2016 International Workshop on Computational Intelligence (IWCI)
12-13 December 2016, Dhaka, Bangladesh

Stroke Matching Based Approach to Recognize Bangla Offline Conjunct Characters

Shamima Nasrin, Tasneem Zerin and Mohammad Mahadi Hassan
Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh
Email: shimu.cse69@gmail.com, tasneem.zerin.chy@gmail.com and mahadi_cse@yahoo.com

***Abstract*—The main target of the approach presented in this paper is to recognize offline Bangla conjunct characters by Stroke Matching Based Approach. At first, the approach simplifies the word into strokes. Later, it matches the strokes with the one that is stored in the dataset. The approach reduces the noise and then works over the strokes. Up to now a lot of recognizers are available to recognize BOC (Bangla Offline Characters). But a recognizer to recognize Bangla Conjunct letters is rare. Here in this research work Stroke Matching Based Approach modifies the existing strokes of Bangla Conjunct Letter to identify it. The recognition of the characters would be done by using neural network system with the matching of strokes of the characters.**

Keywords—Offline characrters; Bangla conjunct characters; Stroke Matching; Artificial Neural Network(ANN)

I. INTRODUCTION

In the natural language rank list Bangla owned its position as number seven as a spoken language since most of the native people of the world communicate using this language. It is originated from Indo-European [1] language. It contains 50 alphabets including 11 vowels that is Shoroborno and 39 consonants that is Banjonborno. And 253 conjunct characters composed of 2, 3 or 4 consonants. Among them 200 characters are formed of 2, 51 characters are composed of 3 and 2 characters are composed of 4 consonants [2]. Bangla language also contains 10 vowel modifiers that is known as Kar, 7 consonant modifier that is known as Fala. Till now Bangla handwritten Pattern Recognition, Character Recognition, Text Summarization etc are not quite rich. And the important part is, in most research, recognition of conjunct letters of Bangla language has been skipped due to its complex patterns. As a result, in many cases, conjunct letters are not clearly recognized. So along with the 50 letters of Bangla, it is also important to recognize them so that they can be recognized easily are stored so that it might turn into a great help for documentation of soft copies.

Most of the offline character recognition are performed using Optical Character Recognition Methods. But the process and way of working is way too hard and lengthy. Though the target of every method is to find a better solution in the character recognition, but every time came with a challenge. Therefore, for the character recognition, the proper segmentation of characters from the text is needed which is difficult because of losing essential data. Bangla Characters also face the same segmentation problem similar to ANN based approaches. So, to identify conjunct letters of Bangla Language Stoke Matching Base Pattern Recognition is established.

This thesis paper work is based on the recognition of Bangla Offline Character Recognition by matching the strokes from an image. The recognition of Bangla characters and text works have been done earlier using different technologies like Optical Character Recognition (OCR), Neural Network , SVM etc. The new approach is Stroke Matching based approach to recognize Bangla Offline Characters for recognizing them by machines. The chapters of this paper represent a go through over the topic in details along with the description of work.

The rest of the paper is organized as follows. Section II is the summarization of some related works. Section III gives the description of methodology of our system. Section IV describes feature extraction of our proposed system. Section V gives the description of experimental analysis and Section VI concludes the paper.

II. LITERRATURE REVIEW

A great amount of work has been done with Offline Character Recognition. The relevant works are briefly discussed below.

An approach proposed in [3] is on the basis of view-based algorithm, where only upper and lower views are analyzed for characters. The character is divided vertically in identical segments depending on the number of points need to be obtained. The calculation of y coordinates decides three elementcharacteristic vector which describes the given character. The found two vectors combining with their two values describes the aspect ratio of the picture which transforms into 8 element vector representing the final character. The accuracy of the proposed scheme was 74.166% and accuracy of recognition is 75%.

Another research work which studies with the Stroke Segmentation and Recognition from Bangla Online Handwritten Text, published in International Conference on Frontiers in Handwriting Recognition, 2012. Here the stroke (collection of points from pen down to pen up) of writing a character is analyzed and Support Vector Machine (SVM) classifier is used for recognizing segmented characters .[3] Correct Segmentation rate 97.89 % and Stroke recognition rate 97.68% [4].

This research work in [5] is on recognition of Bangla Basic character and digits by using convex hull method. They used

978-1-5090-5769-6/16/$31.00 ©2016 IEEE

Graham Scan algorithm for computing the convex hull of numeric pattern. In this research work they used simple 125 features based on different bays and lake attributes of convex hull using MLP based classifier. Recognition rate of this system is 76.86% from 10000 sample of handwritten Bangla text and 99.45% from 12000sample of Bangla numerals.

The proposed approach in this paper is an offline based character recognition system using multilayer feed forward neural networking. A new method "Diagonal based feature extraction" is introduced in this paper [6]. Every character Image of size 90x60 pixels is divided into 54 equal zone which is of 10x10 pixel in size. The extracted zone contain 19 diagonal lines and 19 sub features that obtained from each zone. The recognition system has been implemented using neural network training tool.

This research work proposed a recognition system of printed Bangla text. In this approach the text image is segmented into individual characters and then refined and converted to a designated M x N matrix [7]. Heuristic method is used to obtain better accuracy. Characters are separated into reasonable number of groups by using Heuristic method. The accuracy rate is nearly 75% for this system.

III. STROKE MATCHING BASED APPROACH

The main target of the approach represented in this paper is to recognize off-line Bangla conjunct characters. At first, the approach simplifies the letter into set of strokes. Then it matches these strokes with the one that is stored in the dataset. The approach reduces the noise and then works over the strokes. Up to now a lot of recognizers are available to recognize BOC (Bangla Offline Characters). But a recognizer to recognize Bangla Conjunct letters is rare. SM (Syntactic Method) is one of those popular methods that used to recognize BHC. Here in this research work Stroke Matching Based Approach modifies the existing strokes of SM and uses it to recognize Bangla Conjunct Letters. The proposed system consists of nine steps. The Flowchart of the system is given in Fig. 1.

A. Input Characters

The input characters are inserted as image where the patterns are recognized to identify the character by using Stroke Matching Based Text Recognition which mainly exploits Modified Syntactic Method where it provides a capability for describing a large set of complex patterns by employing small set of simple pattern in strokes.

B. Stroke Generation

After forming the digital 0/1 matrix, the most important step is to contour tracing which means to convert the multi-pixel line into the single pixel line. That means it will filter the line to get a single formed line. The next step is to generate the straight line that is generating the strokes. The Stroke Matching Character Recognizer (SMCR) proposes a to develop an algorithm where offline characters and stroke generation works simultaneously. The pseudo code of the process is shown in *Algorithm 1*.

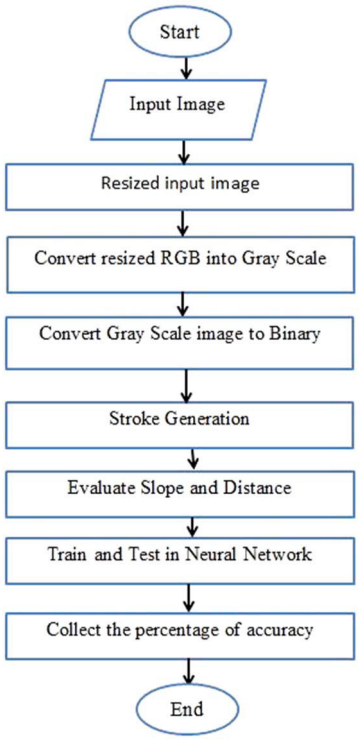


Fig. 1. Flow chart of the Proposed System.

Algorithm 1: Pseudo code for stroke generation

***Step 1*:** Convert the image RGB to Binary.
***Step 2*:** Scaling the image into 120x120 pixel.
***Step 3*:** Apply thinning algorithm.
***Step 4*:** Search '1' in the digital matrix, and set the starting point at this position.
Step 5: Detect the horizontal and vertical line of the image
Step 6: Find the start point and end point of these lines to count them as stroke.
Step 7: Remove the horizontal and vertical lines and go to *Step 6* to evaluate the other strokes.
Step 8: Count slopes and distances between the regarding points.

C. Defined Characters Stored In The Dataset

Every character is defined separately along with their slope and distance between two points in the dataset so that it can be matched with characters that are in the image. Due to this, the confliction with any other character is reduced to low.

D. Recognition

The matching or recognition process includes matching the letter of the image with the one that is stored in the dataset. The process can only be executed after the completion of stroke generation and line drawing.

E. Decision

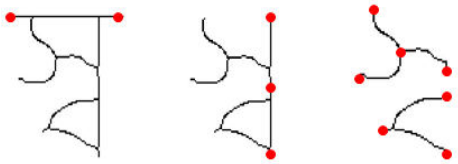
After matching the letter with the dataset letter it comes to an ending decision with a result of percentage of matching. By this the identification process ends here.

IV. FEATURE EXTRACTION

A. Thinning

The thinning algorithm reduces the data size where it extracts the information regarding the shape. It allows working more accurate in recognition. There are different types of methods for thinning. After the process of thinning the horizontal and vertical lines are removed to count the starting and ending point of the strokes.

B. Counting Strokes



(a) Horizontal Stroke (b) Vertical Stroke (c) Other stroke

Fig. 2. Stroke points are showed with red dots in (a), (b) and (c).

Algorithom-2 : Algorithm for counting Strokes is given below

1.FOR i=2 to (Image,1)-1
2. FOR j=2 to (Image,1)-1
3. Pix= certain Pixel of Image
4. IF (Pix(1)==0) &&(Multiplication of Pixel position==0)
5. IF(summation of Pixel Position >= Pix)
6. SET value of i= i
7. SET value of j= j
8. SET count= count+1
9. END
10 END
11. END
12.END

C. Calculating slope

Slope of a stroke can be determine by the coordinates, e. g. (x1,y1) consider as the starting point and (x2,y2) as the ending point of a certain stroke .

- Slope, m = $tan^{-1}(y_2 - y_1)/(x_2 - x_1)$

D. Calculating Distance

Distance between point to point is measured by using Euclidian Method.

- Function Euclidean Distance=CalcDistance(x1, y1, x2, y2)
- Distance=sqrt((x2-x1)^2+(y2-y1)^2)

V. EXPEIMENTAL ANALYSIS

A. Experimental Setup

The recognition process has been worked on Intel Core i5-2.40 GHz processor of 4.0 GBytes of RAM running on Windows 7 operating system. The desktop application has been performed in MATLAB 2013. In recognition, for 4 conjunct characters of 'ত' are represented in a Tree 01 of Fig. 4, a test set of 40(= 4 x 10) character where 10 is the number of sample image has been generated to consider herein. Test set of other characters of total 32 Trees. Since a proper benchmark dataset is not available, we had to make our own dataset where the number of sample image is 10 for each conjunct character.

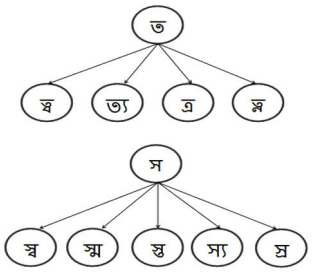


Fig. 3. Sample tree of 2 characters.

The character images were trained and tested in Neural Network to calculate the accuracy of performance. The result has been included in the Experimental Result part.

B. Structure of Neural Network

Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connection.

Here we worked over 50 conjunct characters. The feed forward neural network is trained in a supervised mode. The correct output that corresponds to the data is specified and known as target array. Our vector length is 20. So our input layer has 50 neurons with 20 combinations. So the input is 50×20 array. Since, the output layer depends on the number of characters trained, so in this case the output will be of 50 neurons. The suitable feed forward neural network with 50×20 array input and 50 outputs is shown in the Fig. 4.

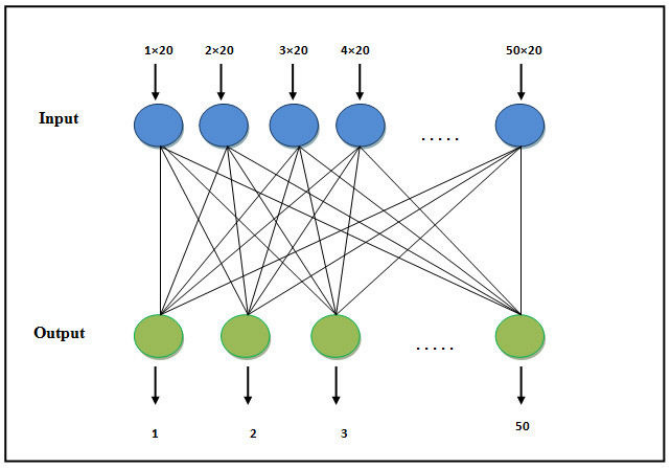


Fig. 4. Design of Neural Network layers for Bangla Conjunct Character recognition.

C. Experimental Results

According to the Tree 01 of Fig. 3, the process at first takes input from scanned file with the extension of JPEG, Bitmap format. The Table I shows the test result with recognition percentage.

TABLE I: TESTING PERCENTAGE USING NEURAL NETWORK

Characters	SAMPLES	RECOGNITION RATE
স্ব	10	98.89 %
স্র	10	91.10 %
স্ম	10	90.5 %
ব্য	10	91.13 %
প্ত	10	91.08 %
প্র	10	90.71 %
ষ্প	10	91.17 %
ম্র	10	90.79 %
ট্ট	10	90.75 %
স্প	10	91.39 %
ত্য	10	91.36 %
গ্ন	10	90.74 %
ত্ত	10	90.18 %
ক্ষ	10	91.10 %

We have worked on 50 conjunct characters with 10 sample of every different characters. And the average training percentage is 98.73 %. And the average testing percentage of the characters are = 90.10 %. A Graphical User Interface is created to confirm the recognition. And for every character a fixed string has been set. Such as, 'SB' has been set for the character 'স্ব'.

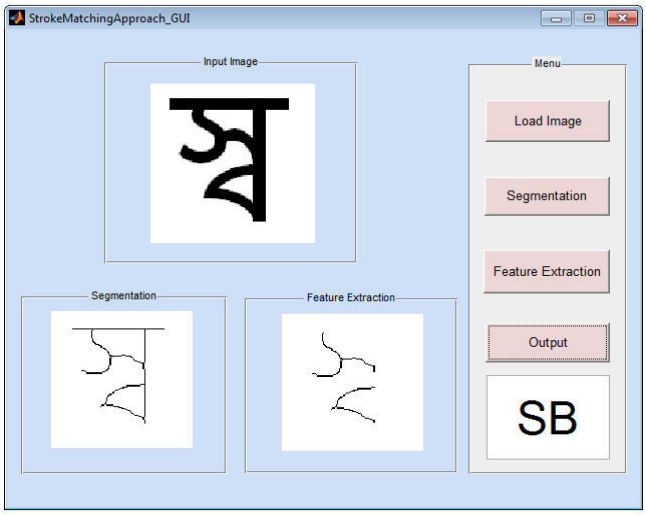


Fig. 5. A GUI to recognize the character.

D. Discussion

Handwritten character comes along with noises. So, stroke matching method at first removes the noise for the easy of character recognition. Then it collects the value points and slopes. It matches with the character stored in the dataset. To complete the recognition Pattern Recognition tool of Neural network is used.

VI. CONCLUSION

The proposed Stroke Matching Based approach is effective for Bangla Offline Character Recognition successfully. Stroke generation and matching completes the process without the loss of information. However, the stroke matching character recognition has some limitations. And after the overcome of maximum limitations the recognition rate stands up to 92.92%.A lot more improvements can be done by curve detection, proper character definition and to recognize all BOC.

REFERENCES

[1] Bhattacharya, Nilanjana, and Umapada Pal. "Stroke segmentation and recognition from Bangla online handwritten text." *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*. IEEE, 2012.

[2] Barman, Sumana, Amit Kumar Samanta, and Tai-hoon Kim. "Design of a view based approach for Bengali Character recognition." *Int. J. Advanced Science and Technology*. 2010.pp-49-62

[3] Mondal, T., et al. "On-line handwriting recognition of Indian scripts-the first benchmark." *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*. IEEE, 2010.

[4] Nasir, Mostofa Kamal, and Mohammad Shorif Uddin. "Hand Written Bangla Numerals Recognition for Automated Postal System." *IOSR Journal of Computer Engineering* 8.6 (2013): 43-48.

[5] Asaduzzaman, A. O. M., Md Khademul Islam Molla, and M. Ganjer Ali. "Printed bangla text recognition using artificial neural network with heuristic method." *Proceedings of International Conference on Computer and Information Technology*. 2002.

[6] Roy, Partha Pratim, et al. "A Novel Approach of Bangla Handwritten Text Recognition Using HMM." *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014.

[7] Badsha, Md Alamgir, et al. "Handwritten Bangla Character Recognition Using Neural Network." *International Journal of Advanced Research in Computer Science and Software Engineering* 2.11 (2012): 307-312.

[8] Zibran, Minhaz Fahim, et al. "Computer Representation of Bangla Characters And Sorting of Bangla Words." *Proc. ICCIT*. 2002.

[9] Mashiyat, Ahmed Shah, Ahmed Shah Mehadi, and Kamrul Hasan Talukder. "Bangla off-line handwritten character recognition using superimposed matrices." *Proc. 7th International Conf. on Computer and Information Technology*. 2004.

[10] Hasan, MA Mehedi, M. A. Alim, and M. Wahedul Islam. "A New Approach to Bangla Text Extraction and Recognition from Textual Image." *8th International Conference on Computer and Information Technology, ICCIT*. 2005.

[11] Hossain, SK Alamgir, and Tamanna Tabassum. "Neural net based complete character recognition scheme for Bangla printed text books." *Computer and Information Technology (ICCIT), 2013 16th International Conference on*. IEEE, 2014.

[12] Rabbani, Masud, et al. "A new stroke matching based approach to recognize Bangla handwritten text." *Computer and Information Technology (ICCIT), 2015 18th International Conference on*. IEEE, 2015.