Word & Character Segmentation for Bangla Handwriting Analysis & Recognition

C. Halder
Dept. of Computer Science
West Bengal State University, Barasat
24 Parganas (N), Kolkata 700126, India
chayan.halderz@gmail.com

K. Roy
Dept. of Computer Science
West Bengal State University, Barasat
24 Parganas (N), Kolkata 700126, India
kaushik.mrg@gmail.com

Abstract— **Segmentation of unconstrained handwritten word into different zones (upper middle and lower) and characters is more difficult than that of printed documents. This is mainly because of variability in inter-character distance, skew, slant, size and curved like handwriting. Sometimes components of two consecutive characters may be touched or overlapped and this situation complicates the segmentation task greatly. In Indian languages such touching or overlapping occurs frequently because of modified characters of upper-zone and lower-zone. In this paper we propose a simple method to segment unconstrained handwritten Bangla word. We achieved 81.41% success rate in the proposed system.**

Keywords— Segmentation, Busy zone, Upper zone, Lower zone, Middle zone.

I. INTRODUCTION

Segmentation of unconstrained handwritten word into different zones (upper middle and lower) and characters is more difficult than that of printed documents. This is mainly because of variability in size, inter-character distance, skew, slant and curved like handwriting. Sometimes components of two consecutive characters may be touched or overlapped and this situation complicates the segmentation task greatly. In Indian languages such touching or overlapping occurs frequently because of modified characters of upper-zone and lower-zone. There are works of segmentation of words & characters using roman script in other countries like USA, UK, Japan etc. like some works of S. N. Srihari et al. [1-4]. But very few works are available in the literature on handwritten Bangla script. Only few complete works of U. Pal et al., K. Roy, and R. Sarkar et al. are available in literature [5-9].

In this paper we propose a simple scheme to segment unconstrained handwritten word of Bangla script into individual characters and its parts. Here, we first compute the histogram of the words to compute the busy-zone of the words. For a printed document we can easily extract the different zones of a word from busy-zone. Here we have used a local approach to segment the word into different zones. Then the characters of those zones are being segmented using a simple procedure. There is another work available of K. Roy et al. where this kind of simple technique is used but the accuracy of the approach is just less than our proposed approach [10].

II. PROPERTIES OF BANGLA SCRIPT

Bangla is the second-most popular language in the Indian sub-continent and fifth-most popular language in the world [11]. The alphabet of the modern Bangla script consists of 11 vowels and 39 consonants [9, 11]. These characters are called basic characters. The Fig.1 exhibits basic characters of Bangla script. Writing style of the Bangla script is from left to right. Unlike Roman script the concept of upper/lower case is absent in Bangla script.

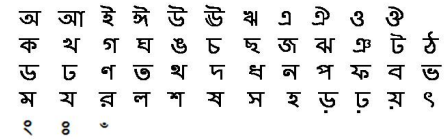


Figure 1. Basic characters of Bangla alphabet. (First 11 are vowels and rests are consonants)

We can observe, in Fig.1 that in Bangla script most of the characters have a horizontal line at the upper part. At the time of writing when two or more characters sit side by side to form a word, these horizontal lines touch each other and generate a long line called head-line (see Fig.2). In Bangla script a vowel following a consonant takes a modified shape, which, depending on the vowel, is placed at the left, right (or both) or bottom of the consonant [9, 11]. These are called modified characters. A consonant or vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be formed using two consonants, as well as consonant and vowel.

A Bangla text line can be partitioned into three zones. The upper-zone denotes the region above the head-line, the middle zone is the region between head-line and base-line, the lower-zone is the region below base-line. Different zones in a Bangla text line are shown in Fig.2.

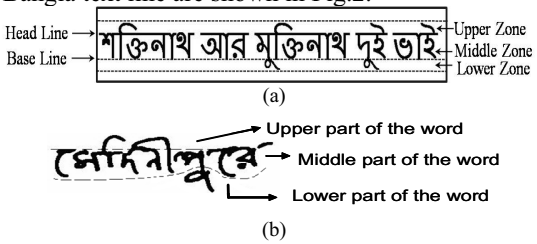


Figure 2: Different zone/part of a (a) Bangla text line and (b) word.

978-0-7695-4599-8/11 $26.00 © 2011 IEEE
DOI 10.1109/NCVPRIPG.2011.68

III. SEGMENTATION

The segmentation of words into different zones and then the segmentation of characters of those zones is necessary for further processing of those individual characters for analysis and recognition of Bangla handwritten words and characters. For this reason the proper extraction of the characters is necessary from the handwritten documents. Here after computing the horizontal histogram we analyze them to find out the busy-zone.

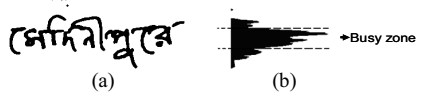
A. Busy-zone Computation Technique

Busy-zone of a word is the region of the word where a maximum portion of its characters lie. Using the following steps busy-zone can be extracted. Here we first compute the peak of the histogram and set as initial marking point. The length and the position are marked. Now from this position we traverse in up-word (down-word) till the length of the histogram is greater than half of the peak length. In this way we get the top and bottom point of the busy-zone. These positions are marked as Head line and the Base line of the word. For example see figure 3(a-b).

The steps are:

- First create histogram of the original image.
- Next identify the Head line and the Base line of the histogram.

Theoretically these three parts should be distinct. For example the upper part should contain upper part of some character like (ই ঈ ঠ ট) and some part of vowel modifiers (ি ী). The middle part should contain the vowel/consonants or its main constituent part like (ন চ ব) or part of modifiers like (া ী) and compound characters like (ঋ). The lower part should contain lower part of some consonants like (ড় র), part of some vowel modifiers (ু ূ ৃ) or lower part of compound character like (ঋ).

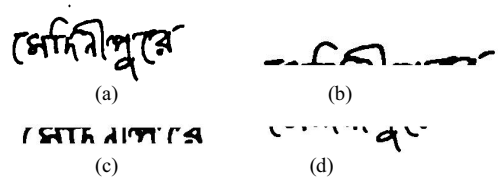


(a) (b)

Figure 3(a-b). (a) Image of a Bengali word. (b) Histogram of the word with busy zone specification.

B. Limitation of Segmentation Technique Using Busy-zone

Using the Busy-zone we can easily compute the upper part, lower part and middle part of a printed document word. But the limitation of this approach is that it will fail in case the handwritten words especially if the word suffers from skew (which may be corrected via skew correction), slant, and variable sized characters, curved or of any other irregular shapes, we would get erroneous or improper result as shown in Figure 4. In Figure 4(b) we can see some noisy characters are present in the upper part cut of the word. In Figure 4(d) we can see some noisy characters are present in the lower part cut of the word. These noisy parts are the actual parts of the middle part characters of the word. For this reason we need to traverse pixel wise, the upper part characters and lower parts characters of the word.



(a) (b)

(c) (d)

Figure 4(a-d). (a)Image of a Bengali word,(b) Upper part of the word after busy zone separation, (c) Middle part of the word after busy zone separation,(d) Lower part of the word after busy zone separation.

C. Proposed approach

In our proposed approach the segmentation of upper-zone, middle-zone, and lower-zone part of a word is considered. Here the zone segmentation is done to segment the characters of each zone properly. Busy-zone of word is used for identifying the upper, lower and middle area of any word. The segmentation procedure of the three part and the characters of any word is described below. Sec. A describes the zone wise segmentation of a word and Sec. B describes the character segmentation of the middle part characters as these are the most significant characters of any word. The segmentation of the upper part characters and lower part characters are done during the upper and lower part zone segmentation.

A. Zone Segmentation

Zone wise segmentation of a word is necessary for separation of the upper, middle, lower part characters.

1) Upper zone: Upper zone of a word is the region of the word where all the upper part characters are located. In Figure 2(b) some upper zone characters of a word are shown. For identifying the upper zone characters properly and to eliminate the noisy characters from the upper part, the head line of the busy-zone of the word is used for horizontal projection technique [9] to identify the area where the lower part of the word exists and along with that a pixel-wise traversal is used to run through the edges of the upper portion characters. The total count of the pixels of the edge of character is used with the stroke width to calculate the approximated value of total number of pixels the character contains. Then a comparison is done with the threshold value (calculated using the stroke width of the word) to get the actual upper part characters of the word.

In the following Figure 5 we can see an example of upper part segmentation done by the system. The upper segmented part of the word shown in Figure 2 is shown here in this Figure 5.

Figure 5. Example of upper zone segmentation done by the system.

2) Lower zone: Lower zone of a word is the region of the word where all the lower part characters are located. In Figure 2 lower zone character of a word is shown.

For identifying the lower zone characters properly and to eliminate the noisy characters from the lower part, the base line of the busy-zone of the word is used for horizontal projection technique [9] to identify the area where the lower part of the word exists and along with that a pixel-wise traversal technique is used to run through the edges of the

lower portion characters. The total count of the pixels contained in the edge of a character is used with the stroke width to calculate the approximated value of total number of pixels the character contains. A threshold value which is calculated using the stroke width of the word is used to compare with the approximated total pixel value of the character to get the actual lower part characters of the word. In the following Figure 6 we can see an example of lower part character of the word shown in Figure 2 after lower part segmentation.

Figure 6. Example of lower zone segmentation done by the system.

3) Middle zone: Middle zone of a word is the region of the word where all the middle part characters are located. In Figure 2 we can see some middle zone characters of a word. The middle zone of a word contains all the characters that are the base characters of the word. The upper and lower zone characters are joined with the middle zone characters to create the original word.

For identifying the middle part characters of the word properly and to remove errors, horizontal projection technique [9] is used between the area of the identified lower and upper zone of the word. After the proper segmentation of lower and upper zone of the word, removal of these parts along with the noisy characters from the original word will give us the actual middle part of the word. In the following Figure 7 we can see an example of middle part characters of the word shown in Figure 2 after lower part segmentation.

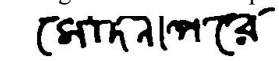


Figure 7. Example of middle zone segmentation done by the system.

B. Character segmentation

After the segmentation of the three parts of a word, character segmentation of the base characters of the word is important for correct recognition of the word. For this reason the middle part characters of the word are also segmented to get isolated characters. For segmentation of middle part characters at first, the inter space between two characters are identified, then the segmentation lines are determined using a vertical projection technique [9]. After identifying the segmentation lines the characters of the middle part are segmented properly using those segmentation lines. In the following Figure 8 we can see an example of separated middle part characters of the word that is shown in Figure 2.

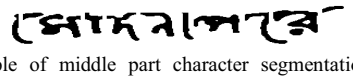


Figure 8.Example of middle part character segmentation done by the system.

IV. Results and Analysis

A. Result on Zone Segmentation

The performance of the system according to zone wise segmentation is given in the following Table 1. Here zone wise segmentation means identifying the 3 zones contained in the word and segmenting these zones properly with characters belong to that zone. It gives 90.76% accuracy. Here, other images that did not give accurate results contain some images with 2 zones that are properly segmented out of 3 zones i.e. containing 1 error. If we consider those 1 error images as good result then the accuracy become 93.76%.

Table I. Zone segmentation result.

Category	*Sample (Total Zones)*	*Properly segmented*	*% of Proper segmentation*
Without Error	2500 (7500)	2269	90.76
With 1 Error	2500 (7500)	2344	93.76

B. Result on Word Segmentation

The table 2 shows the word wise segmentation result. The result shows 54.48% accuracy. The main source of errors which is 45.52% due to words that are written in slanted fashion. Here the accuracy is low because, for any word with a little noisy segmentation of upper, lower, middle part and middle part characters are considered as improper segmentation, thus reducing the accuracy. For example the word shown in Figure 9(f) will be considered as improper segmentation in respect to word wise segmentation but not the word shown in Figure 9(b).

Table II. Word segmentation result.

Sample	*Properly segmented*	*% of Proper segmentation*
2344	1277	54.48%

C. Result on Character Segmentation

The performance of the system according to character wise distribution is given in the following Table 3.Here character wise segmentation means identifying all the characters of the word and segmenting them properly. In the table we can see that total 16957 characters are encountered and it gives results 81.41% accurately. The main source of error which is 18.59% is due to words that are written in slanted fashion and also for not proper positioning of the upper, middle, lower portion characters at the time of writing. Sometime words contain characters which are written very close to each other without any space between the characters these characters also increase errors. Here the numbers of characters accurately segmented were considered. For example the word shown in Figure 9(e) will give approximately 86% accuracy as it segmented properly 6 characters out of 7 characters, but the word shown in Figure 9(b) will give 100% accuracy. The following table also gives the result of 1 error and 2 error segmentations. Here the images that have 1 character improperly segmented are considered as images with 1 error. If we consider these images as good result then the accuracy becomes 89.73%. Likewise if we consider 2 character errors then the accuracy becomes 93.62%.

Table III. CHARACTER SEGMENTATION RESULT.

Category	*Samples (Total Characters)*	*% of correct segmentation*
Without Error	2344 (16957)	81.41%
With 1 Error	2344 (16957)	89.73%
With 2 Error	2344 (16957)	93.62%

D. Analysis

In Figure 9(a), 9(b), 9(c) we can see images that are written almost using a straight line fashion gives proper results. In Figure 9(b) we can see images which have good separation between characters gives proper results.

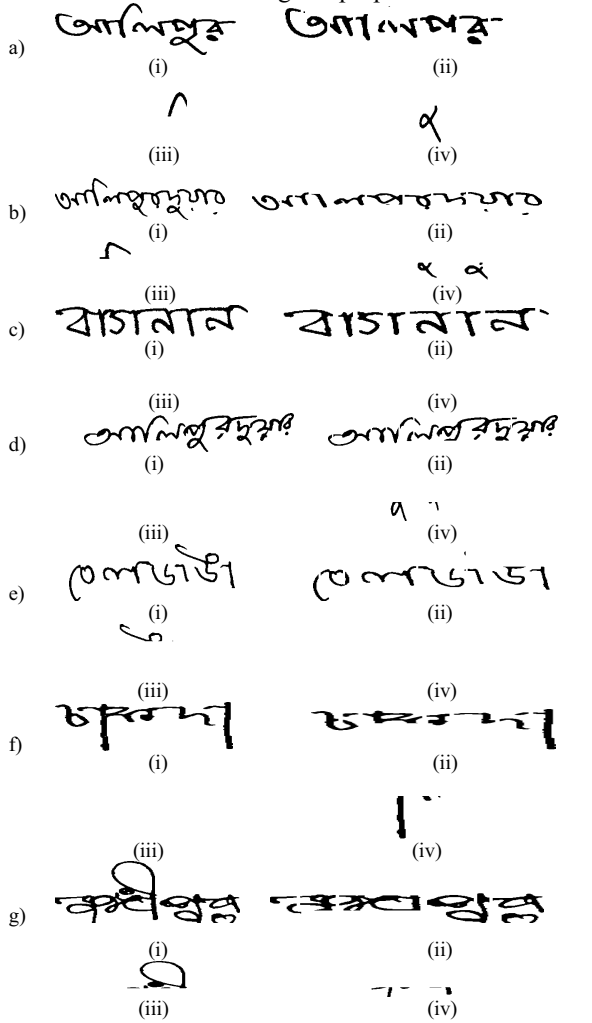


Figure 9. Some sample results of character segmentation. (i)Original image, (ii) Middle part generated by the system, (iii) Upper part generated by the system,(iv) Lower part generated by the system.

The reasons for giving improper results are as follows: Images that are slanted like one shown in Figure 9(d).Images in which middle portion characters of a word appearing at upper portion of the word like one shown in Figure 9(e).Images in which middle portion characters of a word appearing at lower portion of the word like one shown in Figure 9(f).Images in which lower and middle part characters are congested and not clearly identifiable gives improper results like one shown in Figure 9(g).

V. CONCLUSION

In this paper we have proposed a scheme for Segmentation of unconstrained handwritten words into different zones (upper middle and lower) and characters. There are many difficulties in respect to segment handwritten words of Bangla script. Most of these difficulties can be encountered due to the properties of Bangla script. In this paper we propose a simple method to segment unconstrained handwritten Bangla word. We have achieved 81.41% success rate in the proposed system. The word and character segmentation is a kind of a procedure which will be a great help in the field of Bangla handwritten word recognition. In future this segmentation technique can be helpful for developing an automated Bangla handwriting recognizer and analyser. This can also be useful in future for graphological analysis of handwritten documents. We believe this approach will help in development of unconstrained handwriting Optical Character Recognition system for Bangla.

REFERENCES

[1] S. N. Srihari, and E.J. Keubert, "Integration of Hand-Written Address Interpretation Technology into the United States Postal Service Remote Computer Reader System", In Proc. of 4th ICDAR, pp. 892-896, 1997.

[2] U. Mahadevan, and S. N. Srihari, "Parsing and Recognition of City, State, and ZIP Codes in Handwritten Addresses", In Proc. of 5th ICDAR, pp. 325-328, 1999.

[3] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: A comprehensive survey", IEEE Trans. on PAMI, 22, pp. 62-84, 2000.

[4] S. N. Srihari, S-H. Cha, H. Arora and S. Lee, "Individuality of Handwriting", Journal of Forensic Sciences, 47(4), pp. 856-872, 2002. http://www.cedar.buffalo.edu/~srihari/papers/JFS-2002.pdf visited on 22/7/2011.

[5] U. Pal and B. B. Chaudhuri, "Indian Script Character Recognition A Survey.", Pattern Recognition, 37, pp. 1887-1899, 2004.

[6] R. Sarkar, N. Das, S. Basu, M. Kundu,M. Nasipuri and D. K. Basu, "Word level Script Identification from Bangla and Devanagri Handwritten Texts mixed with Roman Script", journal of Computing, 2(2), 2010.

[7] K.Roy, "On the Development of an OCR System for Indian Postal Automation", Phd Thesis, Jadavpur University, 2008.

[8] K. Roy, S. Vajda, U. Pal, and B. B. Chaudhuri, "A system towards Indian postal automation", In Proc. of 9th IWFHR, pp. 580-585, 2004.

[9] U. Pal and S. Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proc. 7th Int. Conf. on Document Analysis and Recognition, pp.1128-1132, 2003.

[10] K. Roy and A. Banerjee, "Segmentation of Bangla Handwritten Character for Indian Postal Automation", In Proc. 2nd National Conference on Advanced Image Processing and Networking, pp. 1-11, 2005.

[11] http://www.ethnologue.com/ethno_docs/distribution.asp?by=country visited on 22/7/2011.