUI for online handwritten data acquisition is a challenging task. The data acquisition GUI must be easy for the user to interact with the acquisition system. The utilization of the GUI depends on the functionalities which should be available in the GUI. The GUI should include appropriately placed text input box and other button facilities which a user needs to execute his/her session of handwriting input.

3.4.3 Human factors

The tools for online handwriting inputs include digitizing tablet with touch sensitive writing surface and a stylus. There are a number of human factors to be considered when substituting a stylus and tablet for pen and paper for online handwritten data acquisition. The paper is still a preferred medium. Therefore, the major human factor involved in the acquisition of online handwriting samples is the inconvenience in writing with the text input surface. Users are not usually accustomed with writing on the touch sensitive surface of the tablet with the stylus. Therefore, a writer has to practice several times to become familiar with online handwriting data acquisition environment before the actual recording of handwritings takes place.

3.5 Conclusion

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. Being 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 importance of the dataset of online handwritten Assamese characters (TU-OHAC Dataset) is evident from its usage by several research groups in performing online handwritten character recognition experiments [99, 100, 101, 102, 103].