Bangla Hand Written Character Recognition

Md. Shahiduzzaman

Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Mirpur, Dhaka, Bangladesh

Abstract: ***This paper discusses the different methods for optical character recognition (OCR), which has been an important field to research from a few decades due its huge necessity to convert paper documents or books in computer readable format. Though Bangla (widely used as Bengali) is one of the top uses language among the other languages, but there is no reliable character recognizer for this. Our work has covered a total process to develop a complete OCR, especially for feature extraction process, which is very important to recognize characters correctly. Here, we have developed and tested many algorithms to identify each ones merits and limitations in various cases for hand written character recognition to make the stage more optimized. Moreover, we have used hidden Markov model (HMM) classifier along with artificial neural network (ANN) to make our classifier more accurate.***

Keywords: OCR; ANN; HMM; features

1. Introduction

In computerization of any language, one of the vital tasks is to develop an efficient and effective optical character recognition (OCR) system for the respected language. In order to store million pages of paper documents into electronic form, OCR is the key tool. Otherwise, if those are entered by typing manually, the efficiency, effectiveness and correctness will drastically fall down. Character recognition system can be of two types: Online and Offline, where online character recognition software is available on the internet and is used directly online by people who need OCR performed on files of various formats, but offline character recognition depends on the documents. Again, there are two types of written document, where one is handwritten document and the other is printed document. Though Bangla (widely used as Bengali) is one of the top uses language among the other languages, but there is no reliable character recognizer for this. In this study, we concentrate on recognition of handwritten Bangla character. It is claim that our study constructs an optical character recognition (OCR) system for handwritten Bangla characters, which focuses on various strategies.

Researchers have been working on Bangla OCR since 1990 [1-4]. A great amount of work has been done by B. B. Chaudhuri and U. Pal [1]. Following them some other researchers have come up with a variety of innovative ideas. Institutes of India and Bangladesh have conducted different projects and research works, but commercially standard Bangla OCR is not available still now. The currently available OCR systems are: BOCRA-2006 [5], AponaPathak-2006 [6], BanglaOCR-2007 [7]. BOCRA has no framework and it does not work for handwritten text. AponaPathak has complete framework but it is not freeware. BanglaOCR is currently the only open source (OCR) software for the Bangla script developed by the Center for Research on Bangla Language Processing (CRBLP) but it does not handle documents with pictures and tables. For this research many published research about Bangla OCR have been studied [8- 16].

This study develops a method for the recognition of handwritten Bangla characters using the neural network, where the preprocessing steps includes segmentation and binarization. In the feature extraction stage, features are extracted using different feature extraction techniques and in the final stage, a multilayer feedforward neural network and recurrent neural network is used to classify and recognize characters. Here, the largest difficulty of handwritten Bengali character recognition is the immense variation of the way handwriting, where many factors can account for the diversity in Bengali handwriting styles, among others the regional origin of the writers, their educational level and their profession. In fact different circumstances such as stress, fatigue, and hurry affect the handwriting of any individual. Thus we especially deal with feature extraction by considering various cases and classification techniques. To have an efficient system we have collected 300 samples written by various persons where the writers were told to write each character naturally. With the development of the system the door of many areas of research works can be opened.

Description on OCR is given in section II. In section III we have discussed about different works on handwritten Bangla OCR. Section IV gives details about corpus. Section V discusses the experimental results and analysis of hand written Bangla OCR. We concluded our paper in section VI with future works.

2. Optical Character Recognition

OCR is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used to convert books and documents into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation, text-to-speech and text mining to it. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. A typical OCR system contains three logical components: an image scanner, OCR software and hardware, and an output interface. The image scanner optically captures text images to be recognized. Text images are processed with OCR software and hardware. The process involves three operations: document analysis (extracting individual character images), recognizing these images (based on shape), and contextual processing (either to correct misclassifications made by the recognition algorithm or to limit recognition choices). OCR software attempts to identify characters by comparing shapes to those stored in the software library or database. The software tries to identify words using character proximity and will try to reconstruct the original page layout. High accuracy can be obtained by using sharp, clear scans of high-quality originals. The output interface is responsible for communication of OCR system results to the outside world.

The OCR system will support various functionalities in terms of proposed specifications. The process of optical character recognition has following five stages given in Figure 1.

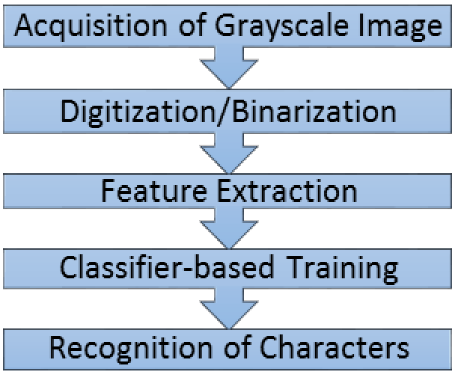

Acquisition of Grayscale Image
Digitization/Binarization
Feature Extraction
Classifier-based Training
Recognition of Characters

Figure 1: Block Diagram of Character Recognition Process.

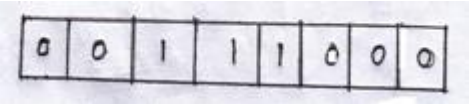
3. Different Works on OCR

As discussed in the previous section feature extraction is the pre-stage of the classification phase. This stage is very important, because final result largely depends for precise recognition. Feature extraction is a bit more challenging for handwritten character recognition due its immense variation of writing font, while printed character font for a particular style is always same. We have tasted a various mechanism to make a reliable feature extraction for various styles of handwriting characters.

- Number of Island Method
- Position Value Method
- Horizontal Vertical Value Method
- Centre Distance Method
- Median Value Method

A single feature extraction method alone is not sufficient to obtain good discrimination power. An obvious solution is to combine features from various feature extraction methods. Different feature extraction methods are used to make the output more robust and accurate.

A. Number of Island Method



0	0	1	1	1	0	0	0

Figure 2: 1x 8 matrixes as a block.

In our binary image, we have a 128 by 128 matrix. By considering 1x8 cells as a chunk as shown in Figure 2 we convert in 128x16 matrix. Then count in a vertical direction according to following function:

$$f(x)= \begin{cases} 1 \text{ , if } \sum \text{ column is greater than 1} \\ 0 \text{ , Otherwise} \end{cases}$$

For each chunk, if the total number of 1 is more than one the function return 1 as output. New converted matrix size is 128x16. Count each column vertically, thus feature vector sixe is 1x16.

B. Position Value Method

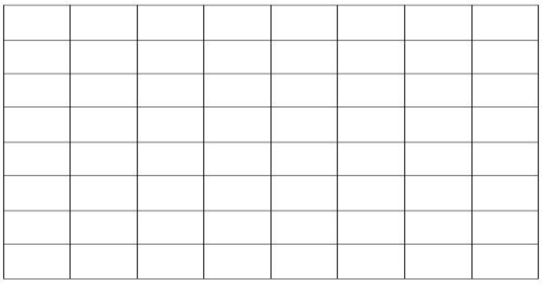
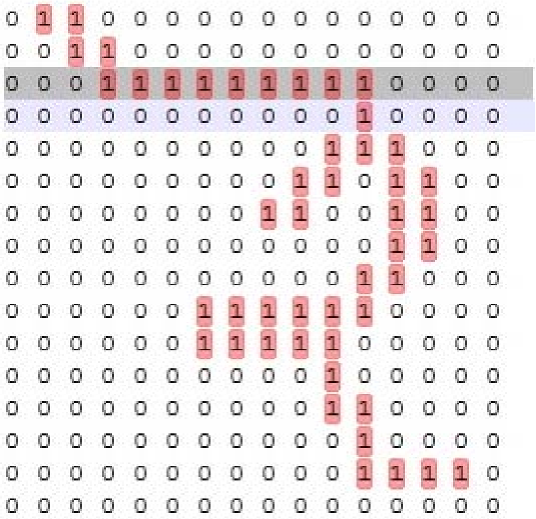


Figure 3: 8x 8 matrixes as a block

From binary file, we have considered 8x8 block shown in Figure 3 to convert new matrix size 16x16. For each 8x8 matrix block, weighted value is used depending column position where the cells contain 1 for following function:
weighted value w(j)= 1 / 10^j
where j is a range of value 1 to 8 , the value of j depending on column position how nearer to the center. Here, j = 8 - column_position % 8 for first half means column position 0 to 64 and for second half, j= 1 + column_position % 8.

$$f(x)= \begin{cases} 1 \text{ , if } \sum \text{ 16x16 weighted box value is} \\ \quad \text{greater than 0.00001} \\ 0\text{, Otherwise} \end{cases}$$

f(x) represents each individual 8x8 matrix block. Thus feature vector size is 16x16 shown in Figure 4.



0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0
0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4: Positional 16x16 feature vector.

C. Horizontal Vertical Value Method

We have counted the number of cell contains 1 both horizontal and vertical direction, by means to convert 128x1 and 1x128 matrixes respectively. Finally merge the two matrixes for 1x256 feature vector.

D. Centre Distance Method

From the binary file, we have considered 16x16 block to convert new matrix size 8x8. For each 16x16 matrix block we have calculated a centroid according to row and column

numbers. Then we find distance among centroid points with each cell contains 1 of the block. Summation of all distances represents the 16x16 block's value. Thus the 128x128 matrix reduces to 8x8 matrix then it is converted as one dimensional 1x64 matrix.
center x=(start index + last index) of a row/2;
center y=(start index + last index) of a column/2;.

E. Median Value Method

From the binary file, we have considered 16x16 block to convert new matrix size 8x8. For each 16x16 matrix block, cells are only considered whose column position in the range of 6 to 11. If any cell of the range contains 1 then the particular row represent 1. Summation of all row's value represents the 16x16 block's value. Thus the 128x128 matrix reduces to 8x8 matrix then it is converted as on dimensional 1x64 matrix.

$$f(x)= \begin{cases} 1 \text{ , if } \sum \text{number of cells contains 1 is greater than 0 in column position 6 to 11} \\ 0\text{, Otherwise} \end{cases}$$

4. Corpus for OCR

We have used Supervised Training method. We have used 300 samples for training the net. We have used both biased and unbiased training strategies. For different feature extraction methods we have changed the layer size several times to check the accuracy level. We have supplied 100 unknown samples to the neural net for testing.

5. Experimental Result and Analysis

In the experiment OCR was performed on 300 individual segmented Bengali character images. The proposed algorithm was implemented using C++ and java in MS visual studio 2010 and Netbeans 7.0 respectively. Table 1 shows the accuracy using the five investigated methods using the Bengali vowel. It is observed from the table that the test gives 80% accuracy for most of the cases using the number of islands method. It is observed that the number of Island method and position value method are the best.

Table 1: Comparison of Outputs Amoung the Five Proposed Feature Extraction Method.

Characters	Percentage of Classification				
	No. of Islands	Position Value	Horizontal Vertical Value	Center Distance	Median Value
অ	80	80	60	100	100
আ	60	50	55	50	60
ই	80	100	55	57	61
ঈ	62	63	53	50	58
উ	80	54	70	67	65
ঊ	80	74	73	63	63
এ	80	60	60	51	60
ঐ	80	100	60	60	62
ও	61	67	69	68	70
ঔ	60	80	68	78	80

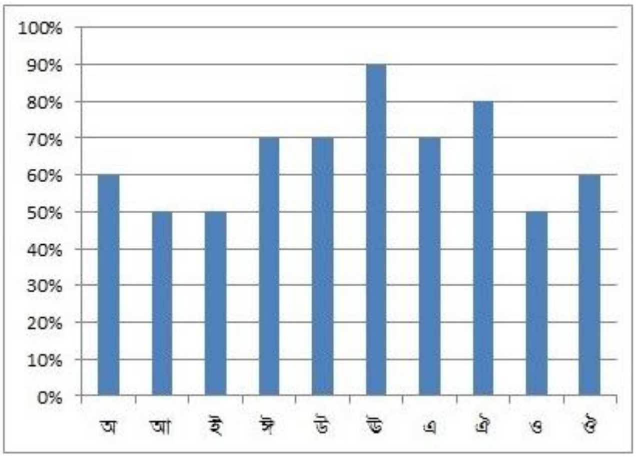


Figure 5: Results using hybrid feature extraction of number of island and position value method.

Then we combined both the Number of Island Method and Position Value Method to obtain the better performance. The accuracy of our combined method is given in Figure 5. It is observed from the figure that ঊ provides 90% accuracy in this case.

We have used Multilayer Neural Network (MLN) as our first classifier. Then we have added Nearest Neighbor (NN) classifier using the Euclidean Distance Measure. The comparison between the hybrid method (MLN+NN) and number of island method is shown in Table 2. From the Table 2 it is shown that the hybrid classifier has improved classification result over the method implemented by a single Multi-layer Neural Network.

Table 2: Head to Head Comparion Between Single Classifier and Hybride Classifier.

Bangla Character	Percentage of Classification	
	Single Classifier (Multilayer Neural Network)	Hybrid Classifier (Multilayer Neural Network + Euclidian distance)
অ	80	*80*
আ	57	*67*
ই	80	*80*
ঈ	56	*71*
উ	66	*67*
ঊ	60	*60*
এ	60	*60*
ঐ	60	*60*
ও	51	*54*
ঔ	71	*80*

6. Conclusion

This paper presented an OCR incorporating the various feature extraction methods based on the different kinds of font shapes due to the enormous variation of writing styles. We have used single and hybrid classifier incorporating Neural Network, Nearest Neighbor and HMM. The research project is not free from any shortcomings. We have experimented only on individual characters. No compound characters have been considered. The system is not working for a complete Bangla document. Characters must be written and processed separately. From the test result 80% accuracy on an average is obtained which is not adequate. The ideal goal of designing a handwritten character recognition method with 100% accuracy is illusionary, because even human beings are not able to recognize every handwritten text. Humans roughly recognize 96%. But we are trying to increase the rate at least 92-95% by implementing more classifiers based on different methodologies.

With the development of a Bangla character recognition system the door of many areas of research works can be opened. It can be used in the fields of Bangla Spelling Checker, Grammar Checker, Font Converter and so on. We have developed our system for single Bangla character recognition. But with some modification it can also recognize the compound words. We have also some recommendations for future work. The percentage of recognition can be improved by combining different classifiers based on Support Vector Machine (SVM) and Recurrent Neural Network (RNN). A complete Bangla document can be recognized with better improvement of this technique. This method can also be used to recognize Bangla numerals. Signatures can be recognized by improving this technique.

References

[1] An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)
[2] Chaudhuri, B.B.; Pal, U. Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR '97) Volume: 2 , 1997 Page(s): 1011 -1015 vol.2.
[3] On Recognition of touching Characters in Printed Bangla Documents: Uptal Garain,B.B. Chaudhuri Proc. Fourth International Conference on Document Analysis and Recognition, ICDAR - 97, Germany, August 18-20, 1997, pp. 1011-1016.
[4] Segmentation of Bangla Hand-written text into characters by recursive contour following: A.Bishnu and B.B. Chaudhuri
[5] Proceedings of Int. Conf. on Document Analysis and Recognition, Bangalore, India, September 20-22, 1999.
[6] Recognition of Handwritten Bengali Characters: M.C. Fairhurst, A.F.R. Rahman, R. Rahman Pattern Recognition Journal May 2002
[7] BOCRA [2006]. http://bocra.sourceforge.net/doc
[8] Apona-pathak[2006].http://www.apona-bd.com/apona-pathak/bangla-ocr-apona-pathak.html.
[9] BanglaOCR[2007]. http://sourceforge.net/projects/blp/files/BanglaOCR
[10] A. K. Roy and B. Chatterjee, "Design of a Nearest Neighbor Classifier for Bengali Character Recognition", J. IETE, vol. 30, 1984.
[11] Indo-Bangladeshi Language", Proc. of 12th Int. Conf. on Pattern Recognition, IEEE Computer Society Press, pp. 269-274, 1994.
[12] B. B. Chaudhuri and U. Pal, "OCR Error Detection and correction of an Inflectional Indian Language Script", Proceedings of ICPR, 1996.
[13] An Optical Character Recognition (OCR) System For Printed Bangla Script", Indian Statistical Institute, 1997.
[14] B. B. Chaudhuri and U. Pal, "A Complete Printed Bangla OCR System", Pattern Recognition, vol. 31, pp. 531-549, 1998.
[15] B.B.Chaudhuri and U.Pal, “A complete Bangla OCR System”, Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, 1998.
[16] Ahmed Asif Chowdhury, Ejaj Ahmed, Shameem Ahmed, Shohrab Hossain and Chowdhury Mofizur Rahman"Optical Character Recognition of Bangla Characters using neural network: A better approach". 2nd International Conference on Electrical Engineering (ICEE 2002), Khulna,Bangladesh.
[17] Jalal Uddin Mahmud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman, "A Complete OCR System for Continuous Bangla Characters", IEEE TENCON-2003: Proceedings of the Conference on Convergent Technologies for the Asia Pacific, 2003.