Online Bangla Handwriting Recognition System

K. Roy
Dept. of Comp. Sc. West Bengal University of Technology, BF 142, Saltlake, Kolkata-64, India

N. Sharma, T. Pal and U. Pal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India

Handwriting recognition is a difficult task because of the variability involved in the writing styles of different individuals. This paper presents a scheme for the online handwriting recognition of Bangla script. Online handwriting recognition refers to the problem of interpretation of handwriting input captured as a stream of pen positions using a digitizer or other pen position sensor. The sequential and dynamical information obtained from the pen movements on the writing pads are used as features in our proposed scheme. These features are then fed to the quadratic classifier for recognition. We tested our system on 2500 Bangla numeral data and 12500 Bangla character data and obtained 98.42% accuracy on numeral data and 91.13% accuracy on character data from the proposed system.

Keywords: Online Recognition, Indian Script, Bangla, Modified quadratic discriminant function

1. Introduction

Data entry using pen-based devices is gaining popularity in recent times. This is so because machines are getting smaller in size and keyboards are becoming more difficult to use in these smaller devices. Also, data entry for scripts having large alphabet size is difficult using keyboard. Moreover, there is an attempt to mimic the pen and paper metaphor by automatic processing of online characters. However, wide variation of human writing style makes online handwriting recognition a challenging pattern recognition problem.

Work on online character recognition started gaining momentum about forty years ago. Numerous approaches have been proposed in the literature [1–5] and the existing approaches can be grouped into three classes namely, (i) structural analysis methods where each character is classified by its stroke structures, (ii) statistical approaches where various features extracted from character strokes are matched against a set of templates using statistical tools and (iii) motor function models that explicitly use trajectory information where the time evaluation of the pen co-ordinates plays an important role.

Many techniques are available for online recognition of English, Arabic, Japanese and Chinese [1–4, 6–9] characters but there are only a few pieces of work [10–13] available towards Indian characters although India is a multi-lingual and multi-script country. Connell et al. [11] presented a preliminary study on online Devnagari character recognition. Joshi et al. [12] also proposed a work on Devnagari online character recognition. Later Joshi et al. [13] proposed an elastic matching based scheme for online recognition of Tamil character recognition. Although there are some work towards online recognition of Devnagari and Tamil scripts but the online recognition work towards other Indian languages is very few. In this paper we propose a system for the online recognition of Bangla character. Recognition of Indian character is very difficult with compare to English because of its shape variability of the characters as well as larger number of character classes. See Figure 1 where samples of four Bangla characters are shown to get an idea of handwriting variability.

There are twelve scripts in India and in most of these scripts the number of alphabets (basic and compound characters) is more than 250, which makes keyboard design and subsequent data entry a difficult job. Hence, online recognition of such scripts has a commercial demand. Although a number of studies [14–16] have been done for offline recognition of a few printed Indian scripts like Devnagari, Bangla, Gurumukhi, Oriya, etc. with commercial level accuracy, but to the best of our knowledge no system is commercially available for online recognition of any Indian script. In this paper we propose a scheme for online Bangla handwritten character recognition and the scheme is robust against stroke connections as well as shape variation while maintaining reasonable robustness.
against stroke order-variations. A quadratic classifier is used here for recognition purpose.

The rest of the paper is organized as follows. In Section 2 we discuss about the Bangla language and data collection. Feature extraction process is presented in Section 3. Section 4 details the classifier used for recognition. The experimental results are discussed in Section 5. Finally, conclusion of the paper is given in Section 6.

Fig. 1. Examples of some Bangla online characters. First three columns show samples of handwritten characters and last column shows samples of a numeral.

2. Bangla Language and online data collection

Bangla, the second most popular language in India and the fifth most popular language in the world, is an ancient Indo-Aryans language. About 200 million people in the eastern part of Indian subcontinent speak in this language. Bangla script alphabets are used in texts of Bangla, Assamese and Manipuri languages. Also, Bangla is the national language of Bangladesh.

The alphabet of the modern Bangla script consists of 11 vowels and 40 consonants. These characters are called as basic characters. Writing style in Bangla is from left to right and the concept of upper/lower case is absent in this script. It can be seen that most of the characters of Bangla have a horizontal line (Matra) at the upper part. From a statistical analysis we notice that the probability that a Bangla word will have horizontal line is 0.994 [14].

In Bangla script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right, both left and right, or bottom of the consonant. These modified shapes are called modified characters. A consonant or a vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters also exists in Bangla. There are about 280 compound characters in Bangla [15]. In this paper we consider the recognition of Bangla basic characters.

To get an idea of Bangla basic characters and their variability in handwriting, a set of handwritten Bangla basic characters are shown in Figure 2. Main difficulty of Bangla character recognition is shape similarity, stroke size and the order variation of different strokes. By stroke we mean the set of points obtained between a pen down and pen up. From the statistical analysis on our dataset we found that the minimum (maximum) number of stroke used to write a Bangla character is 1 (4). The average number of stroke per character is 2.2. We have also seen that the characters ‘ঃ’, ‘ঃ’ and ‘ঃ’ are mostly written by single stroke whereas the character ‘ঃ’ is written by almost all writer by 4 strokes. It also found that the characters ‘ঃ’, ‘ঃ’, ‘ঃ’, ‘ঃ’, ‘ঃ’, ‘ঃ’, ‘ঃ’, ‘ঃ’ etc. are always written by 2 strokes.

Online recognition of these characters imposes several problems like stroke-number, stroke connection and shape variations. Most characters are composed of multiple strokes. Another difficult problem involves in the stroke-order variations. Writing of different stroke sequences of a character is not similar for all persons. During handwriting, some people give upper stroke of a character before giving a lower stroke of that character. Whereas to write the same character, some people give lower stroke before giving the upper stroke of that character. These stroke order variations complicates the development procedure of an online recognition system.
To illustrate this stroke order variation in Bangla script, Figure 3 shows a Bangla character that contains four different strokes. The left-most column shows the first stroke and this stroke is same for all the three samples of three different writers. Stroke-order varies from the second column onwards and the final (complete) character is shown in the right-most column. From the 2nd column of Figure 3, it can be noted that the three samples have different shape. This is because of stroke order variation of the writers. For upper sample of the second column of Figure 3, the writer has given the upper stroke as second stroke. For middle sample of the second column the writer has given the lower stroke as second stroke. For lower sample, the writer has given the middle stroke as second stroke. Similar situation also occurs in 3rd column for Figure 3. Matra is a very common feature in Bangla and its length varies from writer to writer. It is seen that the presence or absence of Matra is the main difference between two characters. For example the character ঠ is a basic character which has Matra but ঠ is a Bangla numeral and it does not have Matra.

For online data collection, the sampling rate of the signal is considered fixed for all the samples of all the classes of character. Thus the number of points M in the series of co-ordinates for a particular sample is not fixed and depends on the time taken to write the sample on the pad. As the number of points in actual trace of the characters are generally large and varies greatly due to high variation in writing speed, a fixed lesser number of points, regularly spaced in time are selected for further processing.

The digitizer output is represented in the format of $p_i \in \mathbb{R}^2 \times \{0,1\}$, where $p_i$ is the pen position having x-coordinate ($x_i$) and y-coordinate ($y_i$) and M is the total number of sample points. Let $(p_i)$ and $(p_j)$ be two consecutive pen points. We retain both of these two consecutive pen points $(p_i)$ and $(p_j)$ if the following condition is satisfied:

$$x^2 + y^2 > m^2$$  \hspace{1cm} (1)

where $x = x_i - x_j$ and $y = y_i - y_j$. The parameter $m$ is empirically chosen. We have set $m$ equal to zero; in Equation 1 to removes all consecutive repeated points.

Analyzing a total of 15000 Bangla characters we found that, for writing Bangla characters, the number of sample points (M) varies from 14 (for the character য) to 176 (for the character শ) points. The average number of sample points in a Bangla character is 72. We also computed the average number of sample points in each character class. We noted that the character class শ has the maximum number of sample points and its average value is 113. The character class প has the minimum number (46) of sample points.

3. Feature extraction

Any online feature is very much sensitive to writing stroke sequence and size variation. Also, in Bangla Matra creates a lot of problem in online recognition. To overcome it, we detect and remove the Matra present in the characters. As Matra of Bangla script
is a digital straight line lies on the upper part of a character. A stroke is called Matra if:

(a) The ratio of distance of 1st and last point of the stroke to sum of individual distances between constituent points of the stroke is less than 1.5.
(b) The ratio of the stroke height to the character height \([\max(y_i) - \min(y_i)]\) is less than 0.35 and all the points of the stroke lie in upper side (we consider top 40% height of the character as upper side) of the character.

These thresholds are decided based on statistical analysis of our dataset. If more than one stroke satisfies the above criteria then higher value of condition 1 is selected.

The features calculated based on Matra are

1. The ratio of average value of x coordinate of the selected stroke to the length of the character,
2. The ratio of average value of y coordinate of the selected stroke to the width of the character,
3. Ratio of the length \((L = \sum l_i, i = 1..M)\) where \(l_i = (x^2 + y^2)\), \(x = x_i - x_{i+1}\) and \(y = y_i - y_{i+1}\) of the stroke to the length of the character,
4. Ratio of the area of the stroke to the character and
5. Ratio of aspect ratio of the stroke to that of character.

A total of 5 features as discussed above are calculated based on Matra. After feature detection from Matra, it is then removed from the character and the rest of the points of the characters are first normalized. The normalization is done in two stages. First the points are re-sampled to a fixed number \(N\) and then they are converted from equal time sample to equal distant points. For example see Figure 4. We have studied several local features, which include a normalized representation of the co-ordinates, a representation of the tangent slope angle, a normalized curvature, the ratio of tangents, etc.

The processed character is transformed into a sequence \(t = [t_1, ..., t_N]\) of feature vectors \(t_i = (t_{i1}, t_{i2}, t_{i3})^T\). Here

1. \(t_{i1} = (x_i - \mu_x)/\sigma_x\) and \(t_{i2} = (y_i - \mu_y)/\sigma_y\) are the pen co-ordinates normalized by the sample mean \(\mu = \frac{1}{N} \sum_{i=1}^{N} p_i\) and standard deviation, \(\sigma_y = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mu_y - y_i)^2}\) of the character’s sample points, and
2. \(t_{i3} = \text{arg}(\arg((x_{i+1} - x_{i-1}) + j \ast (y_{i+1} - y_{i-1})))\), with \(j^2 = -1\) and “\(\text{arg}\)” the phase of the complex number above, is an approximation of the tangent slope angle at point \(i\).

Thus finally, a feature vector sequence is defined as \(t = [t_1, ..., t_N]\), each vector of it as \(t_i = (t_{i1}, t_{i2}, t_{i3})^T\) is obtained. Here the number of points \((N)\) in which the character is normalized is 50. So a total of 155 (50 X 3 [3 for each point] + 5 [5 features based on Matra]) features are used in our experiment.

![Figure 4](image-url)  
Fig. 4. Feature extraction from a sample of character is shown. (a) Original image, (b) its normalized point used as feature (mapped into 50 points), (c) the normalized character.

4. The Classifier

Based on these 155 dimensional features, recognition of characters in quadratic classifier [17] is carried out by using the following discriminant function:

\[
g(X) = (N + N_0 - n - 1) \ln[1 + \frac{1}{N_0 \sigma^2} ||X - M||^2] - \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i + \frac{N_0}{N} \sigma^2} \{\Phi_i^T (X - M)\}^2] + \sum_{i=1}^{k} \ln(\lambda_i + \frac{N_0}{N} \sigma^2).
\]

where \(X\) is the feature vector of an input character; \(M\) is a mean vector of samples; \(\phi_i^T\) is the \(i^{th}\) eigen vector of the sample covariance matrix; \(\lambda_i\) is the \(i^{th}\) eigen value of the sample covariance matrix; \(n\) is the feature size; \(\sigma^2\) is the initial estimation of a variance; \(N\) is the number of learning samples; and \(N_0\) is a confidence constant for \(\sigma\) and \(N_0\) is considered as 3N/7 for 155 dimensional feature. We do not use all eigen values and their respective eigen vectors for the classification. We sort the eigen values in descending order and take first 70 eigen values and their respective eigen vectors for classification.

Rejection in the system is done if for a character the difference of 1st and 2nd value of \(g(X)\) is smaller than a threshold.
5. Results and Discussion

The experimental evaluation of the above techniques was carried out using isolated Bangla character and numerals (digits). The data was collected from people of different background both by using mouse as well as Wacom tablet. The mouse is used as an input medium to facilitate the user with the flexibility of using this common input medium. The data obtained by mouse is in general poorer than Wacom tablet. Our another aim to use this mouse as one of the input devices is to test our system on these poorer data. If we get encouraging results from the data capture by mouse then there is no need to use any additional hardware like Wacom tablet for data capture. A total of 15,000 characters (2500 digits and rest are characters) are collected for the experiment. Out of them 66.7% of the characters (digits) are used for the training of the classifier for the present work and rest is used for the testing purpose. The recognition accuracy obtained from our classifier is shown in Table 1.

Table 1. Recognition results on Bangla character and numeral (here rejection is not considered).

<table>
<thead>
<tr>
<th>Data</th>
<th>Recognition rate</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
<td>91.13%</td>
<td>8.87%</td>
</tr>
<tr>
<td>Numeral</td>
<td>98.42%</td>
<td>1.58%</td>
</tr>
</tbody>
</table>

From the experiments we noted that the overall recognition accuracy of the proposed scheme was 91.13% for Bangla characters and 98.42% accuracy for Bangla numerals when rejection is not considered. Accuracy of 96.23% and 99.58% is obtained for Bangla characters and numerals respectively, if we consider first two choices of the recognition results. The detail recognition results are given in Table 2.

Table 2. Recognition results based on different choices from top.

<table>
<thead>
<tr>
<th>Top choices</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Character</td>
</tr>
<tr>
<td>1</td>
<td>91.13%</td>
</tr>
<tr>
<td>2</td>
<td>96.23%</td>
</tr>
<tr>
<td>3</td>
<td>97.68%</td>
</tr>
<tr>
<td>4</td>
<td>98.26%</td>
</tr>
<tr>
<td>5</td>
<td>98.39%</td>
</tr>
</tbody>
</table>

We also analyzed the error versus rejection rate of the classifier and the results are presented in Table 3. Table 3 depicts that for Bangla numeral, 1.57% error occurs at a rejection of 0.13% and only 0.85% error occurs when rejection was 1.10%. Table 3 also depicts that for Bangla character 6.56% error occurs at a rejection of 3.79% and only 2.64% error occurs when rejection was 17.27%.

Table 3. Rejection versus error rate obtained for the Characters and Numerals.

<table>
<thead>
<tr>
<th>Characters</th>
<th>Numerals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rejection</td>
<td>Error</td>
</tr>
<tr>
<td>3.79%</td>
<td>6.56%</td>
</tr>
<tr>
<td>7.44%</td>
<td>5.12%</td>
</tr>
<tr>
<td>12.28%</td>
<td>3.69%</td>
</tr>
<tr>
<td>17.27%</td>
<td>2.64%</td>
</tr>
</tbody>
</table>

Recognition accuracy of about 96% (99.6%) is obtained in character (numeral) recognition, if first two choices are considered. These different choices in recognition accuracy will be very helpful in designing a complete recognition system on word/sentence level where we will be able to take the help of dictionary or we can predict alternate results. From the experiment we noted most of the errors occurred because of similar shape characters. Maximum error occurred between ‘’ and ‘’ and it is noted that about 1.067% cases they mis-recognized one as the other. Main difference between the above two characters is that there is a small loop in left bottom side of one of the characters. Some times during handwriting people do not give this loop and hence mis-recognition occurs. The next erroneous character pair in Bangla is ‘’ and ‘’. They mis-recognized in 0.93% cases. The main difference between the above characters is also the presence of loop in the lower half of the character and sometimes people do not give this loop during handwriting. The five most erroneous character pairs obtained from our experiment in Bangla are shown in Figure 5. Here erroneous character pairs are shown in the first row. In the second row the percentage of error and how many times they are mis-recognized are given. For example character ‘’ is mis-recognized as ‘’ in 19 cases, and the characters ‘’ is mis-recognized as ‘’ in 11 cases. So they mis-recognized between them in 30 (19+11) cases and their percentage of error is 1.067%.
We compared our work with the work of Garain et al. [10]. For numeral recognition they have used a dataset of 400 numerals for their work. They captured their data by Wacom tablet and did not used mouse for data collection. For feature extraction they have used angular variation in 8 directional code and Euclidian distance of each sample point. Nearest neighbor classifier has been used by them for recognition. A comparative result on numeral recognition is shown in Table 4. Garain et. al. obtained 97.43% accuracy on training data of size 400 and they did not report any test result. We obtained 99.86% accuracy on training data and 98.42% on test data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garain et. al. [10]</td>
<td>400</td>
<td>97.43%</td>
</tr>
<tr>
<td>Proposed</td>
<td>2500</td>
<td>99.86%</td>
</tr>
</tbody>
</table>

Comparative results of character recognition is shown in Table 5. Garain et. al. [10] have used a dataset of 2440 character collected using Wacom tablet for their work and obtained 96.34% accuracy on training data. We obtained 98.73% accuracy on training data and 91.13% on test data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garain et. al. [10]</td>
<td>2440</td>
<td>96.34%</td>
</tr>
<tr>
<td>Proposed</td>
<td>12500</td>
<td>98.73%</td>
</tr>
</tbody>
</table>

6. Conclusion

This paper presents a quadratic classifier based system for the recognition of online Bangla handwriting. A modified and robust feature extraction technique is proposed here which can be used for the scripts with Matra/Shiorekha at the upper part. We tested the proposed system on 15000 samples and obtained encouraging results. Not much work has been done towards the online recognition of Indian scripts in general and Bangla in particular. So this work will be helpful for the research towards online recognition of other Indian scripts.

References