IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-1 E-ISSN: 2347-2693

An Innovative Approach of Dehooking for Online Handwritten Bengali Characters and Words

Gouranga Mandal

Department of Computer Science and Engineering, FST, The ICFAI University, Tripura, India

[*] *Corresponding Author: gourangamandal@yahoo.com*

Available online at: www.ijcseonline.org

Received: 02/Jan/2018, Revised: 10/Jan/2018, Accepted: 23/Jan/2018, Published: 31/Jan/2018

***Abstract*—** For the last few decades several researches have been conducted on Online handwriting analysis. But scholars have unanimously agreed to the fact that it is challenging research area. To recognize with perfect prediction some pre-processing steps are essential. In this paper an honest endeavor is made to present dehooking as one of the important pre-processing steps. Here Bengali online handwritten Characters and words are considered as samples for removing hooks. Hooks are basically common artifacts used by people during fast writing. Hooks are very common issues present at the beginning in very rare case and the end of character stroke in maximum case and are generated by the pen-down and pen up movements respectively. Dehooking is the process of eliminating such unwanted strokes that appear due to inaccuracies in pen down position. Dehooking algorithms are applied to remove hooks. Here, strokes are detected by comparing the number of points with a threshold value. If the value is greater than the threshold value, the mark is retained or it is removed otherwise. In this new and innovative approach we focus on the dehooking at the end of character stroke and consider last 20 percent of each stroke for checking, according to distance from the co-ordinate of the first pixel. In last 20 percent of a stroke, we calculated angle among three consecutive pixels. If in a particular point, angle among three consecutive pixels is falling suddenly then immediately we pointed out that point. After pointing out the angle falling place we checked the entire remaining pixel after that point, whether all the remaining points are getting fade slowly or not. If it is found that all the remaining points are getting faded slowly then it can be assumed that it is a hook. After detecting the hook of a particular stroke we remove all the remaining pixels from the falling angle place so that hook can be removed and the handwritten character remains hook less. I have tested 4000 Bengali online handwritten characters and have got 97.02 percent of accuracy.

***Keywords*—**Online, Handwriting, Character, Angle, Fade, Hook

I. INTRODUCTION

Online Handwriting recognition is a procedure of a computer to detect and understand intelligible handwritten characters, words, sentence or paragraph input from a touch sensitive or pen sensitive input sources such as Pen tablets, PDA, touch-screens or other devices. The movements of the pen tip may be sensed "on line", but it is comparatively difficult task to recognize with great accuracy because in case of online handwriting only co-ordinate values are known to us. Handwriting recognition principally entails optical character recognition. A complete handwriting recognition system contains many pre-processing steps, formatting, performs correct segmentation into characters, normalization and finds the most plausible Characters and words. By On-line handwriting recognition system, any handwritten text can be detected and converted to any high level natural language, where as a sensor picks up the pen-tip movements as well as pen-up/pen-down switching and covert into vectors or matrix form. This kind of data is known as digital ink and can be considered as a digital representation of handwriting. The obtained signal is converted into letter codes which are usable within computer and text-processing applications.

For online word recognition there are some preprocessing steps. One of the most important pre-processing steps is dehooking as because without dehooking of a handwritten stroke actual recognition is not possible. Only few works have been done on dehooking of online handwritten characters and words. Here a new algorithm is proposed for dehooking of online handwritten characters and words. This algorithm can be applied for any Indian as well as foreign languages. Here we tested the algorithm on Bengali handwritten characters and words. Rest of the paper is organized as follows-
Section II contains Bengali script and online data collection,
Section III deals with the related work of dehooking process,
Section IV explains the proposed methodology, Section V

describes results and discussion, Section VI concludes the research work with future directions.

II. Bengali Script AND ONLINE DATA COLLECTION

The Bengali alphabet or Bengali script [1] is the writing system for the Bengali language and, together with the Assamese alphabet, is the fifth most widely used writing system in the world. The script is used for other languages like Meithei and Bishnupriya Manipuri, and is also used to write Sanskrit within Bengal. Besides, Bengali is the national language of Bangladesh. From a classificatory point of view, the Bengali script is an abugida, i.e. its vowel graphemes are mainly realized not as independent letters, but as diacritics attached to its consonant letters. It is written from left to right and lacks distinct letter cases. It is recognizable, as are other Brahmic scripts, by a distinctive horizontal line running along the tops of the letters that links them together which is known as matra. From a statistical analysis we notice that the probability that a Bengali word will have horizontal line is 0.994.The Bengali script is however less blocky and presents a more sinuous shape.

The alphabet of the modern Bengali script consists of 11 vowels and 40 consonants. These characters are called as basic characters. In Bengali 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 is called 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 Bengali. There are about 280 compound characters in Bengali. In this work the recognition of Bengali basic characters are considered.

The online data collection involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements X (t), Y (t) as well as pen-up/pen-down switching. That kind of data is known as digital ink and can be regarded as a dynamic representation of handwriting (see Figure 1). The ink signal is captured by either: A paper-based capture device a digital pen on patterned paper a pen-sensitive surface such as a touch screen the information on strokes and trajectories are mathematically represented in an ink signal composed of a sequence of 2D points ordered by time. No matter what the handwriting surface may be, the digital ink is always plotted according to a matrix with x axis and y axis and a point of origin. Online data acquisition captures just the information needed, which is trajectory and strokes, to obtain a clear signal. This effective information makes the data easier to process.

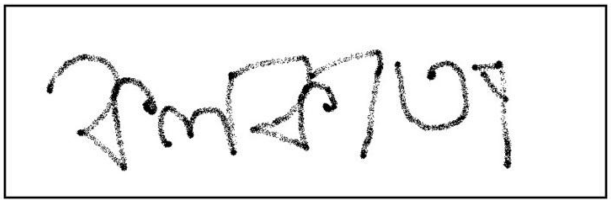

Figure 1. Example of a online bengali handwritten word

III. Related Work

Hooks (see Figure 2) are very common artifacts found at the ends of the strokes. They are generated during fast writing, when pen-down and pen-up events are generated with a delay, such that the events do not match with the real touch and lifting of the stylus. Hooks affect the efficiency of process of recognition. Hooks are mostly found at the end of the input but sometimes can be seen at the beginning of the writing as well. The way hooks are detected in strokes consists of locating abrupt changes of the turning angle.

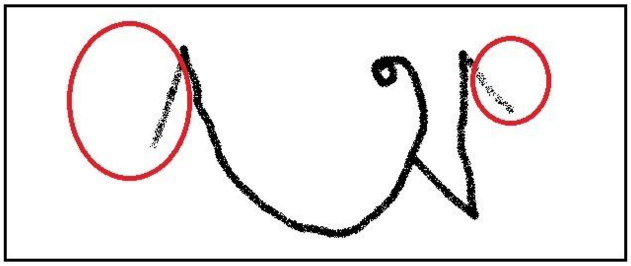

Figure 2. A tripical example of hook in bengali character

Hooks could occur due to the excessive writing speed, style or inexperience in writing. If Hooks are not removed, it will be difficult to detect original ligature. Lesser are the unwanted parts, greater is the recognition rate.

A. Writing Direction, and Zoning Information

This approach was applied on Roman Handwritten and Numeral script. This approach [2] is based on Structural features-the change of writing direction and zoning information to create a single global feature vector/ Neural Network. By this approach the highest test result was 86.63% for digit by using 40 hidden units.

B. De-hooking and Preservation of shape features

This approach was applied on Urdu language which has lots of different variations in writing. It has loops, sharp edges, cusps and curves etc. In order to preserve edges, cusp points, loops present in stroke and to retain the original shape, scholars working in the field have uphold a new algorithm. They test angle between each pair of co-ordinates, if the angle is 25 degree less than the last angle then that will be considered as hook point. As a result they observe that in addition with removing extra parts by this algorithm is also preserving its shape, while Cusp points remain same. [3]

C. Dehooking based on a threshold value

Hooks occur at the beginning and the end of character stroke and are generated by the pen-down and pen up movements. Dehooking is the process of eliminating such unwanted strokes that appear due to inaccuracies in pen down position. Dehooking algorithms are applied to remove hooks. Here, strokes are detected by comparing the number of points with a threshold value. If the value is greater than the threshold value, the mark is retained or it is removed otherwise. [4]

D. Dehooking based generated chain codes

In this approach (applied for Urdu script) the hooks occurring at the beginning and the end of the stroke are removed with the help of the generated chain codes. If variation in the chain code (last 6 chain code in length) at the beginning or end is less than the specified threshold, then that part is considered as a hook and is removed either by discarding it or by replacing the respective co-ordinates with the neighbouring ones. Hooks are generated by users either he is experienced or inexperienced. As there is a very small lines that is added at start and end but some problems exist in removing these hooks, it may be possible that small up of ‘jeem’ is removed instead of hooks which is the most important part in the detection of ‘jeem’. So to avoid de-hooking in ‘jeem’ de-hooking at beginning is not performed on those ligatures which are written from left to right for some length like. The isolated ‘jeem’ is written from left to right and the ligature is also written left to right at beginning while the remaining ligature is written from right to left. [5]

IV. METHODOLOGY

The algorithm we have implemented is very interesting one and it is based on checking angle inside 1st 20% and last 20% of each stroke where the possibility of hooks are very high. If angle meets a certain criteria then we have checked whether average distances of pixels are fading slowly or not. The algorithm for dehooking is as follows:

Algorithm: *Dehooking*

step 1: *Consider the whole word and split each stroke based on third variable ‘z’, i.e. - z=0*

step 2: *Consider a single stroke and calculate the distance of all the pixels from the stating pixel of stoke, i.e. the distance of (x2, y2) from (x1, y1) is*

$$distance \;=\; \sqrt{(x1-x2)^2+(y1-y2)^2}$$

step 3: *Calculate the total_distance of all the pixels. I.e. total_distance = total_distance+distance (Where total_distance is initialize to 0, and when a new stoke starts the total_distance again initialize to 0)*

step 4: *Calculate 1st 20% and last 20% of total_distance.*

step 5: *Now calculate all the angles among the three adjacent pixels which are inside 1st 20% and last 20%. i.e.:*

$$\theta = cos^{-1}\left|\frac{\{(x2-x1)\times(x3-x2)\}+\{y2-y1)\times(y3-y2)\}}{\sqrt{(x2-x1)^2+(y2-y1)^2}\times\sqrt{(x3-x2)^2+(y3-y2)^2}}\right|$$

(Where θ is the angle, (x1, y1), (x2, y2) and (x3, y3) are the coordinate of three adjacent pixels (see Figure 3) *of handwritten word)*

step 6: *Check if there any value of angle (θ) is less than 40 degree or not inside 1st 20% and last 20%, if found then note that point as hook point1 (for 1st 20%) and hook point2 (for last 20%).*

step 7: *Now check all the pixels before hook point1 and after hook point2 if pixel’s distances are fading slowly or not. i.e.: split a hook in two part as per length and count average distance of two parts and check whether average distance will increase slowly or not.*

step 8: *If pixels found fading then remove all the pixels before hook pont1 and after hook point2.*

step 9: *Repeat the process for each stroke.*

step 10: *Whole word will be hook less.*

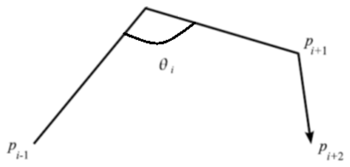

Figure 3. Angle among three adjacent pixel

V. RESULTS AND DISCUSSION

The experimental evaluation of the above algorithm is carried out using online handwritten words. The data is collected from the people with different backgrounds. Total of 4,000 Bengali handwritten words are collected as samples for the experiment. Out of them 41.2% of the words are used for the training of the classifier for the present work and rest is used for the testing purpose. 2352 Bengali words with hook have been tested (see Figure 3) in our system and around 97.02% accuracy is obtained *(e.g., Table 1)*. The dehooking accuracy obtained from the classifier is shown in Table:

Table 1: *Result of Dehooking Algorithm*

Total Words	Words With Hook	Hook Removed	Hook Not Removed	Dehooking Accuracy In %
4000	2352	2282	70	97.02%

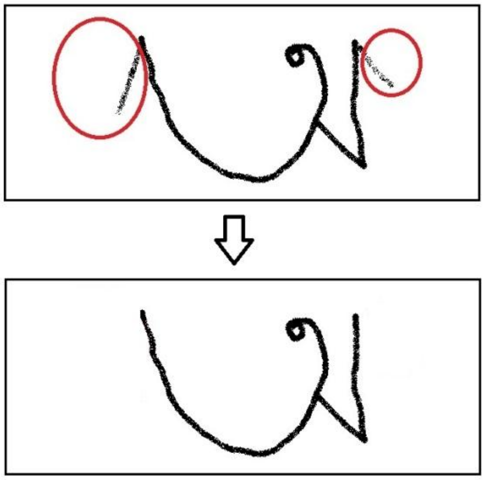

Figure 4. Bengali Character before dehooking and after dehooking

VI. CONCLUSION AND FUTURE SCOPE

This paper tends to present an innovative technique for detecting the hook and removing the hook in Bengali handwritten characters and words. By using the stated technique different online handwritten characters and words can be made hook less. If dehooking of every stroke can be done successfully then characters segmentation will be apt and the recognition will be more perfect. This result will be helpful for recognizing Bengali as well as words of other Indian languages.

We tested the proposed system on 4000 data out of them 2352 are words with hook and got the encouraging result. Not much work has been done towards the online recognition of Indian scripts in general and Bengali in particular. So this work will be helpful for the research towards online recognition of other Indian scripts as well as for Bengali in the level of word, text and so on. This technique in turn will be very much effective in the process of conversion of handwritten Bengali or other Indian script into the conventional system recognized fonts, which has a very wide range of application, especially in the Indian service sectors. In fact the work for online recognition of Bengali handwritten word is going on by us and hopes that work can be completed successfully by taking the help of the current proposed work.

REFERENCES

[1] Mazumdar, Bijaychandra,"*The history of the Bengali language*" (Repr. [d. Ausg.] Calcutta, 1920. ed.). New Delhi: Asian Educational Services. p. **57**. ISBN **8120614526**, **2000**.

[2] Aini Najwa Azmi, Dewi Nasien, Siti Mariyam Shamsuddin,"*A review on handwritten character and numeral recognition for Roman, Arabic, Chinese and Indian scripts*", International Journal of advanced studies in Computer Science and Engineering, Vol **2** issue **4, 2013**

[3] Fareeha Anwar, Muhammad Adnan Aftab, Dr.Syed Afaq Hussain, Dr.Ayyaz Hussain "*Preprocessing of Online Urdu Handwriting for Mobile Devices*", International Journal of Computer Science and Network Security, Vol.**17** No.**10**, October **2017**

[4] Anitha Mary M.O. Chacko, Dhanya P.M.,"*Handwritten Character Recognition in Malayalam Scripts– a Review*", International Journal of Artificial Intelligence & Applications (IJAIA), Vol. **5**, No. **1**, January **2014**

[5] Muhammad Imran Razzak, Syed Afaq Hussain, Muhammad Sher, Zeeshan Shafi Khan, "Combining Offline and Online Preprocessing for Online Urdu Character Recognition", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009, Vol **I** IMECS **2009**, March **18 - 20, 2009**, Hong Kong

Authors Profile

Mr. Gouranga Mandal pursed Bachelor of Technology in Information Technology from West Bengal university of Technology, Kolkata in 2009 and Master of Technology in Computer Science & Engineering from West Bengal university of Technology, Kolkata in year 2012. He is currently working as Assistant Professor in Computer Science & Engineering Department, Faculty of Science & Technology, The ICFAI University Tripura since 2017. He has published many research papers in reputed international journals. His main research work focuses on Online Document Image Processing, Optical Character Recognition, Natural Language Processing and analysis in Bengali and other Indian Script and Online Handwritten Document Recognition. He has 6 years of teaching experience.