Bangla off-line Handwritten Character Recognition Using Superimposed Matrices

Ahmed Shah Mashiyat
Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh.
mashiyatkucse@yahoo.com

Ahmed Shah Mehadi
Pioneer computers, Sheltech Sierra, New elephant road, Dhaka-1205, Bangladesh.
a_mehadi@yahoo.com

Kamrul Hasan Talukder
Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh.
kamrul9375@yahoo.com

Abstract

This paper presents an off-line recognition system for Bangla handwritten characters using superimposed matrices. It is observed that, in all cases, the same character written by different individuals shows at least a minimum level of similarity. In this system, the Bangla text, accepted as an image file, is first segmented into lines and words and then each word is segmented into characters. Then the boundary of each character is determined. The characters are scaled to a standard size using an image scaling algorithm and are stored in a 32X32 matrix. This matrix is then compared with a knowledge base where all recognized characters given by various persons are stored in superimposed form. Finally, depending on the similarity of the character with the stored one, the system recognizes the character to use in the output. This system is suitable to convert handwritten texts into printed documents.

Keywords

Text segmentation, character recognition, superimposed matrices, pattern recognition, water reservoir principal.

INTRODUCTION

Bangla is one of the richest languages of the world. More then 200 million people use Bangla as their medium of communication. So, scientists all over the world are trying to computerize the Bangla language. And in this trend character recognition plays an important role. Although printed Bangla character and numerals recognition is on its way to being solved, producing excellent recognition rates, researchers concentrating on the recognition of handwritten words cannot boast the same success. This has been ascribed to the difficult nature of unconstrained Bangla handwriting, including the diversity of character patterns, ambiguity and illegibility of characters and the overlapping nature of handwriting. But the research is on its way. The character recognition process is accomplished by using DP approach [1], Intelligent regional search [3], Artificial neural network [4,5,8,9], Self-Organizing Maps [10], Hidden Markov models [11] with satisfactory accuracy.

In this paper, we have proposed a system which use supper imposed matrices for character recognition. Due to the cursive ness and touching characteristics of the handwritten Bangla text, we have focused on the segmentation phase of the text document which is to be processed. Our system takes the page of a handwritten document as an input, detect the boundary of the document, segment the document into lines, words and characters, detect the boundary of the character, remove noise, Scale the character, extract the features, identify the class, search for matching and finally print the recognized character. The system block diagram is given in Figure1

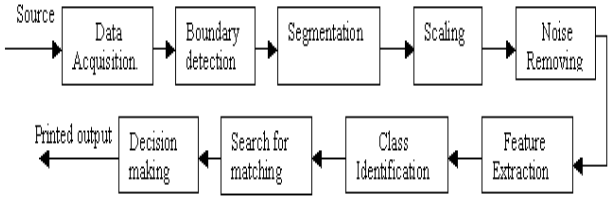

Figure 1: A block diagram representing the system

BANGLA CHARACTER SET IN OUR VIEW

Bangla has 50 letters in alphabet of which 11 letters are vowel (Sorborno) and 39 letters are consonant (Banjonborno). There are also 10 vowel modifiers (i.e Kar) and 7 consonant modifiers (i.e Fala) and 10 digits in Bangla character set. Besides these, there are more than about 253 compound character composed of 2, 3, or 4 consonants (200 compound characters composed of 2 consonants, 51 compound characters composed of 3 consonants and 2 compound characters composed of 4 consonants) [6]. As a result, the total number of pattern to be recognized is more than 310. It is very difficult to recognize a single character form the large number of characters. To get better performance of the system we have employed a grouping concept. We defined a group as a class. There are many similar features in the characters. However, some very distinct features have also been seen in some characters that make them completely different from others [5]. These features could help us in forming the classes. We have classes with 'Matra', left vertical line, right vertical line, upper part, disjoint part, lower part and with their composites. Some classes are shown in Figure 2.

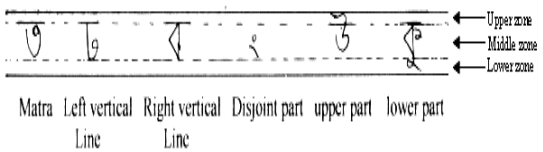

Figure 2: Features of some class

RECOGNITION SYSTEM

The main phase of the recognition system can be divided as segmentation of text into characters and identify the characters. The whole process is described as follows:

Data Acquisition

The input data of the recognition system is acquired by scanning a plain white paper containing black handwritten Bangla text. The scanned paper is then saved as .bmp or .jpg extension. Scanning is done in monochrome mode. The image is then passed for boundary detection.

Boundary Detection

It is necessary to find out the processing area of the text document to speed up the system. We employ horizontal and vertical scanning from the upper left and the bottom right to extract a processing window. The scanning is halted when it faces a single pixel. After boundary detection the processing area is passed to the segmentation phase.

Segmentation of Text

Text segmentation is a process where the text is partitioned into its elementary entities. i.e. characters. The total performance of the handwritten Bangla character recognition process depends on the accuracy of the segmentation process of the text into the characters. In the segmentation phase, first the document is segmented into text lines. Then the text lines are segmented into words and the words are divided into characters.

Segmentation into Lines

The global horizontal projection method computes sum of all black pixels on every row and constructs corresponding histogram. Based on the peak / valley points of the histogram individual lines are segmented [13]. Though this process is suitable for printed document segmentation, it fails to segment the handwritten one. Because: (a) the lines may have deviations (b) two consecutive lines can be so close or overlapping. To overcome this problem we have employed a piecewise projection method. In this process we divided the document into N vertical strips. By horizontal histogram of the strips we determine the local baseline and local headline of text lines. We define the local lines as vertices. For consecutive two lines we determine the lower vertices of the upper text line L_1= (X_1, X_2, X_3……X_n) and the upper vertices of the lower text line L_2= (Y_1, Y_2, Y_3……Y_n). We took the lower vertex of the 1st line and the upper vertex of the 2nd line of a stripe and determine a rectangle. If we get no pixel value in the rectangle then the vertices are shifted inward d/2 distance vertically where d is the distance between the vertices. If there is a pixel value then the upper vertex is omitted. We have employed the process in all stripes and get a single vertex between two consecutive lines for each stripe.

Now we start joining the vertices from the upper left vertex of the document and segmented the document into lines. Figure 3 illustrates the process.

Segmentation into Words

We have observed that in Bangla handwritten text there is a minimum gap between two consecutive words. For word segmentation from the text line we compute the corresponding histogram of the vertical projections of the text line. Depending on the threshold (d) distance between two consecutive zero break cost of the histogram we determine the words. The process is illustrated in the Figure 4.

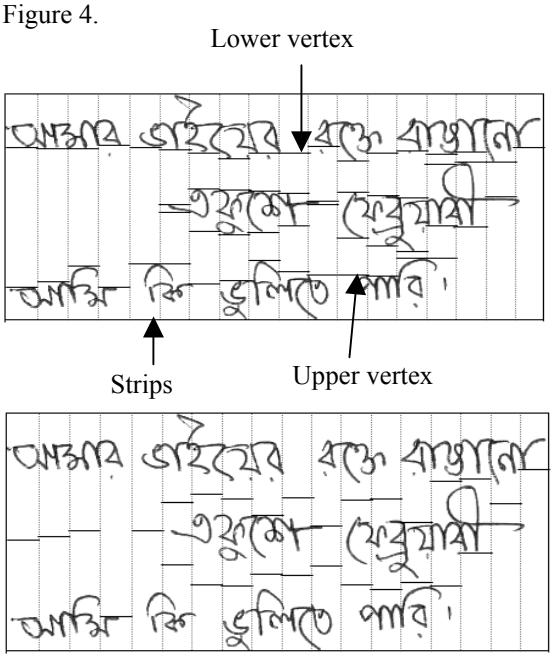

Figure 3: segmentation of document into lines

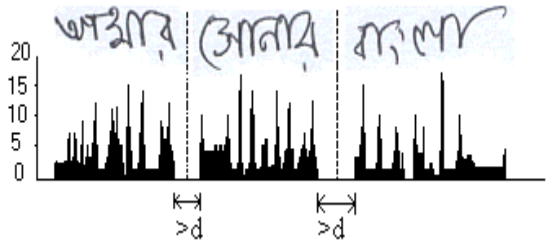

Figure 4: word segmentation

Segmentation into characters

For character segmentation from word we have used water reservoir principle describe in [13]. In our analysis we have observed that in Bangla handwritten text two different characters will not touch each other in two points. In 98.2% cases connected characters touch near "matra". So Bangla words create large bottom reservoir in the middle zone. If any character creates no bottom reservoir with its left or right characters then it can be segmented easily. On the other hand if reservoirs are created then the base area points of the reservoirs are located. The line between two consecutive feature points of a reservoir is just divided to extract the characters. We have employed a threshold height (H_T) to take the base area points. The illustration is in the Figure 5.

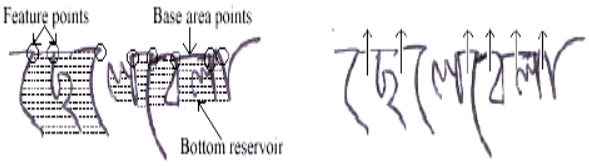

Figure 5: Character Segmentation

Scaling

After getting the segmented character we have detected the boundary area of the character and scaled it to a uniform size of 32X32. For this purpose we have employed two simple equations

$$X_{new} = X_{old} * F_x$$

$$Y_{new} = Y_{old} * F_y$$

Where horizontal scaling factor, $F_x = 32/image\_height$; X is any horizontal unit and

Where vertical scaling factor, $F_y = 32/image\_width$; Y is any vertical unit.

Feature Extraction

Feature extraction is the process of extracting essential information content from the image segment. It plays an important role in the whole recognition process. We generate a 32X32 matrix according to the height and weight of the bitmap image by the color value of the pixels. The black pixels are considered as 1's and the white pixels are as 0's. Figure 6 shows a converted feature matrix of 16x16 matrix format.

Figure 6: Segmented character and its feature matrix (16X16) representation

Noise Removing

Noise is a random error in pixel value. We have filtered out the noise from the character image for better recognition. It is observed that for Bangla text we cannot remove wide pixels from the upper or lower part of the character. Because removing pixels from the lower part eliminates the difference between 'ব' and 'র' and for some other characters. So we have just removes the wide portions in the middle zone.

Class Identification

The heuristic method is guided by intuition and experience. Identification of the character class is a pre-recognition step. A heuristic search through the feature matrix of the segmented character and finding

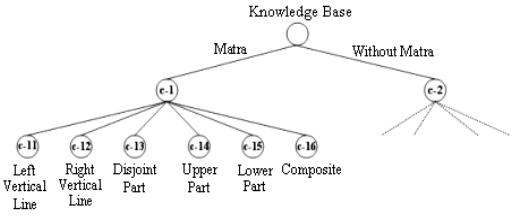

Figure 7: class identification tree

the noted high level feature, the recognition process can easily determine that the segmented character belongs to which group [5]. This grouping is used in the matching phase. Figure 7 demonstrates the class identification.

Knowledge Base

We have designed the knowledge base based on the feature matrix of various characters. In order to build the knowledge base we have repeat the step 4.1 to 4.4 for each of the 310 characters. Characters written by 10 different individuals are taken for knowledge base. The feature matrices of a character are supper imposed in a 32X32 matrix and stored at the knowledge base as a reference. The characters are stored in the knowledge base in a sequential manner considering their pixel density from top to bottom. This is ascribed to employ a binary searching in the matching phase.

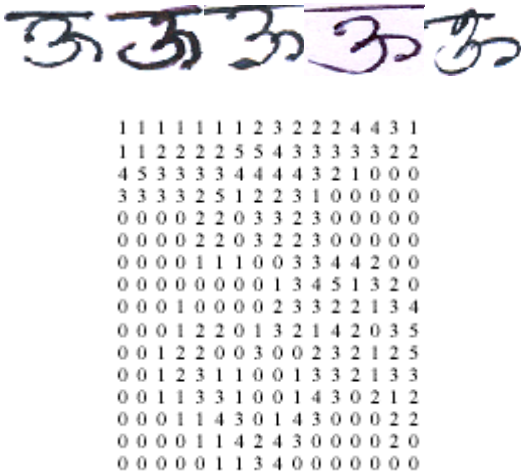

Figure 8: 5 Reference Characters and their superimposed resultant matrix

Matching

The matching technique for the character recognition can be divided into three steps.

1. Initial matching.
2. Horizontal matching and
3. Vertical matching.

Initial Matching

Initial matching is performed only by superimposition of the feature matrix of the test character upon the character matrices in the knowledge base stored earlier. The superimposition is performed to count the number of times a value 1 in the feature matrix as against the resultant matrix position which is of a value 0. If the number of count is N and the summation of all positions of the resultant matrix is M then percentage of error can be calculated as $\in_i = (N/ (M/10))*100\%$ for each of the resultant matrices of the knowledge base. (M/10 is used because M holds the summation of all positions having value 1 for the 10 sample image.) We take the lowest erroneous character as our target character. And move to next phase.

Horizontal Matching

We divided the feature matrix and the initially recognized resultant matrix into k groups where each group is composed of 32/k rows for performing horizontal recognition. In the experiment we have taken K=8. Let ϕ_i and ψ_i be the summation of all position values found in the i^{th} group of the feature matrix and the resultant matrix respectively. Here i=1,2,3,...,K. Since we have selected K=8, so i might be at least 1 and at most 8 and in any of the group, there will be exactly 4 rows which is illustrated by Figure 9.

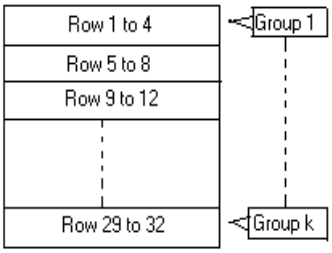

Figure 9: Horizontal matching method.

For a true character ϕ_i must be close to (ψ_i /10). (Since ψ_i is the summation of all position values found for 10 identical feature matrices of group i) Now percentage of horizontal error for group i can be calculated as

$$\in_{hi} = (1/(M/10))*\Sigma(abs((\psi_i/10) - \phi_i))*100\%.$$

Here i=1,2,3,…….,K. Where abs( ) indicates absolute value.

Again we consider percentage relative error $\in_{hc}$ and if $\in_{hi} > \in_{hc}$ then we consider the i^{th} group is discarded. Now we count the total number of discard groups found. If the number of count is N then we calculated with the equation

$$\in_{hg} = (N/K)*100\%.$$

There is another percentage relative group error consideration $\in_{gc}$ and if $\in_{hg} > \in_{gc}$ then the resultant matrix is declared as no matched and go to 4.6.a for another otherwise go to section 4.6.c .

Vertical Recognition

Like previous section vertical recognition is performed. The only difference is that instead of grouping the rows here we group columns. As usual percentage of vertical error $\in_{vi}$ along with error consideration $\in_{vc}$ for each group i as well as group error $\in_{vg}$ along with group error consideration $\in_{gc}$ is taken. Now if $\in_{vg}$ is below $\in_{gc}$, i.e. $\in_{vg} <= \in_{gc}$ then we declare the character as recognized and pass for print.

EXPERIMENTAL RESULTS AND DISCUSSION

We have divided our system in two main phases: segmentation and matching. So the overall performance of the system directly depends on the performance of the two individual phases. In the segmentation phase we have tested handwritten text document of our classmates and teachers. The performance of the segmentation is very satisfactory. The experiment results are shown in the Table 1.

Table 1: Segmentation Result

No. of text document	Line segmentation rate	Word segmentation rate	Character segmentation rate
5	<=96%	97%	94%
15	97.5%	98.7%	93%
20	98.4%	99%	94.5%
25	99.7.%	99.4%	95.6%

In our system we have used very simple matching technique. So at the initial stage the recognition rate was very poor. So we have employed a grouping concept and get better result. The recognition rate was increasing with the number of class. With the same data for segmentation we have done the matching experiment. The result of the experiment is show in the table 2.

Table 2: Recognition Result

No. of sample	No. of class	Correct recognition	Wrong recognition	Unrecognized
120	4	55%	25%	20%
120	6	67%	18%	15%
120	8	75%	11%	14%
120	10	81%	8%	11%
120	14	90%	4%	6%

Some documents are so cursive and unconstrained that we can not even recognize them manually. We omitted them for the experiment.

CONCLUSION

The ultimate goal of the handwritten recognition system is to achieve 100% accuracy. But it is illusionary to reach the target because even human beings are not able to recognize every handwritten text without any doubt, e.g. it happens to most people that they sometimes cannot even read their own notes. There will always be an obligation for the writer to write clearly. Human roughly recognizes 96% [1].

In this paper, we have presented a recognition system emphasizing on the segmentation phase. The matching technique employed here is very simple. But the grouping concept reached the recognition system at a satisfactory accuracy level. Our main achievement is that we have processed all the Bangla character (“Shorborno”, “Banjonborno”, Numerals and compound character) in a single platform.

REFERENCES

[1] Al Mamunur Rasid and Muhammad Masroor Ali, *On Line Recognition of Handwritten Characters Using 1-Dim, 1-Dim DP approach*, Proc. ICCIT’1998 , 18-20 December, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.pp.86-90.

[2] A. F. R. Rahman and M. A.Sattar, *An Analysis of Bengali characters: Detection of Some Characteristic*

Features, Proc. ICCIT'1998 , 18-20 December, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.pp.142-145.

[3] Mohammed Ali Asger Moshad and Muhammad Masroor Ali, *Recognition of Handwritten Bangla Digits by Intelligent Regional Search Method*, Proc. ICCIT'2001, 28-29 December, Dhaka University, Dhaka, Bangladesh.pp.297-302.

[4] Md. Morshedul Arefin, Md. Khademul Islam Molla, Md. Latifur Rahman and M. Ganjer Ali, *Size Independent Bangla Optical Character Recognition System*, Proc. ICCIT'2001, 28-29 December, Dhaka University, Dhaka, Bangladesh.pp.314-318.

[5] A.O.M. Asaduzzaman, Md. Khademul Islam Molla and M. Ganjer Ali, *Printed Bangla Text Recognition Using Artificial Neural Network with Heuristic Method*, Proc. ICCIT'2002 , 27-28 December, East West University, Dhaka, Bangladesh.

[6] Minhaz Fahim Zibran, Arif Tanvir, Rajiullah Shammi and Ms. Abdus Sattar, *Computer Representation of Bangla Characters And Sorting of Bangla Words*, Proc. ICCIT'2002 , 27-28 December, East West University, Dhaka, Bangladesh.

[7] Shariful Hasan Shaikot, Fahim Kawsar, Md. Shahariar Saikat, *Bangali Digit Recognition System using 3-layer Backpropagation Neural Network*, Proc. ICCIT'2003, Jahangirnagar University, Dhaka, Bangladesh.pp. 327-331.

[8] A. O. M Asaduzzaman, Mst. Shayeala Parven and M. Ganjer Ali, *Detection of Bangla Number Using Artificial Neural Network*, Proc. ICCIT'2003, Jahangirnagar University, Dhaka, Bangladesh.pp. 347-350.

[9] Md. Rezaul Bashar, Mirza A. F. M Rashidul Hasan, MD. Farukuzzaman Khan, *Bangla off-line Handwritten Size Independent Character Recognition Using Artificial Neural Networks Based On Windowing Technique*, Proc. ICCIT'2003, Jahangirnagar University, Dhaka, Bangladesh.pp.351-354.

[10] Md. Badrudoza, *Recognition of Bengali Handwritten Letters Using Self-Organizing Maps(SOM)*, Proc. ICCIT'2003, Jahangirnagar University, Dhaka, Bangladesh.pp.355-359.

[11] Md. Shafiul Azam Howlader, Md. Mostafa Jamal and Gahangir Hossain, *New Approch of Bangla Handwriting Recognition*, Proc. ICCIT'2003, Jahangirnagar University, Dhaka, Bangladesh.pp.370-373.

[12] Jalal Uddin Mahmud and Chowdhury Mofizur Rahman, *Handwritten Bengali Digit Recognition by Improved feature Analysis and MLP network*, Proc. ICCIT'2003, Jahangirnagar University, Dhaka, Bangladesh.pp.380-384.

[13] U. Pal and Sagarika Datta, *Segmentation of Bangla Unconstrained Handwritten Text*—Proceedings of the Seventh International Conference on Document Analysis and Recognition(ICDAR-2003).

[14] Mahbub Murshed, S. M. Saifur Rahman, K. M. Hasan Al Noor- *Hand written signature verification system using Artificial Intelligence* B.Sc. thesis paper, Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh.

[15] J. K. Parker, *Algorithms For Image Processing And Computer Vision*, Wiley Computer Publishing.1st Edition, 1997. pp. 275-303.