

Bangla Character Recognition System is Developed by using Automatic Feature Extraction and XOR Operation

Md. Mojahidul Islam¹, Md. Imran Hossain² and Md. Kislu Noman³

¹ Islamic University, Bangladesh.

Received: 15 December 2012 Accepted: 4 January 2013 Published: 15 January 2013

Abstract

This paper presents off-line bangla character recognition system using automatic feature extraction and XOR operation. In this system, the Bangla text is accepted as an image file which is first segmented into lines and words and then each word is segmented into characters. The pixels outside the boundary of the character are eliminated. The characters are scaled to a size equal to the database image. A XOR operation is performed between the scaled image and the database image and the error (

Index terms— character recognition; character segmentation; automatic feature extraction; XOR operation.

1 Introduction

II.

2 Implementation of Character Recognition System

The character recognition system can be divided as segmentation of text document into character and recognition of the character. The whole process is shown in the subject of character recognition has been receiving considerable attention in recent years due to the advancement of the automation process. Automatic character recognition improves the interaction between man and machine in many applications like office automation, cheque verification, mail sorting, and a large variety of banking, business and data entry applications. We are concerned here with the recognition of character in Bangla language. Bangla is the mother language of Bangladesh and approximately 10% of the world's population speaks in Indian, Chinese and other languages trying to develop the complete character recognition system. In our country, research works in this field have achieved a limited success so far as compared to the other foreign languages. Though, the achievement in this fascinating field is not enough to reach the ultimate goal. But the progress of such research with Bangla language is still in an initial level. This research is a simple flourish to implement that dream as the initial step to convert the Bangla text to computer readable form that is development of complete Bangla Character Recognition system. Individual Bangla characters were recognized using various techniques such as geometric shape analysis, black runs and concavity measurement technique. The input images are acquired from documents containing text by using scanner as an input device or using Adobe Photoshop or Paint. Acquired images are then stored in Hard Disk in JPG picture format. This image is then passed for preprocessing.

3 Source

4 b) Pre-Processing

The scanned image is converted into binary image. At first, the RGB image is converted into grayscale image and then binary image i.e. an image with pixel 0 (white) and 1 (black). After converting the image, the unnecessary pixels (0s) from the original image is removed.

5 c) RGB to Grayscale and Gray to RGB Conversion

In practical cases most of the images are generally color (RGB), but it is complex to work with a three-dimensional array. So it needs to convert the RGB image into the grayscale image. The RGB to grayscale conversion is performed by MATLAB command.

6 $I = \text{rgb2gray}(f)$

For ease of analysis, the grayscale image is converted into binary image by using the following MATLAB command.
 $BW = \text{im2bw}(I)/III.$

7 Text Segmentation

Text segmentation is a process where the text is partitioned into its elementary entities i.e. characters [10]. The total performance of the character recognition process depends on the accuracy of the segmentation process of the text into the characters. In the segmentation phase, first the document is segmented into text lines, the text lines are segmented into text words and then the words are segmented into characters.

8 a) Line Segmentation

Text line segmentation is performed by scanning the input image horizontally. Frequency of black pixels in each row is counted to separate the line. The position between two consecutive lines, where the number of black pixels in a row is zero denotes a boundary between the lines [13]. The output image is shown in In English text there is a minimum gap between two consecutive characters and two consecutive words. The minimum gap between two consecutive words is greater than two consecutive characters. Although maximum characters in Bangla text line are connected by matra line with each other, the same case occurs if the gap exists between them. For word segmentation from the text line, the vertical scan is performed. If there exists n consecutive scan that find no black pixel, we denote it to be a marker between two words. The value of n is the minimum gap between two consecutive words which is taken experimentally. The output is shown in For character segmentation from the word, the vertical scan is performed. The starting boundary of a character is the first column where the first black is found. After finding the starting boundary of a character, it continues scanning until a column without any black pixel is found, which is the ending boundary of the character being processed [14]. Fig. 4 shows a single segmented character and its corresponding binary format. The knowledge base is designed based on the feature matrix of various characters. In order to build the knowledge base, first, the RGB character image is converted into grayscale image then it is converted into binary image. After getting the binary image, the unnecessary pixels from the character boundary is eliminated.

9 e) Feature Extraction

Feature extraction is the process of extracting essential information content from the image segment. It plays an important role in the whole recognition process [10].

10 f) Scaling

Depending on the height and width of the database image the segmented characters are scaled. If the size of the segmented character is higher than the database character then the system will be scaled down all the segmented characters to the size of the database character, otherwise scaled up. If C be the segmented character then the scaled image S is obtained by the following MATLAB command: $S = \text{imresize}(C, [\text{height}, \text{width}])$.

Where, height and width is the dimension of the database character. Character recognition performance depends on the scaling. If the segmented character is too higher or too lower than the database image then the character recognition performance is reduced. The character recognition procedure is described in following Algorithm: BEGIN 1. Calculate total pixel = height $\times$ width. 2. Take XOR between first database character and scaled character S . 3. Calculate no. of correct pixels (0 is the correct pixel), correct pixel. 4 In this way, for all database character the error (%) calculation is repeated. If the database character exactly or approximately matches with the segmented character then the error (%) will minimum. So base on the minimum error, the system gives the corresponding output character.

11 IV. Result and Performance Analysis

The system is divided in two main phases: segmentation and character recognition. So the overall performance of the system directly depends on the performance of the two individual phases. The accuracy of this system is measured as the success rate for the recognition of characters. It is measured using Eq. (?? The segmentation performance of this system is shown in Table 1. V.

12 Discussion and Conclusion

The aim of this system is to recognize Bangla characters. This system can recognize these characters with slight limitations. The limitations are discussed in the following section.

13 a) Limitation

The performance of this system depends on the segmentation and recognition. If the characters of text are in very close or overlap to each other, then the system fails to segment the characters. For Bangla characters, different font size is possible in practical. It is not possible to store all the font size in database. So it needs to scale the character which causes distortion in character shape. It should create a problem but the system should not fail always. b) Further Scope Due to the limitations described in previous section the system is not suitable for on-line applications. The overlapping character can be segmented by using Flood fill and Boundary fill algorithm. It is further target to perform this work.

14 c) Conclusion

In this paper the off line bangle character recognition system is developed by using automatic feature extraction and XOR operation. The efficiency of this system is not so high. In future, MLP and SVM classifier can be used for character recognition.



Figure 1:

জীবনের বড় একটা সময় চার

দেয়ালের মধ্যেই কেটেছে তাঁর।

গণতন্ত্রের জন্য উৎসর্গ করেছেন নিজের জীবন।

1

Figure 2: Figure 1 :

জীবনের বড় একটা সময় চার

Figure 3:

দেয়ালের মধ্যেই কেটেছে তাঁর।

২ গণতন্ত্রের জন্য উৎসর্গ করেছেন নিজের জীবন।

Figure 4: GlobalFigure 2 :

৩ জীবনের বড় একটা সময় চার

Figure 5: Figure 3 :

৪ জীবনের

Figure 6: Figure 4 :

৫ বড় একটা সময় চার

Figure 7: Figure 5 :

৬ অ

Figure 8: Figure 6 :

1

No. of Lines in a	Line Segmentation	Word Segmentation	Character Segmentation
Text Document	Accuracy (%)	Accuracy (%)	Accuracy (%)
4	100	97	89.05
5	100	97.5	92.50
6	100	94	90.71
7	100	96.67	92.69
8	100	94	90.32

b) Segmented Character Recognition Performance
For character recognition, this system uses XOR operation which is a very simple matching technique. The character recognition performance of this system is shown in Table 2 for Shoroborno and table 3 for Numerical Character.

Figure 9: Table 1 :

2

No. of Test Sample	Total No. of Characters	Total No. of Success	Success Rate (%)	Average Success Rate (%)
1	120	90	75	76.96368
2	150	116	77.33333	
3	125	94	75.2	
4	130	102	78.46154	
5	170	134	78.82353	

Figure 10: Table 2 :

3

No. of Test Sample	Total No. of Characters	Total No. of Success	Success Rate (%)	Average Success Rate (%)
1	50	42	84	83.27363
2	70	53	75.71429	
3	65	56	86.15385	
4	40	33	82.5	
5	50	44	88	

Figure 11: Table 3 :

-
- [Jalal Uddin Mahmud et al. ()] ‘A Complete OCR System for Continuous Bangla Characters’. Mohammed Jalal Uddin Mahmud , Chowdhury Mofizur Feroz Raihan , Rahman . *Proceedings of the Conference on Convergent Technologies for the Asia Pacific*, (the Conference on Convergent Technologies for the Asia Pacific) 2003.
- [Md and Chowdhury] *An Approach to Implement Signature Recognition System Using Neural Network and Genetic Algorithm*, Md , Chowdhury . RUET, Rajshahi, Bangladesh.
- [Haque et al.] *Automatic Detection and Translation of Bengali Text on Road Sign for Visually Impaired*, S M Haque , Shahida Arbi , Tabassum Tamanna , Sadia Mahsina , Itu . http://www.daffodilvarsity.edu.bd/library/opac/DIUJSTPapers/Vol2Iss2/CR01_30090609.pdf
- [Shah Mashiyat et al. (2004)] ‘Bangla off-line Handwritten character Recognition Using Superimposed Matrices’. Ahmed Shah Mashiyat , Ahmed Shah Mehadi , Kamrul Hasan Talukder . *7th International Conference on Computer and Information Technology (ICCIT 2004)*, (Dhaka, Bangladesh) 26-28 December, 2004. BRAC University
- [Rahman et al. (1998)] ‘Bangla Sorting Algorithm: A Linguistic Approach’. Md Rahman , Md Iqbal , Zafar . *Proceedings of International Conference on Computer and Information Technology*, (International Conference on Computer and Information Technology Dhaka) 18-20 December 1998. p. .
- [Haralick and Shapiro ()] *Computer and Robot Vision*, Robert M Haralick , Linda G Shapiro . 1992. Addison-Wesley. 1.
- [Minhaz et al. (2002)] ‘Computer Representation of Bangla Character and Sorting of Bangla Words’. Fahimm Minhaz , Tanvir Zibran , Shammi Arif , Abdus Raziullah , Md . *Proceedings of 5th ICCIT*, (5th ICCIT Dhaka, Bangladesh) 2002. December 2002. p. .
- [Shapiro and Stockman ()] *Computer Vision*, Linda G Shapiro , George C Stockman . 2001. New Jersey: Prentice-Hall. p. .
- [Gonzalez et al.] Rafael G Gonzalez , Richard E Woods , Steven L Eddins . *Digital Image Processing Using MATLAB*, Pearson Education, Inc.
- [Sonka et al. ()] *Image Processing Analysis and Machine Vision*, M Sonka , V Hlavac , Boyle . 1998. PWS Publishing.
- [Image processing Toolbox User’s Guide-For Use with MATLAB, Version 2, The Math Works (1997)] *Image processing Toolbox User’s Guide-For Use with MATLAB, Version 2, The Math Works*, May 1997.
- [References Références Referencias] *References Références Referencias*,
- [Sayeed Md et al. (2004)] *Rotation Independent Image Object Recognition Using Automatic Feature Extraction and Artificial Neural Networks*, Abu Sayeed Md , A A M Sohail , M A Haque , Mottalib . ICCIT-2004. December 2004. Dhaka. p. .