Comparative analysis of connected component for Bengali Character recognition.

Pritilata Biswas
1103017

Sumaia Afrin Prity
1103057

Parvez Abedin Siddique
1103094
Section A
Level 4, Term 2
Department of Computer Science and Engineering
BAIUST

October 13, 2019

1 Topic Characteristics

Segmentation of Bangla handwritten text is the pre-process of handwritten text. The bengali script is also used for representing other language like Meithei and Bishnupriya Manipuri, and there is time when it used to write Sanskrit within Bengal [1]. Most of the Bengali characters are topologically connected [2]. They can be easily classified using this connectedness property. This classifying or segmentation of Bengali character is the most challenging task in making a Bengali Optical Character Recognition (OCR). In Bengali there is a combination of basic, modified and compound characters. The combination of characters can be more than 300. Here, Bengali character segmentation is the most challenging task. So, we need some robust techniques that can classify these compound characters.

We know Bengali script characters are connected with a Matra line, so the characters can be easily recognized by the connected component analysis [2]. In this paper our main contribution is

to improvise the method of analyzing connected component using some common image processing operations.

In this 21st century of digital generation, OCR is one the most important technique to interface between human and machines. OCR of printed text has already achieved a huge success. ABBYY is one of the major examples of most fine reader that is commercially available [3]. A little works has been done on Bengali Handwritten character recognition [9], [10], [11], [12]. Now, an efficient Bangla OCR system is important to be developed. Many researchers have already developed many methods of segmentation of Bengali script.

Chaudhuri et al [9] first proposed an OCR system using HPP and VPP and headline removal technique they segmented the documents into lines, words, and characters. Mahmud et al [10] first proposed an that supports multi font characters. Mahmud et al [11] used Depth First Search (DFS) algorithm for overlapped character segmentation. Hasan et al [12] proposed a hamming network for Segmentation. All of these method, they divide the words into three regions (matra line, upper zone, lower zone). For dividing them into regions one of the best approaches is to detect the Matra line and remove them. But removing them incurs some major problems in most of the characters as they are incomplete without their Matra line. So, in this paper we proposed a method that may counter this problem.

2 Objective

- To enhance Bengali connected component analysis
- To improvise Profile Projection with Opening, Closing operations
- To implement Opening, Closing operations, Profile Projection with Erosion, Dilation operations are needed.
- Comparative analysis between existing and proposed method.

3 Working Hypothesis

A typical OCR system consists of preprocessing, segmentation, recognition and post-processing parts. In this paper we mainly

focused on the segmentation process which is the back bone of the recognition phase.

At first the input image will be converted into grayscale image using grayscale filter and using bilateral filter we will remove the noise in the input image. Then using bit plane slicing the input image will be binarized and will be converted into two-tone image. Some input image may be skewed by a few degrees. For skew correction we will use the skew angle technique [4].

In Bengali script, lines are almost of the same height. There is a horizontal gap between two lines. We will detect those gapes by horizontal projection profile (HPP). Vertical projection profile (VPP) will be used in word segmentation from extracted lines. In this process the number of black pixels in each column will be calculated. In vertical scan if no black pixel found then it will be considered as the gap between words.

Bengali words are connected through Matra line. By removing them, characters get topologically disconnected. Then using the connected component method, individual components of a word can be separated. Now we will apply erosion and dilation to remove the problems that will occur from the removal of Matra line. Finally, we will analyze the connected component (each character) how close they are.

4 Methodology

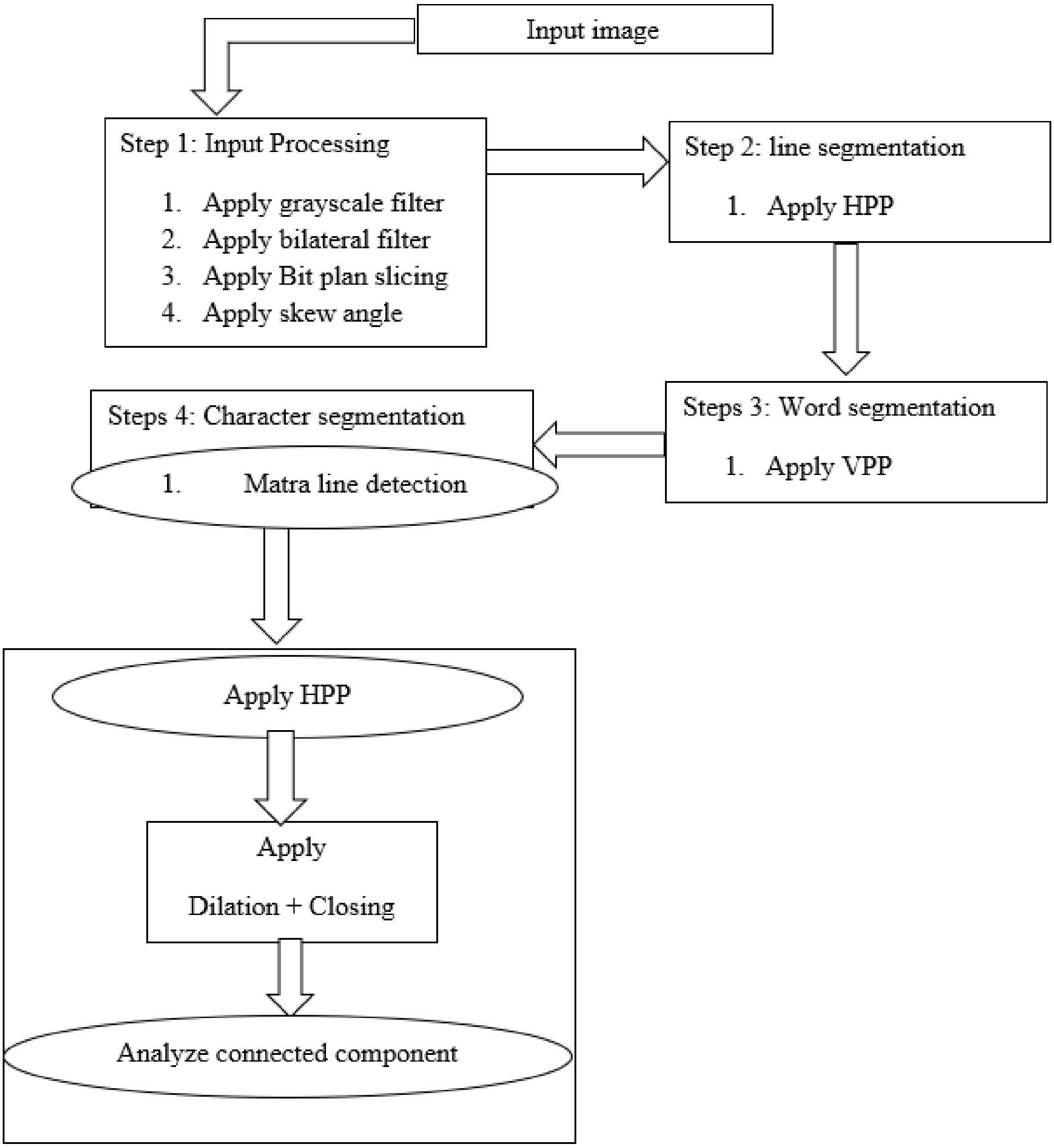
Input image
Step 1: Input Processing
1. Apply grayscale filter
2. Apply bilateral filter
3. Apply Bit plan slicing
4. Apply skew angle
Step 2: line segmentation
1. Apply HPP
Steps 3: Word segmentation
1. Apply VPP
Steps 4: Character segmentation
1. Matra line detection
Apply HPP
Apply
Dilation + Closing
Analyze connected component

Figure 1: The proposed methodology

5 References

1. https://en.wikipedia.org/wiki/Bengali_alphabet

2. T. Zahan, M.R. Selim, M.S. Rahman, M.Z. Iqbal, (2018). *"Connected Component Analysis Based Two Zone Approach for Bangla Character Segmentation."* 1-4. 10.1109/ICBSLP.2018.8554684.

3. https://www.abbyy.com/.Lastaccessed07Dec2016

4. M. A. Obaida, T. K. Roy, M. A. Horaira and M. J. Hossain, *"Skew Correction Function of OCR: Stroke-Whitespace based Algorithmic Approach,"* International Journal of Computer Applications 28(8):7-12, August 2011

5. S. S. Goswami, *"Identification of Matra Region and Overlapping Characters for OCR of Printed Bengali Scripts,"* Intelligent Computing and Information Science, Volume 135 of the series Communications in Computer and Information Science, 2011, pp. 606-612

6. https://docs.opencv.org/2.4/doc/tutorials/imgproc/erosion_dilatation/erosion_dilatation.html

7. https://homepages.inf.ed.ac.uk/rbf/HIPR2/open.htm

8. https://homepages.inf.ed.ac.uk/rbf/HIPR2/close.htm

9. B. B. Chaudhuri and U. Pal, "A Complete Printed Bangla OCR System," Pattern Recognition, Vol. 31, No. 5, pp. 531—549, 1998.

10. J. U. Mahmud, M. F. Raihan and C. M. Rahman, *"A complete OCR system for continuous Bengali characters,"* In: Proceedings of the TENCON, 1372–1376, 2003.

11. S. M. M. Mahmud, N. Shahrier, A.S.M D. Hossain, M. T. M. Chowdhury, and M.A. Sattar, *"An Efficient Segmentation Scheme for the Recognition of Printed Bangla characters,"* Proceedings of ICCIT, pp 283-286, 2003.

12. M. A. M. Hasan, M. A. Alim, and M. W. Islam *"A New Approach to Bangla Text Extraction and Recognition from Textual Image"*, Proceedings of ICCIT, 2005.