I.J. Information Technology and Computer Science, **2013, 05, 38-43**
Published Online April 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2013.05.05

A New Technique for Segmentation of Handwritten Numerical Strings of Bangla Language

Md. Aktaruzzaman
Assistant Professor, Dept. of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh
Email: mazaman_iuk@yahoo.com

Md. Farukuzzaman Khan
Professor, Dept. of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh
Email: mfkhanbd2@gmail.com

Dr. Ahsan-Ul-Ambia
Associate Professor, Dept. of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh
Email: ambiaiu@yahoo.com

***Abstract*—** Segmentation of handwritten input into individual characters is a crucial step in connected handwriting recognition systems. In this paper we propose a robust scheme to segment handwritten Bangla numbers (numerical strings) against the variability involved in the writing style of different individuals. The segmentation of digits from a number is usually very tricky, as the digits in a Bangla number are seldom vertically separable. We have introduced the concept of Degenerated Lower Chain (DLC) for this purpose. The DLC method was proved efficient in case of segmenting handwriting digits in our experiments. Ten pages of handwritten Bangla numerical strings containing 2000 individual digits that construct 700 numbers written by five different writers of variable ages were segmented by the developed system. The system achieves more than 90% segmentation accuracy on average.

***Index Terms*—**Bangla, Handwriting Style, Degenerated Lower Chain, Connected Digits

I. Introduction

With increasing the interest of computer applications, the Bangla alphanumeric recognition of typed or handwritten text bears high importance. Bangla language has 49 letters in its alphabet and 10 digits in decimal number system. There is no capital letters. The Bangla optical character recognition (OCR) has numerous applications including the digitization of old and rare Bangla books, newspapers, journals, etc., which would save a lot of time compared to manually typing all the words and numbers in those books and articles. Other important applications of handwriting recognition involves in many areas are census taking, postal code reading, and mail routing, etc [1]. A document image can contain every type of characters (alphabet and numerical digit). With the recent research development in this field, the recognition of printed text has reached an exploitation level. However the recognition of handwritten text is still in the development stage. Though the achievement in this fascination field is not enough to reach the ultimate goal, the progress of such research in English is significant. But the progress of such research in Bangla is in initial level. Bangla is spoken by about 245 million people of Bangladesh and two states of India. The people of Bangladesh, West Bengal and Tripura (two states in India) speak and write Bangla as their first language [2]. Bangla is the fifth most popular language in the World [2]. An online Bangla handwritten recognition system has been reported in[3] that uses artificial neural network for feature selection and extraction, and achieves an average recognition rate about 90%. The segmentation for Bangla character plays an important role in recognition because it allows the recognition system to classify the characters more accurately and quickly. In this paper, we have given emphasis for segmentation of numerical strings (numbers) into individual digits.

Segmentation is a major problem in the handwritten text recognition and there are very few such algorithms for Bangla described in some articles [4][5][6] in recent years. The usual approaches for segmentation are:

The Curved Pre Stroke Cut (CPSC) segmentation algorithm evaluates a large set of cuts through the image of the input string using dynamic programming and selects a small“optimal” subset of cuts for segmentation [7].In Pixel scanning method [5], certain

range of intensity values are searched through a row or column of the bitmap of an image. Segmentation based on Vertical Projection Profile scans a text line vertically, if in one vertical scan two or less black pixels are encountered the scan is denoted by 0 (zero), else the scan is denoted by the number of black pixels [8]. In this way a vertical projection profile is constructed. Another technique of Segmentation based on water reservoir principle described in some literature [9]. In this technique, if we pour water on the top of a character, the positions where water will accumulate are considered the reservoirs. For detecting segmentation points, the considerations are the direction of water overflow from the reservoir, the height of the water level in the reservoir and position of the reservoir with respect to the character-bounding box.

Among the segmentation methods as discussed above, some are not so efficient for handwriting text segmentation and some of them are only fit for printed text segmentation. In the recognition of handwritten Bangla numbers, segmentation plays an important role as most of the digits of a number are generally keeping connected in handwriting styles. In this paper, we propose a new technique to segment cursive handwritten Bangla numerical strings into digits.

II. Handwriting Styles

Handwriting occurs in different kinds of styles. The two broad classes are hand printed handwriting and cursive handwriting. In a cursive handwriting style, characters are deliberately linked together, while in a hand printed style, characters are generally generated as more or less distinct units but may touch accidentally1. The properties of some individual Bangla digits are shown in the Figure 1.

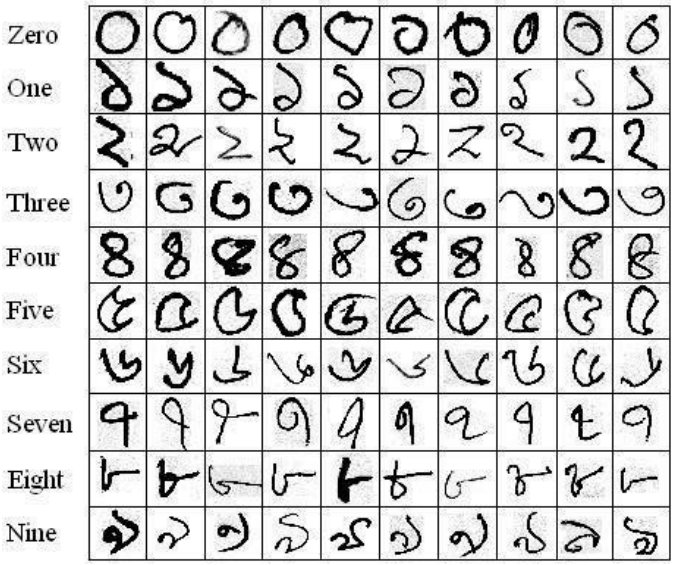


Fig. 1: a sample set of styles of Bangla numerals (digits)

III. Segmentation of Handwritten Bangla Numbers

Before segmentation, a number is converted from bitmap image format to a text format that contains only 0's (black pixels) and 1's (white pixels) using a threshold value. Thus we get a two dimensional matrix of 0's and 1's from a bitmap image. The digits in a number may have the following states:

i. Isolated: it is not connected or overlapped with it's neighbors digits.
ii. Connected: there is no blank space between two adjacent digits. They have touched each other in at least one point.
iii. Overlapped: the digits are not connected but they are not vertically separated. There is no vertical blank space lie between two consecutive digits.

These situations are described in Figure 2.

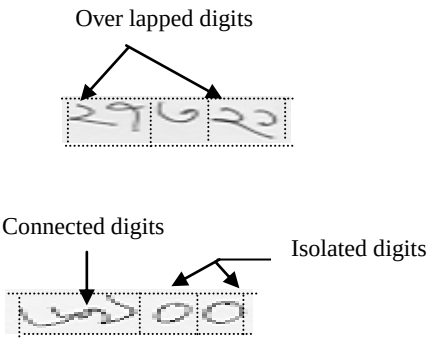


Fig. 2: Various states of digits in a number

3.1 Isolated digit segmentation

For segmentation of a number into it's isolated digits, the binary matrix of the number is scanned vertically from left to right for a column that has at least one 0 (zero). This column is taken as the left boundary of a digit. The process is continued through the consecutive columns of the number until it is found a column that has no 0's (zeros). The immediate previous column of this column is considered as the right boundary of the digit. Now the region between left and right boundary is segmented. In this way, vertically separated digits or substrings (consists of connected and overlapped digits) from a number are segmented.

So, this vertical scanning technique fails to segment connected and overlapped digits. For segmenting these types of digits, we have proposed a new technique that is degenerated lower chain (DLC).

3.2 Degenerated lower chain (DLC) technique

The idea of DLC used in this paper has been derived from the concept of degenerate polygon and lower chain of a polygon [10]. The binary matrix of a number consists of only two Bengali digits 6 and 9 connected in a single point is shown in Figure 3.

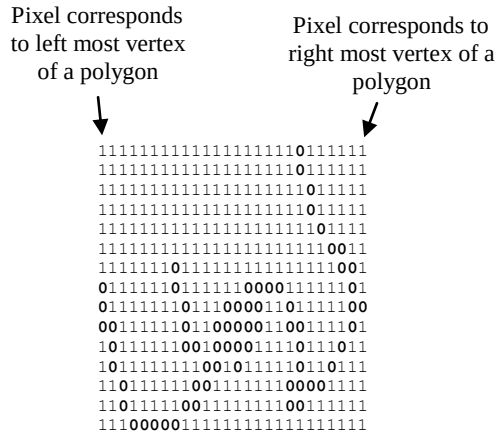


Fig. 3: Binary matrix of (69)

In the binary matrix of a number as shown figure 3, if the bottom most 0s of each consecutive column are joined by a virtual line segments. Then we get a line between leftmost and rightmost 0's that represent leftmost and rightmost vertex, respectively of a lower chain of a degenerate polygon. In this paper, this line is referred to as degenerated lower chain. The degenerated lower chain for 69 ( ) of Figure- 3 is shown in Figure-4.

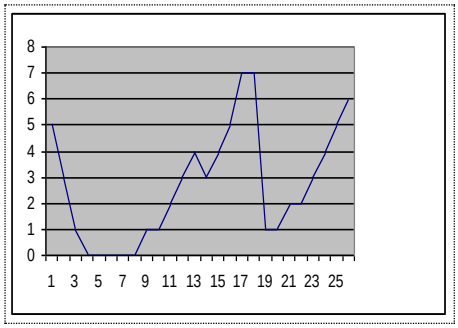


Fig. 4: DLC for Bengali number (69) where 6 and 9 are connected

3.3 Overlapped digit segmentation

After isolated digit segmentation, it is assumed that all digits are virtually isolated. The width of each isolated digit is calculated and if it is greater than the average width, W_{avg} of a digit, then it may be either overlapped or connected digit. The overlapping of digits in a Bengali number occurs very rare. To segment overlapped digits, a search for a white channel of at least one bit (1) is made at each row of the number. If such a channel is found then the digit lies on the left and right banks of the channel are separated using 8-connected paths[11]. The segmentation of two overlapped digits is described by Figure 5 and figure 6. The white channel is described by the bold 1's in the region between the two digits.

Fig. 5: Before segmentation of overlapping Bengali digits 2 and 7 in a number 27 (27)

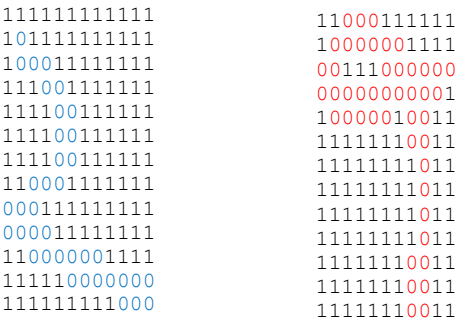


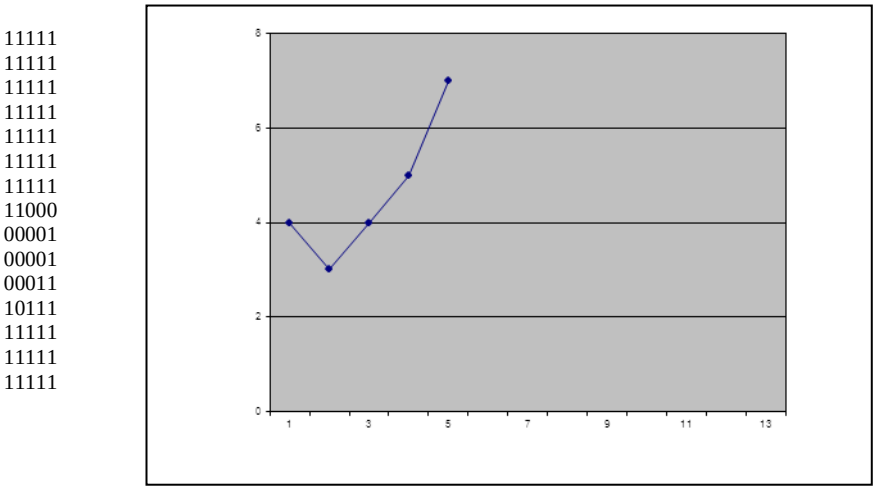
Fig. 6: segmentation of 2 and 7 from 27

3.4 Determination of connected digits

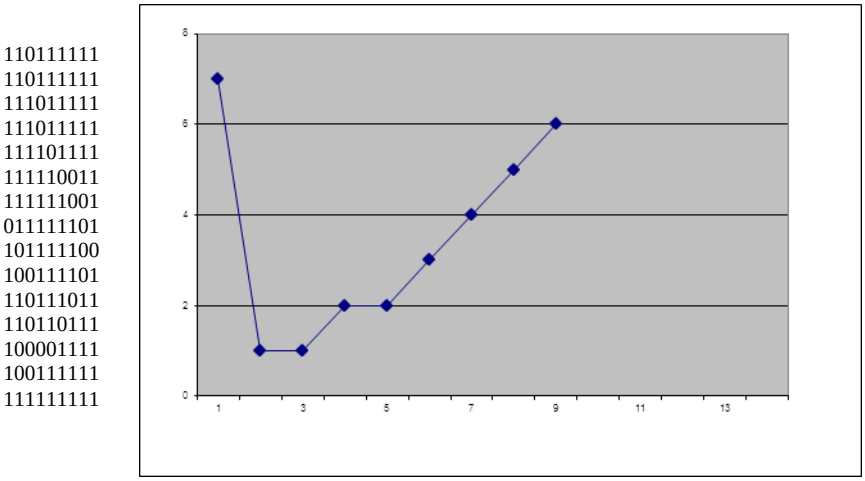
In our research work, the determination of connected digits was done by the nature of DLC. If the DLC shown in Figure-4 is traversed from left to right, with increasing X-coordinate of a vertex, the Y-coordinate of the vertex is either increased or decreased or remains constant. Then DLC for a single digit is defined by the following regular expression:

Y-coordinate of DLC=D(D)*C*I(I)* (1)

Where, D=decrease in Y-coordinate by 1, C=Y-coordinate remains constant by 1, and I=increase in Y-coordinate by 1.The number of patterns described in (1) in a number or numerical string determines the number of connected digits.



(a): Left segmented part of 9 (9) and it's LC



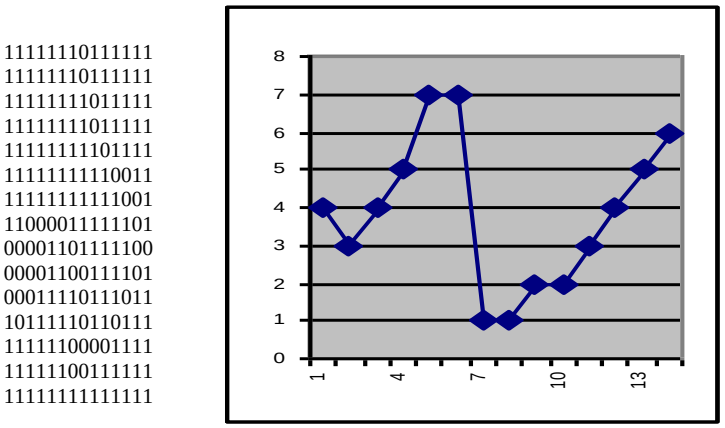
(c): Right segmented part of 9 (9) and it's DLC

Fig. 7: Over-segmentation of 9 (9)

3.4.1 Connected digit segmentation:

To segment connected digits, DLC is scanned from left to find a lower transition point. After this the scanning is continued for an upper change point. This upper change point determines the last column of pixels of left digit of the connected or overlapped digits in consideration. The next column of pixels is the starting column of the next digit. However, in this method digit 9 and digit 7 are sometimes over segmented .The over segmentation of digit 9 is illustrated in Figure-7. To eliminate this over segmentation, the height of the segmented parts are calculated and compared with the maximum height of the connected or overlapped digits in consideration. If the height of a segmented part is less than the half of the maximum height then it is

considered as wrong segmentation and the segmented parts are merged together. The merging of two over segmented parts is shown in Figure-8.



(a) (b)

Fig. 8: (a) binary form and (b) DLC of 9 (9) after merging of it's over segmented parts

IV. Results and Discussion

The DLC method described in this paper has been proved as an efficient technique to segment Bangla handwritten numbers of all styles. To study the performance of this method a software system was developed and ten pages of Bangla numerical text written by five different writers were used to test the performance of the system. 2000 individual digits that construct 700 numbers of different lengths were contained in these pages. The experimental results are summarized in Table-1.

The developed method has successfully segmented Bangla digits from numbers containing both connected and overlapped digits with some limitations. The system cannot segment correctly if the digits are both connected and overlapped. This situation rarely happens in Bangla text. In future, the system can be combined with water reservoir technique to improve the segmentation performance.

Table 1: The average accuracy rate of segmentation of the system

Writer No.	No. of Correct Segment Expected	No. of Correct Segment Obtained	Segmentation Accuracy (%)
1	313	294	93. 92
2	457	416	91.02
3	342	309	90.35
4	400	361	90.25
5	288	254	88.19
Average Segmentation Accuracy Rate (%)			90.75

References

[1] M. S Islam. Research on Bangla Language Processing in Bangladesh: Progress and Challenges 8th International Language and Development Conference, 23-25, June 2009, Dhaka, Babgladesh

[2] U. Pal and Sagarika Datta. Segmentation of Bangla Unconstrained Handwritten Text. Proceedings of the Seventh International Conference on Document Analysis and Recognition ICDAR 2003, IEEE, 0-7695-1960-1/03, 2003.

[3] M Badruddoza, Reocgnition of Bangla handwritten letters using self-organizing map(SOM). Proceeding of 6th International Conference on Computer and Information Technology(ICCIT),PP. 357-360.

[4] A. O. M. Asaduzzaman, Mst Shayeala Parveen, and M Ganjer Ali. Detection of Bangla Numbers

Using Artificial Neural Network. 6th ICCIT-2003, Jahangirnagar University, Dhaka, Bangladesh, pp347-350, 2003.

[5] Md. Farukuzzaman Khan, Md. Mizanur Rahman, Md. Aktaruzzaman, Md. Robiul Hoque, Md. Monirul Islam. System Development For Optical Character Recognition in Bangla, Journal Of Applied Science And Technology, Islamic University Studies, Vol-3, Part-1

[6] Md. Khademul Islam Molla and Kamrul Hasan Talukder. Bangla Number Extraction and Recognition from Document Image. 5th ICCIT 2002, East West University, 27-28 December 2002.

[7] Thomas M. Brueuel. Segmentation of Handprinted Letter Strings using a Dynamic Programming Algorithm. Website: tbreuel@parc.xerox.com, Xerox PARC, Palo Alto, CA, USA.

[8] Md. Rafiul Hasan, Mohammad Azizul Haque and Syeda Umme Farhana Malik. Bangla Optical Character Recognition System. International Conference on Computer and Information Technology, ICCIT'99, SUST, Bangladesh, pages 164-168, December 1999.

[9] B B Chaudhury, U Pal and M Mitra. Automatic Recognition of Printed Oriya script", Sadhana Vol. 27, Part 1, pp. 23-34. India, February 2002.

[10] Michael J. Laszlo. Computational Geometry and Computer Graphics In C++. Second Edition, Prentice-Hall of India Private Ltd., February-2002.

[11] Rafael C. Gonzalez. Digital Image Processing. 3rd Edition, PHI Learning Private Limited, New Delhi, 2008.

Authors' Profiles

Md. Aktaruzzaman: Assistant Professor of the department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh. His research interest includes Medical Image Processing, Pattern Recognition and Computer Vision.

Md. Farukuzzaman Khan: Professor of the department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh. His research interest includes Speech Processing, Pattern Recognition.

Dr. Ahsan-Ul-Ambia: Associate Professor of the department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh. His research interest includes Medical Imaging, Bioinformatics and Pattern Recognition.

