ONLINE BANGLA HANDWRITTEN COMPOUND WORD RECOGNITION BASED ON SEGMENTATION

Sumanta Daw

Department of CSE, Hooghly Engineering & Technology College, Hooghly
sumanta.daw@hetc.ac.in

ABSTRACT

In this paper I propose a scheme for "Online Bangla Handwritten Compound Word Recognition" based on segmentation of word into its constituent characters with more accuracy. The goal of this Paper is to develop a system for segmentation of Bengali Compound Word into its constituent characters or basic strokes and then to recognize each character individually based on stroke generation, thus the recognizer can recognize the entire word. I achieved the correct segmentation rate of 87% and the overall recognition rate of 73% on a dataset of 4200 Bangla Compound Words.

KEYWORDS

Compound Word, Segmentation, Under Segmentation, Over Segmentation.

1. INTRODUCTION

Online handwriting recognition provides a dynamic means of communication with computers through a pen like stylus, as it is natural writing instrument and this seems to be an easier way of entering data into computers. However, wide variation of human writing style makes online handwriting recognition a challenging pattern recognition problem.

In this work, major part is the segmentation of a word into its component characters or valid basic strokes [9]. So, this phase should be proper otherwise determining the combination of strokes to determine the boundary of a particular character in a word will be ambiguous. I tried to solve the problem by the various modules such as Online Data collection, Preprocessing or Stroke Extraction, Segmentation of Online Handwritten Compound words into basic stroke, Features generation of Compound stroke, Basic Compound stroke training to the classifier, Recognition of individual basic stroke [10]. The recognition result for the previous work [11] was just around 43% due to some under and over segmentation problems; however the segmentation rate was nearly 83%. In this paper the modified segmentation algorithm enhanced the segmentation rate up to 87% and mostly overcome the under and over segmentation problems for some characters by which the recognition rate also increased.

1.1 Brief comparison between offline and online approaches:

- Online recognition system: the system accepts the movement of pen from the hardware such as graphic tablet, wacom tablet, light pen, A4 takes note; and there is a lot of information during the

Rupak Bhattacharyya et al. (Eds) : ACER 2013,
pp. 69–76, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3207

input process available such as: current position, movement's direction, stopping points, starting points, strokes order.

- Offline recognition system: the system accepts image as input from scanner, offline recognition is more difficult than online recognition: because of not availability of contextual information and prior knowledge like text position, size of text, order of strokes, stop points, and start points. Furthermore there are noises in image while the noises in online recognition near to be absent.

2. DATA COLLECTION

On-line handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or A4 take not**e** where a sensor picks up the pen-tip movements X (t), Y (t) as well as pen-up/pen-down P either with 0 or 1 switching. That kind of data is known as digital ink and can be regarded as a dynamic representation of handwriting. The ink signal is captured by either:

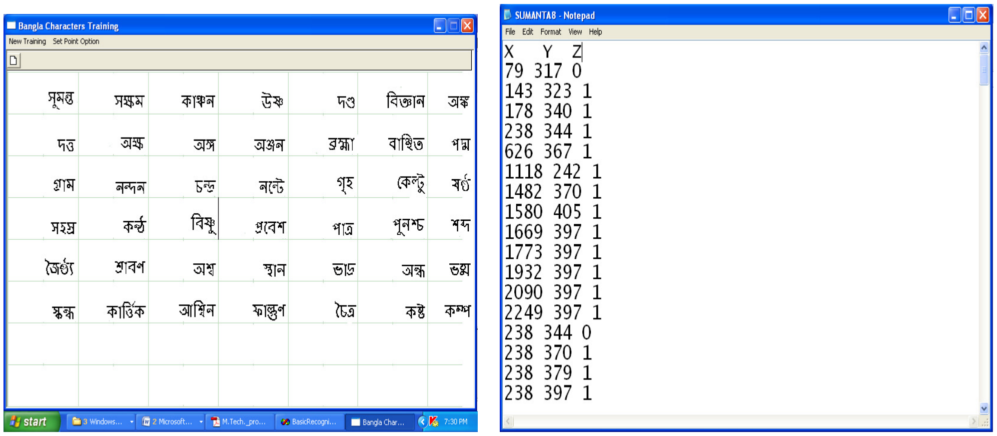
A paper based capture device,
A digital pen on patterned paper,
A pen-sensitive touch screen,

To collect the data (Word) I used A4 take note or the datasheets. Here we used datasheets.
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 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* € R 2 X{0,1}*; i* = 1*:M*, where *pi* is the pen position having x-coordinate and y-coordinate and M is the total number of sample points. Let (*pi*) and (*pj*) be two consecutive pen points. We retain both of these two consecutive pen points (*pi*) and (*pj*) if the following condition is satisfied:

$$x2 + y2 > m2 \text{ .... (i)}$$

where $x = xi - xj$ and $y = yi - ¡yj$. The parameter m is empirically chosen. I have set m equal to zero in Equation (i) to removes all consecutive repeated points.

Analyzing a total of 4200 Bangla compound words we found that, for writing Bangla characters, the number of sample points (M) varies from 14 (for the character ত) to 176 (for the character ষ্ফ) points. The average number of sample points in a Bangla character is 72. I also computed the average number of sample points in each character class. I noted that the character class (ঞ্জ) has the maximum number of sample points and its average value is 113. The character class (চ) has the minimum number (46) of sample points. Figure 1 shows the online collected data in form of text and the datasheet of 42 compound words.



সুমন্ত	সক্ষম	কাঞ্চন	উষ্ণ	দণ্ড	বিজ্ঞান	অঙ্ক
দত্ত	অক্ষ	অঙ্গ	অঞ্জন	ব্রহ্মা	বাঞ্ছিত	পদ্ম
গ্রাম	নন্দন	চন্দ্র	নন্টে	গৃহ	কেল্টু	ষষ্ঠ
সহস্র	কন্ঠ	বিষ্ণু	প্রবেশ	পাত্র	পুনশ্চ	শব্দ
জৈষ্ঠ্য	শ্রাবণ	অশ্ব	স্থান	ভাদ্র	অন্ধ	ভস্ম
স্কন্ধ	কার্ত্তিক	আশ্বিন	ফাল্গুন	চৈত্র	কষ্ট	কম্প

```
X    Y   Z
79  317  0
143  323  1
178  340  1
238  344  1
626  367  1
1118  242  1
1482  370  1
1580  405  1
1669  397  1
1773  397  1
1932  397  1
2090  397  1
2249  397  1
238  344  0
238  370  1
238  379  1
238  397  1
```

Figure 1: Datasheet for Collection Data and Text format of collected Data

3. Stroke Extraction

By stroke we mean the set of points obtained between a pen down and pen up. In other words the number of sample points collected by a continuous writing of the pen without lifting it. Main difficulty of Bangla character recognition is shape similarity, stroke size and the order variation of different strokes. From the statistical analysis on our dataset we found that the minimum and maximum number of stroke used to write a Bangla compound character is 1 and 6. Bangla compound characters also may be written by using all of these basic strokes. So in Bangla language apart from the simple 66 strokes with compound characters there are mainly 72 strokes available. All of these strokes also written by the combination of basic 66 strokes. Although in case of combination we consider that 66 + 72 basic stroke in Bangla, so a total of 138 basic strokes. The list of compound basic strokes is in Figure 2:

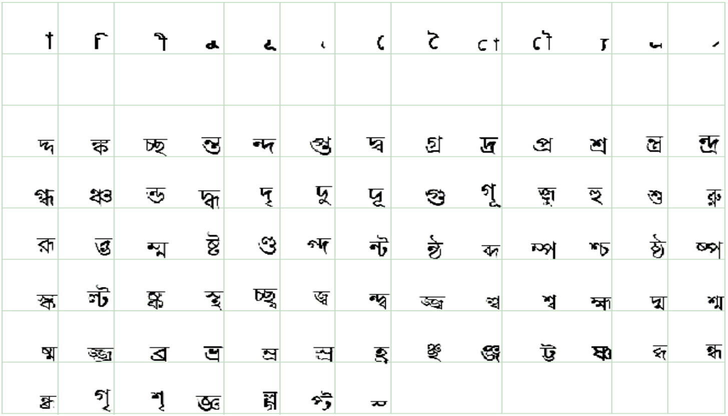


Figure 2: Compound basic Stroke

4. Compound Word Segmentation

There are about 280 compound characters in Bangla. Main difficulty of Bangla character recognition is shape similarity, stroke size and the order variation of different strokes. I know that in Bengali handwriting the movement of each stroke is generally downside. By keeping this concept in mind it has been seen that in a downside movement stroke the point from where that downside movement starts [10, 11] at that point I have to split that stroke. This should be done only in the upper zone i.e. first 33% portion of the total height of the image. In the remaining 67% of the image segmentation is not needed. But the compound characters mostly prepared by using the two different simple characters. By considering that feature of compound characters, these characters also may be segmented from middle portion also (i.e. 50% of total height). People write any word in Bangla, such a manner where more than one alphabet is joined with one another. This joining is generally found in the upper 1/3rd. portion of the character (exception in few cases) [11]. The modified segmentation algorithm is as follows:

Step 1: Store each pixel of the online data in three variables corresponds to X and Y coordinates and pen feature value of 0 or 1 in third variables for identifying strokes.

Step 2: For each third variable value 0 separates each strokes scanning pixels of the word. Calculate 30% of the height for a simple and 50% of the height for a compound character.

Step 3: Select at which point of stroke segmentation is needed based on the previous output. We have to finally segment those points of same or different strokes which required to be segmented. So, we use one function to check at which pixel it is feasible to segment a stroke. We have to check few features of Bangla characters for this process such as:

i) Each pixel's distance from the start and end of the stroke,
ii) The width of the stroke up to the pixel in question from the start and end of the stroke,
iii) The height of the stroke up to the pixel in question,
iv) Total stroke distance,
v) Total width of the word. After finding these features we have to take some ratio of
(a) Each pixel's distance & Total stroke distance,
(b) The width of the stroke up to the pixel in question & Total width of the word and thus to decide at which pixel of a particular stroke segmentation is feasible.

Step 4: Now if at a particular pixel it is feasible to segment the stroke, then first we check whether that pixel's *y* co-ordinate value is 30% of the *height* or not. If it is not then there will be no segmentation. If it is, then we check whether at that pixel *downside* movement of the stroke starts or not. For this checking I am taking two points p_{i-1} and p_{i-2} before the point in question and similarly two points p_{i+1} and p_{i+2} after that point. If the y-coordinate of p_{i-1} is <= p_{i-2} and p_i <= p_{i-1} and simultaneously if the y-coordinate of p_{i+1} >= p_i and p_{i+2} >= p_{i+1} (i.e. *downside movement* of stroke) then only at p_i stroke is splitted. If at a particular point stroke is splitted then I skip next 9 or 10 pixels for checking of feasibility of segmentation.

Step 5: Repeat step 3 and 4 for each pixels and each strokes of the entire word.

By this approach I tried to segment all the compound words covering all the vowels and consonants modifiers and also covering all the alphabets in Bangla language and the result is in Figure 3.

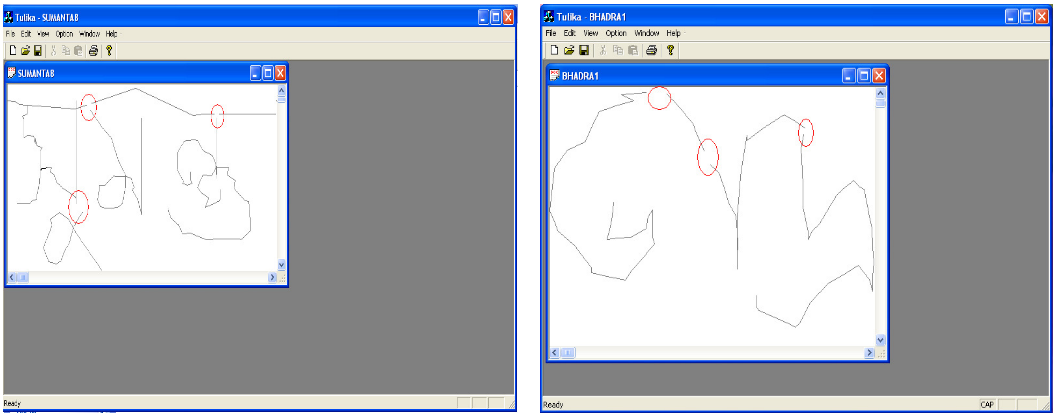

Tulika - SUMANTA8
File Edit View Option Window Help
SUMANTA8
Ready
Tulika - BHADRA1
File Edit View Option Window Help
BHADRA1
Ready
CAP

Figure 3: Result after Segmentation

5. Feature Generation

Any online feature is very much sensitive to writing stroke sequence and size variation. Total 233 features (90+15+128) are used [9].

The features used are:

- *Point based feature(90),*
- *Structural features (15),*
- *Directional feature (128),*

The processed character is transformed into a sequence $t = [t_1 \ldots t_N \; t_{N+1} \ldots. \; t_{N+15} \; t_{N+15+1} \ldots..t_{N+15+128}]$ of feature vectors $t_i = (t_{i1}, \; t_{i2}, \; t_{i3})^T$ (Where I <= N).

6. Training To The Classifier

In this step the extracted features are to be fed to the classifier by the training to the classifier using the concept of Neural Network. Based on the above-normalized features, a Multilayer Perception Neural Network based scheme was used for recognition of the strokes. The Multi Layer Perception Network (MLP) is, in general, a layered feed-forward network, pictorially represented with a directed acyclic graph. Each node in the graph stands for an artificial neuron of the MLP, and the labels in each directed arc denote the strength of synaptic connection between two neurons and the direction of the signal flow in the MLP [8]. For pattern classification, the number of neurons in the input layer of an MLP is determined by the number of features selected for representing the relevant patterns in the feature space and output layer by the number of classes in which the input data belongs. The neurons in hidden and output layers compute the sigmoid function on the sum of the products of input values and weight values of the corresponding connections to each neuron.

In this work the number of neurons in input and output layers of the perception is set to 278 and 138; respectively since the number features are 278 and the number of possible classes in hand written stroke considered for the present case is 138. The number of hidden units was set to 90, back propagation learning rate and acceleration factor is set to suitable values, based on trial runs.

7. WORD RECOGNITION

Each word will be constructed with the help of its recognized strokes. I have taken 42 different Bengali words covering all the vowels & consonants in Bengali and all the modifiers using which a Bangla word is written. Those words are collected online using A4 take note. Now to construct each of those 42 words we have to send to the recognition module the correct combination of basic strokes (which are obtained from segmentation) to recognize each character as each character will be constructed with the help of its recognized strokes. So, if we can recognize each and every character in a word as well as modifiers (if exists) individually, then the entire word will be recognized. To do so, all the probable sequences of strokes are stored in a tree data structure that makes a valid character into a database. To build this a database report is generated from the raw data (characters), from which the sequences of strokes of the characters are gotten, that are generally used by people.

8. RESULT AND DISCUSSION

The experimental evaluation of the above techniques was carried out on isolated Bangla words. The data was collected from people of different background. A total of 42 different words are collected for the experiment covering all vowels & consonants as well as modifiers of bangle script. Each word's 100 instances are taken from people of different background. So, a total of 4200 words have been collected still now and worked.

Data	Segmentation Rate	Segmentation Error
Compound Word	Approx 87%	Approx 13%

Table 1: Segmentation Result

Data	Under & Over Segmentation	Error
Compound Word	Approx 9%	4%

Table 2: Under and Over Segmentation Result

From the experiment it was found that the overall accuracy of the proposed scheme considering segmentation is satisfactory. I have checked that accuracy of segmentation is around 87% by applying the modified segmentation algorithm shown in Table 1. Over and under segmentation problem is reduced by using the new segmentation algorithm shown in Table 2.

I have checked that the classification accuracy of compound word is near around 73% by applying my modified segmentation algorithm. Rejection rate of data is 24%. The system is unable to classify the data nearly about 3%. Classification result is in Table 3.

Total Data	Classified Data	Rejected Data	Misclassified Data
4200	3066	1008	126

Table 3: Classification Result

Figure 4 and Figure 5 shows some under and over segmented data result for some compound words which is unable to segment properly by the algorithm.

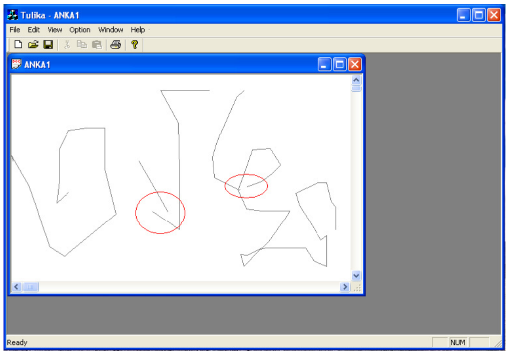


Figure 4: Over Segmented

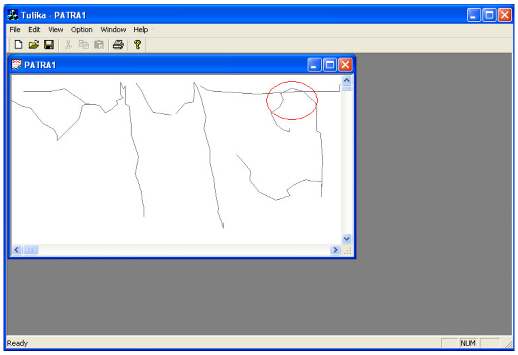


Figure 5: Under Segmented

Table 4 shows the under and over segmented characters suffers from the segmentation algorithm.

Over Segmented		Under Segmented	
Character	Word	Character	Word
ঙ্ক	অঙ্ক	ত্ত	দত্ত
হ্ম	ব্রহ্মা	ত্র	পাত্র
ঙ্গ	বাঙ্গা		

Table 4: Over and Under Segmented Characters

9. Conclusion

This work describes a new approach of Bangla Compound Handwritten Word Recognition. The work is done on 42 different predetermined Bangla compound words based on different vowels with a modified segmentation algorithm specially for Bangla Compound Word from various writing style which made the system more dynamic with a higher recognition rate than the previous. The system now can work with any Bangla basic and compound words; so it is dynamic in nature. The problem of over and under segmentation error decreased but still it suffers from few characters (such as 'ঙ্গ', 'হ্ম', 'ঙ্ক') because of their complex physical structure.

Not much work has been done towards the online compound word recognition of Indian scripts in general and Bangla in particular. I think this work can be helpful for Bangla signature verification or Paragraph Verification. So, there are many scope remains in this field based on this proposed work.

Acknowledgements

I like to thankfully acknowledge my god, colleagues and family members for their all-round of support towards the present work.

REFERENCES

[1] C. C. Tappert. C. Y. Suen, and T. Wakahara. "The State of the Art in On-Line Handwriting Recognition," IEEE PAMI. Vol. 12, No. 8, pp. 179-190, 1990.

[2] E. J. Bellagarda, J. R. Bellagarda, D. Nahamoo and N. S. Nathan, "A probabilistic Framework for Online Handwriting Recognition," 3rd IWFHR, pp. 225-234, 1993.

[3] B.B. Chaudhuri and U. Pal. "An OCR System to Read Two l" Indian Language Scripts: Bangla and Devnagari." Proc. 4th ICDAR, pp. 1011-1015, 1997.

[4] I. Guyon, M. Schenkel, and J. Denker, "Overview and Synthesis of On-Line Cursive Handwriting Recognition Techniques", Handbook of Character Recognition and Document Image Analysis, pp. 183-225, 1997.

[5] B.B. Chaudhuri and U. Pal, "A Complete Printed Bangla System," PR, Vol. 31, pp. 531-549, 1998.

[6] R. Plamondon and S. N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey". IEEE PAMI, vol. 22, No. 1, pp. 63-84, 2000.

[7] C. Bahlmann and H. Burkhardt, “The Writer Independent Online Handwriting Recognition System frog on hand and Cluster Generative Statistical Dynamic Time Warping”, IEEE PAMI, Vol. 26, No. 3, pp. 1-12, 2004.

[8] K. Roy, C. Chaudhuri, M. Kundu, M. Nasipuri and D. K. Basu, "Comparison of the Multilayer perceptron and the nearest neighbor classifier for handwritten digit Recognition", JISE, 21, 1245 (2005).

[9] K. Roy, N. Sharma, T. Pal, U. Pal “Online Bangla Handwritten Recognition System”- October 13, 2006, WSPC.

[10] Rajib Ghosh, Debnath Bhattacharyya, Samir Kumar Bandyopadhyay, “Segmentation of Online Bangla Handwritten Word”, 2009 IEEE International Advance Computing Conference (IACC 2009), Patiala, India.

[11] Sumanta Daw and Rajib Ghosh, “Online Bangla Handwritten Compound Word Recognition”, International Conference for Computing and System – 2010, November 19-20, 2010, Burdwan University, pp. 221-226.

AUTHOR

Mr. Sumanta Daw is an Assistant Professor in the Department of CSE at Hooghly Engineering & Technology College in Hooghly. He was born on 20th April, 1979. He obtains his M.Sc. in Software Engineering in the year of 2003 and MTech in CSE in the year of 2010. He worked in industry for one year and as an academician for last 9 years. He has already published several national and international papers, including one journal paper in the fields of Pattern Recognition and Network Security.