American Journal of
Electronics & Communication

Poojarini Mitra et. al., American Journal of Electronics & Communication, Vol. I (3), 8-11

Character Segmentation for Handwritten Bangla Words using Image Processing

Poojarini Mitra, Kaustav Bhattacharjee, Anirban Das, Sayan Kumar Dey, Deepjyoti Chakraborty, Aritra Ghosal, Shadab Akhtar

Computer Science Engineering

University of Engineering and Management

Kolkata, India

poojarini.mitra@uem.edu.in, kaustuv.bhattacharjee@uem.edu.in, anirbandas01011984@gmail.com, sayankumardey908@gmail.com, deepjyoti19982018@gmail.com, aritra.ghosal.96@gmail.com, shadabdsw@gmail.com

***Abstract*--- Character segmentation has long been a critical area of the OCR process. The higher recognition rates for isolated characters vs. those obtained for words and connected character strings well illustrate this fact. A good part of recent progress in reading unconstrained printed and written text may be ascribed to more insightful handling of segmentation.**

To take care of variability involved in the writing style of different individuals in this paper we propose a robust scheme to segment unconstrained handwritten Bangla texts into lines, words and characters. For line segmentation, at first, we divide the text into vertical stripes. Stripe width of a document is computed by statistical analysis of the text height in the document. Next we determine the horizontal histogram of these stripes and the relationship of the minimal values of the histograms is used to segment text lines. Based on vertical projection profile lines are segmented into words. Segmentation of characters from handwritten words is very tricky as the characters are seldom vertically separable. Segmentation of cursive handwriting is the challenging step of Optical Character Recognition (OCR).The recognition accuracy will highly depend on the good segmentation. Segmentation of cursive handwriting is very difficult. The segmentation can be done on the basis of zoning, a line segment of text, a word segment from line and character segment from word. This can be done by the use of horizontal, vertical methods. This paper reviews many basic and advanced techniques of handwritten word segmentation.

I. INTRODUCTION

Optical character recognition is a program that translates a scanned image of a document into a text document that can be edited. Segmentation of cursive handwriting is very difficult. Character segmentation is an operation to decompose an image into the sub- image of individual symbols. There are mainly three phases of a character recognition system, namely pre- processing, segmentation and recognition. Pre-processing aims to produce data that is easy for the OCR system to work accurately. It reduces noise and distortion, removes skewness and performs skeletonizing of the image, thereby simplifying the processing of the rest of the stages. The next stage is segmenting the document into its sub-components. It separates the different logical parts, like text from graphics, line of a paragraph, and character of a word. Segmentation of unconstrained handwritten text lines is difficult because of inter-line distance variability and base-line skew variability. Components of two consecutive text-lines may be touched or overlapped in unconstrained handwritten text. These overlapping or touching characters complicate the line segmentation task greatly. In Bangla touching or overlapping occurs frequently because of modified characters of upper-zone and lower- zone. Most of the characters in Bangla handwritten words are touching and segmentation of touching characters is the main bottleneck in the handwritten recognition system. Many techniques have been proposed on touching character segmentation. One class of approaches uses contour features of the component for segmentation. Some researchers use profile features for touching character segmentation. Thinning based methods are also reported for touching characters segmentation. Combined features based methods are also used for the touching string segmentation. Although many methods on handwritten line, word and character segmentation have been published in the literature for Roman, Chinese, Japanese and Arabic scripts only one report is available on character segmentation from Bangla handwritten isolated words. They used recursive contour following technique for character segmentation from a word. In this paper we propose a robust scheme to segment handwritten texts of Bangla script into lines, words and characters. Bangla is the second-most popular language in the Indian sub-continent and fifth-most popular language in the world. For line segmentation we divide the text into vertical stripes and determine horizontal histogram projections of these stripes. The relationship of the minimal values of the histograms is used to segment text lines. Based on vertical projection profiles, lines are segmented into words. Segmentation of characters from handwritten words is difficult as the characters are mostly connected in a word. For character segmentation we first detect isolated and touching characters in a word.

II. PROPERTIES OF BANGLA SCRIPT

The alphabet of the modern Bangla script consists of 11 vowels and 39 consonants. These characters may be called basic characters. The basic characters of Bangla script are shown in Fig.1. Writing style in the script is from left to right. The concept of upper/lower case is absent in Bangla script.

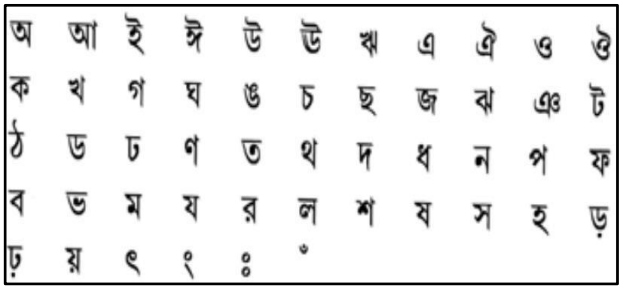
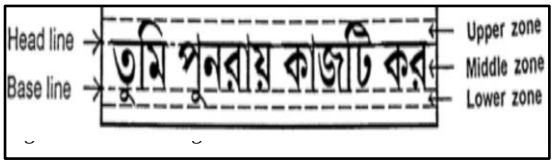


Fig.1. Basic characters of Bangla alphabet. (First 11 are vowels and rest are consonants)

From Fig.1.it is noted that most of the characters have a horizontal line at the upper part. When two or more characters sit side by side to form a word, these horizontal lines touch and generate a long line called head-line. We use these head-line characteristics for isolated and connected character identification and character segmentation. From a statistical analysis we notice that the probability that a Bangla word will have at least one character with head-line is 0.994. Hence, the use of headline based features is justified for the purpose. 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. 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 combinations of consonant and consonant, as well as consonant and vowel. A Bangla text line can be partitioned into three zones. The upper-zone denotes the portion above the head-line, the middle zone covers the portion between head-line and base-line, the lower-zone is the portion below base-line.



III. LINE, WORD AND CHARACTER SEGMENTATION

Segmentation processes, including following processes:

A. *Line segmentation*

Line segmentation is the process in which from the image, we extract only lines or differentiate the lines. Horizontal projection of a document image is most commonly used to extract the lines from the document. The horizontal projection will have separated peaks and valleys for the lines that are well separated and are not tiled, which serve as the separators of the text lines. These valleys are easily detected and used to determine the location of boundaries between the lines. Word segmentation is the process in which from the line segmentation, we extract only words. As we know that there is a distance between one words another word, this concept is used for word segmentation.

B. *Word segmentation*

Word segmentation is a process of dividing a string into its component words. Word splitting is the process of parsing concatenated text to infer where word breaks exist. By using a vertical projection profile, one can get column sums. By looking for minima in the horizontal projection profile of the page, we can separate the lines and then separate words by looking at minima in the vertical projection profile of a single line. By using the valleys in the vertical projection of line image, one can extract words from a line and also extract individual characters from the word. The global horizontal projection method computes the sum of all black pixels on every row and constructs a corresponding histogram. Based on the peak/valley points of the histogram individual lines are segmented. This method has some drawbacks: (a) it will not work on skewed texts (b) it will not work on overlapping situations (c) some diacritical points in Bangla can generate false separating lines. To take care of above drawbacks we use a modified *piece-wise projection* method suitable for Bangla script. Here we assume that a document page is in portrait mode. In this method we divide the text into vertical stripes of width W. Width of the last tripe may differ fromW. If the text width is Z and the number of stripe is N then the width of the last stripe is [Z-W*(N- 1)]. Computation of W is discussed later. Next we compute Piecewise Separating Lines (PSL) from each of these stripes. We compute the row wise sum of all black pixels of a stripe. The row where this sum is zero is a PSL. We may get a few consecutive rows where the sum of all black pixels is zero. Then the first row of such consecutive rows is the PSL. The PSLs In character segmentation, we extract only characters from word. Character segmentation is a difficult step of OCR systems as it extracts meaningful regions for analysis. This step decomposes the images into classifiable units called characters. A poor segmentation process leads to incorrect recognition or rejection segmentation process carried out only after the preprocessing of the image of different stripes of a text are shown by black horizontal lines. All these PSLs may not be useful for line segmentation. We choose some potential PSLs among these. We compute the normal distances between two consecutive PSLs in a stripe. So if

there are *n* PSLs in a stripe we get *n-1* distances. This is done for all stripes. We compute the statistical mode (MPSL) of such distances. If the distance between any two consecutive PSLs of a stripe is less than MPSL we remove the upper PSL of these two PSLs. PSLs obtained after this removal are the potential PSLs. The potential PSLs of Fig.4(a) are shown in Fig.4(b). We note the left and right coordinates of each PSL for future use. By proper joining of these potential PSLs we get individual text lines. It may be noted that sometimes because of overlapping or touching of one component of the upper line with a component of the lower line, we may not get PSLs in some regions. We will take care of these situations during PSLs joining.

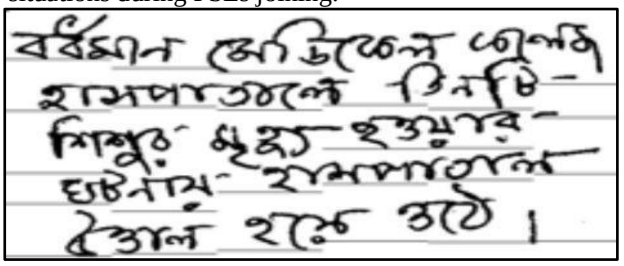


Fig.3: PSLs and potential PSLs are shown in (a) and (b)respectively, for a Bangla handwritten text

C. Character Segmentation

In character segmentation, we extract only characters from words. Character segmentation is a difficult step of OCR systems as it extracts meaningful regions for analysis. This step decomposes the images into classifiable units called characters. A poor segmentation process leads to incorrect recognition or rejection segmentation process carried out only after the preprocessing of the image.

IV. SKEW DETECTION AND CORRECTION

Prior to the segmentation it is necessary to preprocess all word images. Initially the images are in a new variable and are used to create the histogram. Each element of this variable represents the number of black pixels of the main word image. Black pixels are then plotted in an image file to get the histogram grayscale format. The gray scale is median filtered and then using the algorithm is used to binarize the images. Document skew is a distortion that often occurs during document scanning and copying. This mainly concerns the orientation of text lines and with no skew the lines are horizontal or vertical, depending on the language. This effect visually appears as a slope of the text lines with respect to the X axis. Document skew is an unavoidable effect because of the complex structure of handwritten words and the copying and scanning process especially when digitizing the huge amount of documents.

V. CHARACTER SEGMENTATION

After word segmentation and skew correction we are getting each word as different images. Now on each word we are segmenting each character for recognition. In the beginning inputting each word in grayscale, it is noticed that some pixels in the images are gray. To eliminate those pixels, the image is converted into black and white.

Fig.4: Original Word

The output is then saved as an image variable.

On that image variable, the black pixels are counted for each row and it is saved in a new variable. This new variable is used to create the histogram. Each element of this variable represents the number of black pixels of the main word image. Black pixels are then plotted in an image file to get the histogram.

Fig.5: Histogram of Original Word

In the histogram it is noticed, in the “Matra” region the occurrence of black pixels increased drastically if compared to the change in occurrence of black pixels for the rest of the image. So we calculated the highest change in occurrence to determine the “Matra” region of the image. Now from that very point (region) we segregated the upper zone and middle zone of the word and saved the segregated images into files. The upper zone is saved as “upper” and the middle zone is saved as “mid”.

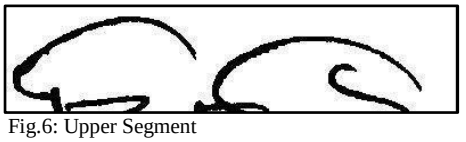


Fig.6: Upper Segment

Fig.7: Mid Segment

As we already got the middle zone now it's all about segmenting the middle zone and upper zone for character segmentation. To do so we copied the image “mid” into a new variable. In most of the cases after

segregation from the "Matra" region the "Matra" is eliminated. We noticed on multiple words that there is a minimum of 4 pixel gap in between characters. So we drew some black lines on the image "mid" in a way so that if there is any black pixel occurrence on those drawn lines, those pixels will become white.

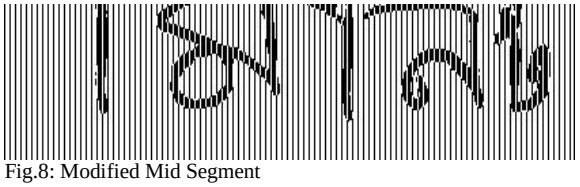


Fig.8: Modified Mid Segment

Now in this new image only those lines will black, which does not go through any character. The column index of those black lines is saved in a list variable. There can be multiple black lines one after the other. To eliminate those extra black lines we calculated, if the difference of two corresponding elements of the list is 4, the first element is removed. In this way we got the segmentation points. Using these segmentation points the "mid" image is segmented and each segmented image is saved in image files. The image files are numbered depending upon the segregation number.

VI. CONCLUSION

In this paper we have proposed a scheme for Segmentation of unconstrained handwritten words into different zones(upper and lower) and characters. There are many difficulties in respect to segmenting 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 Bangla words. We achieved some success rate in the proposed system. The word and character segmentation is a kind of a procedure which will be of great help in the field of Bangla handwritten word recognition. In the future this segmentation can be helpful for developing an automated Bangla handwriting recognizer and analyzer. This can be also helpful in future for graph logical analysis of handwritten documents.

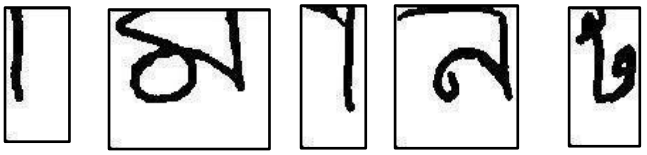


Fig.9: Character Segmentation of Mid Segment

VII. REFERENCES

[1] U. Pal and Sagarika Datta "Segmentation of Bangla Unconstrained Handwritten Text"

[2] T. K. Bhowmik, A. Roy and U. Roy, "Character Segmentation for Handwritten Bangla Words using Artificial Neural Network"

[3] B.Chaudhari and U.Pal, "Skew angle detection of digitized Indian Script documents".

[4] A.Bishnu and B.B.Chaudhari "Segmentation of Bangla handwritten text into characters by recursive contour following".

[5] R.G. Casey, E. Lecolinet, "A survey of methods and strategies in Character segmentation.

[6] K. Kurino, Y. Nakano, H. Fujisawa, "The book of segmentation methods for character recognition from segmentation to document structure analysis.