Global Journal of Computer Science and Technology
Graphics & Vision
Volume 13 Issue 2 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

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

By Md. Mojahidul Islam, Md. Imran Hossain & Md. Kislu Noman
Islamic University, Bangladesh

Abstract - This paper presents off-line bangle 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 (%) is calculated. Finally, depending on the minimum error, the system recognizes the character to use in the output. The average recognition accuracy rate of the system was about 80%.

Keywords : *character recognition; character segmentation; automatic feature extraction; XOR operation.*

GJCST-F Classification: *I.5.0*

Strictly as per the compliance and regulations of:

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

Md. Mojahidul Islam [α], Md. Imran Hossain [σ] & Md. Kislu Noman [ρ]

Abstract - This paper presents off-line bangle 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 (%) is calculated. Finally, depending on the minimum error, the system recognizes the character to use in the output. The average recognition accuracy rate of the system was about 80%.

Keywords : character recognition; character segmentation; automatic feature extraction; XOR operation.

I. Introduction

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.

II. 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 Fig 1.

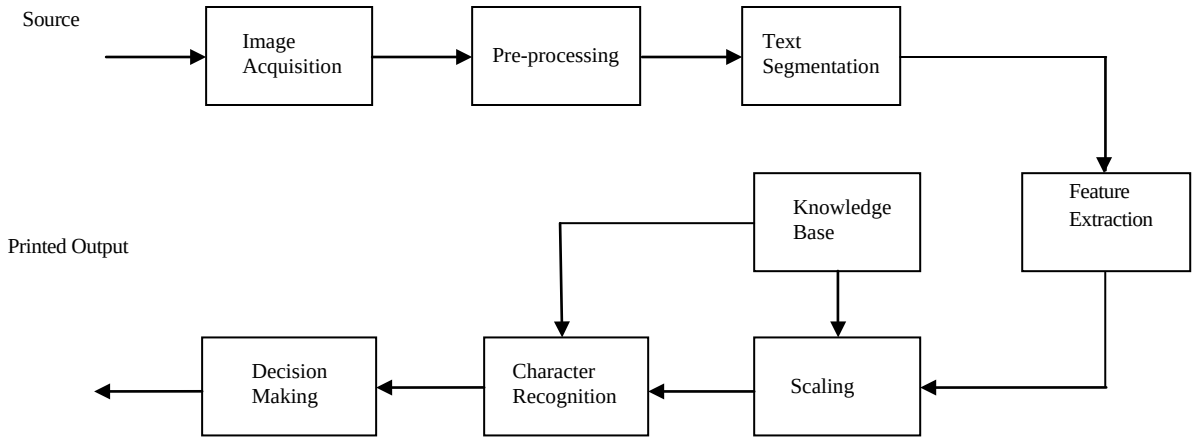


Figure 1 : Block diagram of character recognition system

Author α : Assistant Professor, Department of CSE, Islamic University, Kushtia-7003, Bangladesh. E-mail : mojahid.cse@gmail.com
Author σ : Lecturer, Dept. of ICE, Pabna Science and Technology University, Pabna, Bangladesh. E-mail : imran05ice@gmail.com
Author ρ : Lecturer, Dept. of CSE, Pabna Science and Technology University, Pabna, Bangladesh. E-mail : md.k.noman@gmail.com

a) Image Acquisition

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.

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.

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.

I = rgb2gray(f)

For ease of analysis, the grayscale image is converted into binary image by using the following MATLAB command.

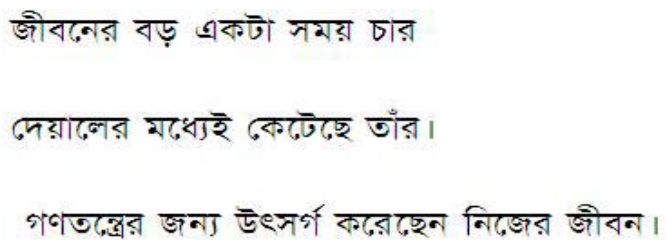
BW = im2bw(I)

III. Text Segmentation

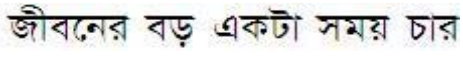
Text segmentation is a process where the text is partitioned into its elementary entities i.e. characters [10]. The total performance of the character recognition process depends on the accuracy of the segmentation process of the text into the characters. In the segmentation phase, first the document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

a) Line Segmentation

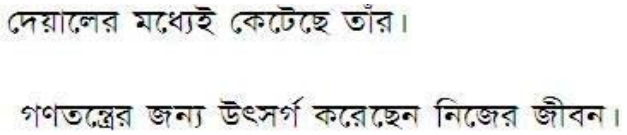
Text line segmentation is performed by scanning the input image horizontally. Frequency of black pixels in each raw is counted to separate the line. The position between two consecutive lines, where the number of black pixels in a raw is zero denotes a boundary between the lines [13]. The output image is shown in Fig 2.



(a)



(b)

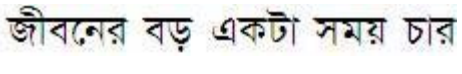


(c)

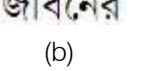
Figure 2 : Line Segmentation (a) Bangla input text image, (b) Image of first segmented line and (c) Text image without first line

b) Word Segmentation

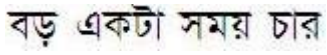
In English text there is a minimum gap between two consecutive characters and two consecutive words. The minimum gap between two consecutive words is greater than two consecutive characters. Although maximum characters in Bangla text line are connected by matra line with each other, the same case occurs 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 Fig 3



(a)



(b)

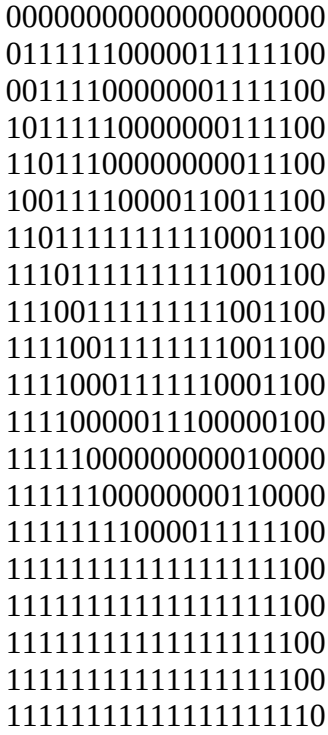


(c)

Figure 3 : Word Segmentation (a) Bangla Text Line, (b) Image of first segmented word and (c) Image without first word

c) Character Segmentation

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.



(a)

Figure 4 : (a) Binary Form of a Segmented Character

d) Knowledge Base

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.

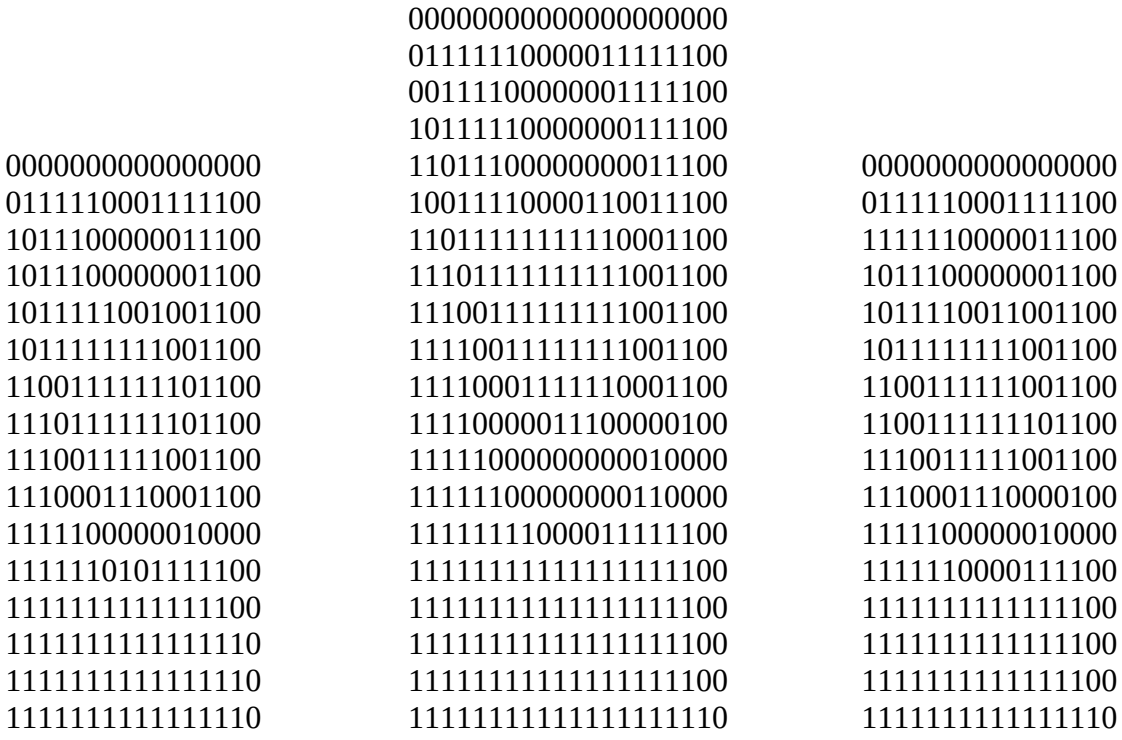
e) Feature Extraction

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].

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 = imresize(C, [height, width]).

Where, height and width is the dimension of the database character. Fig 5(a) shows a database character অ whose size is 16×16. The segmented character অ of size 20×20 shown in Fig 5(b) is scaled down to the size of database character 16×16 shown in Fig 5(c). This is repeated for all database characters and finally for all segmented characters.



(a) (b) (c)

Figure 5 : Image scaling (a) Database image of size 16×16, (b) Segmented image of size 20×20, and (c) Scaled image of (b) of size 16×16

g) Character Recognition

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×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. Calculate percentage of error using $\text{error}(\%) = \frac{\text{total}_{\text{pixel}} - \text{correct}_{\text{pixel}}}{\text{total}_{\text{pixel}}} \times 100\%$ and save error (%).
5. Repeat Step 1 to Step 4 for all database characters.
6. Calculate minimum error (%) (e_{min}) obtaining from Step 4 for database characters.
7. Define a error tolerance, $error_{tolerance}$.
8. If $e_{min} < error_{tolerance}$
   Compare e_{min} for all %error
   If e_{min}=error(i) (%)
   Then print the i^{th} character
   endif
   else
   Print 'the character is not recognized'
   endif.
9. Repeat Step 1 to Step 8 for all segmented characters
10. End

0000000000000000	0000000000000000	0000000000000000
0111110001111100	0111110001111100	0000000000000000
1011100000011100	1111110000011100	0100010000000000
1011100000001100	1011100000001100	0000000000000000
1011111001001100	1011110011001100	0000001010000000
1011111111001100	1011111111001100	0000000000000000
1100111111101100	1100111111001100	0000000000100000
1110111111101100	1100111111101100	0010000000000000
1110011111001100	1110011111001100	0000000000000000
1110001110001100	1110001110000100	0000000000001000
1111100000010000	1111100000010000	0000000000000000
1111110101111100	1111110000111100	0000000101000000
1111111111111100	1111111111111100	0000000000000000
1111111111111110	1111111111111100	0000000000000010
1111111111111110	1111111111111100	0000000000000010
1111111111111110	1111111111111110	0000000000000000
(a)	(b)	(c)

Figure 6 : Character recognition (a) Database image of size 16×16, (b) Scaled image of size 16×16, (c) Image after XOR between (a) and (b)

Total number of pixels, $total_{pixel}$ = 16×16 = 256
Total number of correct pixels (0_s), $correct_{pixel}$ = 221

$$\text{error}(\%) = \frac{\text{total}_{\text{pixel}} - \text{correct}_{\text{pixel}}}{\text{total}_{\text{pixel}}} \times 100\%$$

$$= \frac{256 - 221}{256}$$

$$= 13.6719\%$$

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.

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. (1):

$$\text{Accuracy}(\%) = \frac{\text{Number of Success}}{\text{Number of Test}} \times 100\%$$

a) Segmentation Performance

The segmentation performance of this system is shown in Table 1.

Table 1 : Text Document Segmentation Result

No. of Lines in a Text Document	Line Segmentation Accuracy (%)	Word Segmentation Accuracy (%)	Character Segmentation 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.

Table 2 : Bangla Character (Shoroborno) Recognition Result

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	

Table 3 : Bangla Numerical Character Recognition Result

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	

V. 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.

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 front 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.

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.

References Références Referencias

1. Rahman, Md. Shahidur Iqbal, Md. Zafar, "Bangla Sorting Algorithm: A Linguistic Approach". Proceedings of International Conference on Computer and Information Technology, Dhaka, 18-20 December 1998, pp: 204-208.
2. Fahimm Minhaz, Zibran, Tanvir Arif, Shammi Rajiullah and Abdus Md., "Computer Representation of Bangla Character and Sorting of Bangla Words". Proceedings of 5th ICCIT 2002, Dhaka, Bangladesh, December 2002, pp: 191-195.
3. Md. Jamil Chowdhury, "An Approach to Implement Signature Recognition System Using Neural Network and Genetic Algorithm", RUET, Rajshahi, Bangladesh.

4. Haralick, Robert m., and Linda G. Shapiro, Computer and Robot Vision, Volume 1. Addision-Wesley, 1992.
5. Sonka. M, Hlavac. V, Boyle. R, Image Processing Analysis and Machine Vision, PWS Publishing, 1998.
6. Image processing Toolbox User's Guide- For Use with MATLAB, Version 2, The Math Works, May 1997.
7. Linda G. Shapiro and George C. Stockman (2001): "Computer Vision", pp 279-325, New Jersey, Prentice-Hall, ISBN 0-13-030796-3.
8. http://en.wikipedia.org/wiki/Segmentation_(image_processing)
9. http://en.wikipedia.org/wiki/Pattern_recognition
10. Ahmed Shah Mashiyat, Ahmed Shah Mehadi and Kamrul Hasan Talukder, "Bangla off-line Handwritten character Recognition Using Superimposed Matrices", 7th International Conference on Computer and Information Technology (ICCIT 2004), 26-28 December, 2004, BRAC University, Dhaka, Bangladesh.
11. http://en.wikipedia.org/wiki/Feature_extraction
12. Rafael G. Gonzalez, Richard E. Woods, and Steven L. Eddins,"Digital Image Processing Using MATLAB", Pearson Education, Inc.
13. Jalal Uddin Mahmud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman, "A Complete OCR System for Continuous Bangla Characters", IEEE TENCON-2003: Proceedings of the Conference on Convergent Technologies for the Asia Pacific, 2003.
14. S.M. Anamul Haque, Shahida Arbi, Tabassum Tamanna and Sadia Mahsina Itu, "Automatic Detection and Translation of Bengali Text on Road Sign for Visually Impaired". http://www.daffodilvarsity.edu.bd/library/opac/DIUJSTPapers/Vol2Iss2/CR01_30090609.pdf
15. Abu Sayeed Md. Sohail, A.A.M. Mahmudul Haque and M.A. Mottalib, "Rotation Independent Image Object Recognition Using Automatic Feature Extraction and Artificial Neural Networks", pp-504-509, ICCIT-2004, Dhaka, December 2004.