SEGMENTATION OF CHARACTERS WITHOUT MODIFIERS FROM A PRINTED BANGLA TEXT

Fakruddin Ali Ahmed

Department of Computer Science & Engineering
Global Institute of Management & Technology, West Bengal, India
fakruddin_1@rediffmail.com

ABSTRACT

Optical Character Recognition (OCR) is one of the fundamental research areas of image processing and pattern recognition field. The performance accuracy of an OCR system depends on the proper segmentation of the characters. This paper is concerned with the segmentation of printed bangla characters without modifiers for optical character recognition (OCR) system. The basic steps needed for developing an OCR system also have been discussed.

KEYWORDS

Optical Character Recognition (OCR), Preprocessing, Segmentation, Classification.

1. INTRODUCTION

There was an ancient dream of researchers to invent a machine which can read text document written in a language. Today the researchers are able to develop Optical Character Recognition (OCR) system which can read the text document. We can convert an image of handwritten, typewritten or printed text to a machine editable text using OCR. OCR systems are two types, Offline and Online. In offline the recognition is performed after the writing or printing has been completed whereas in online the characters are recognized as they are drawn. In case of offline the source is either an image or a scanned form of the document whereas in online the successive points are represented as a function of time and the order of stroke [1][2]. Now a day's various commercial OCR systems are available for various applications. Some practical applications of OCRs are: data and text entry, aid for blind, automatic number-plate readers, signature verification and identification, preserving documents in electronic format, desktop publication, library cataloging etc.

2. AN OVERVIEW OF BANGLA SCRIPT

Bangla is a language of the eastern Indian subcontinent. Most of the people of West Bengal, Bangladesh, Tripura and Assam use this language. Bengali has 'Sadhubhasa' and 'Chaltibhasa' literary styles. The differences between the two styles are forms of pronouns and verb conjugations. Bangla alphabets are used to write Bengali, Assamese, Manipuri, Garo and Mundari. In Bangla, the number of characters is large and two or more characters combine to form new character shapes called compound characters. As a result, the total number of characters to be recognized is about 300. The Bengali alphabet is derived from the Brahmi alphabet. It is also closely related to the Devanagari alphabet.There are 50 basic characters (11 vowels and 39 consonants) in modern Bangla script. The basic characters of Bangla script are

Rupak Bhattacharyya et al. (Eds) : ACER 2013,
pp. 11–24, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3202

shown in Figure-1. The concept of upper/lower case is absent in this script. Bangla script is written from left to right and there is no upper/lower case in writing. Most of the characters in Bangla script have a horizontal matra line at the upper part. There may be modified shaped of a vowel depending on the position of it whether it is to the left, right (or both) or bottom of the consonant(see Figure- 2). Some vowels may take different modified shapes when attached to some consonant characters (see Figure- 3). In some cases a consonant following (proceeding) a consonant is represented by a modifier called consonant modifier (see Figure-4). There may be upper zone, middle zone and lower zone in a bangla word. The imaginary line which separates middle and lower zone is called the base line. Mostly a modified or a part of a modified character sits in the upper zone and lower zone of a line. A typical zoning is shown in Figure-5. Sometimes a consonant or vowel following a consonant forms a different shape character. This character is called compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Combination of three or four characters also exists in the Bangla script. To get an idea about Bangla compound characters some examples of compound characters formed by two and three characters are shown in Figure-6.

অ	আ	ই	ঈ	উ
ঊ	ঋ	এ	ঐ	ও
ঔ	ক	খ	গ	ঘ
ঙ	চ	ছ	জ	ঝ
ঞ	ট	ঠ	ড	ঢ
ণ	ত	থ	দ	ধ
ন	প	ফ	ব	ভ
ম	য	র	ল	শ
ষ	স	হ	ড়	ঢ়
য়	ৎ	ং	ঃ	ঁ

Figure-1. Basic characters of Bangla script.

Vowel	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
Modified Shape	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ো	ৌ
When attached to constant ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ

Figure-2. Vowel Modifiers

Constant	গ	র	শ	হ	র	হ
Vowel	ঊ	ঊ	ঊ	ঊ	ঊ	ঋ
Compound Character	গূ	রূ	শূ	হূ	রূ	হৃ

Figure-3. Exceptional cases of vowel modifiers

Constant	য	র	র
Modified Shape	্য	্র	র্
When attached to constant দ	দ্য	দ্র	র্দ

Figure-4. Consonant modifiers.

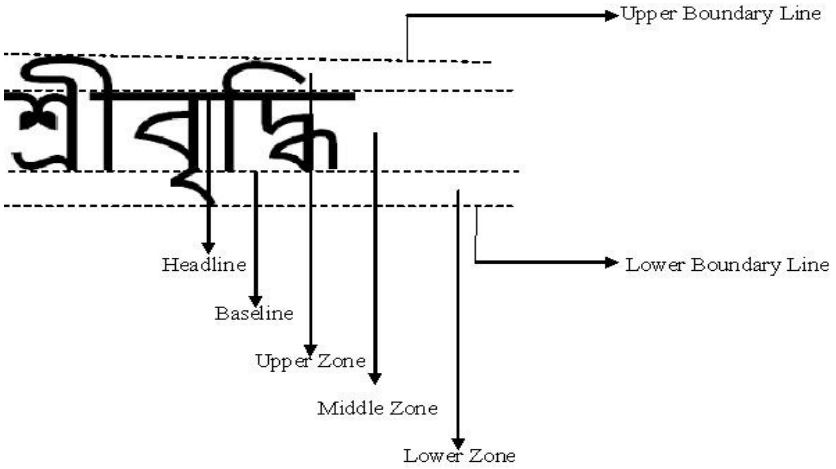


Figure-5. Various zones of a Bangla word.

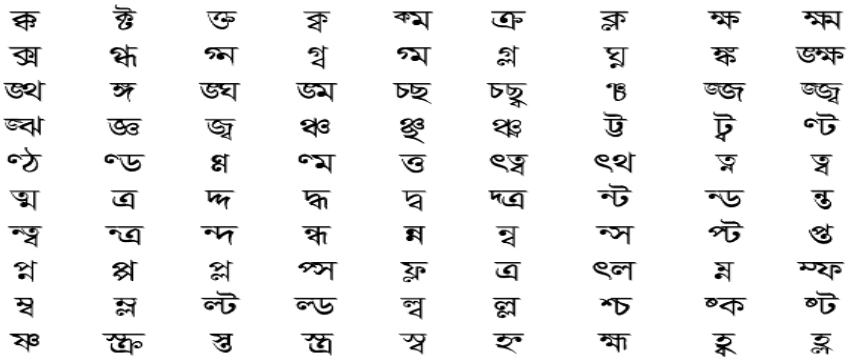


Figure-6. A set of 90 compound characters.

3. METHODOLOGY

There are various steps for developing an efficient bangla OCR system of printed bangla text. A general model of these OCR systems is shown in Figure-7. The steps used by these models are:

- Scanning } Image Acquisition

- Binarization
- Noise Detection and Removal
- Skew Detection and Correction
- Line, Word and Character Segmentation

} Preprocessing

- Feature Extraction and Selection
- Classification } Recognition

These steps can be characterized as Image Acquisition, Preprocessing and Recognition respectively.

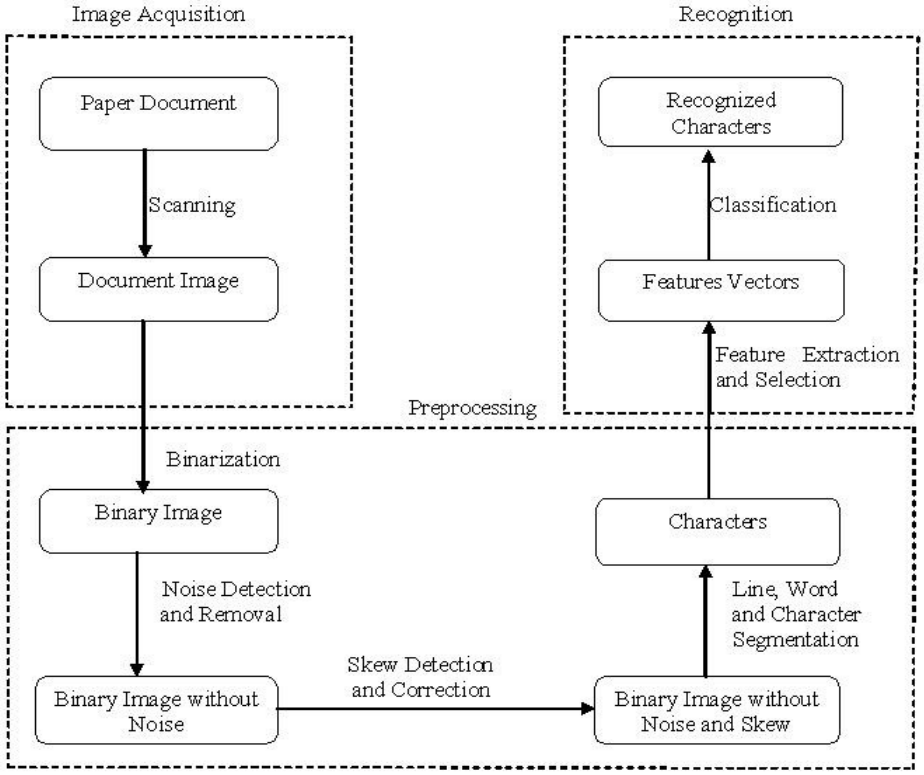
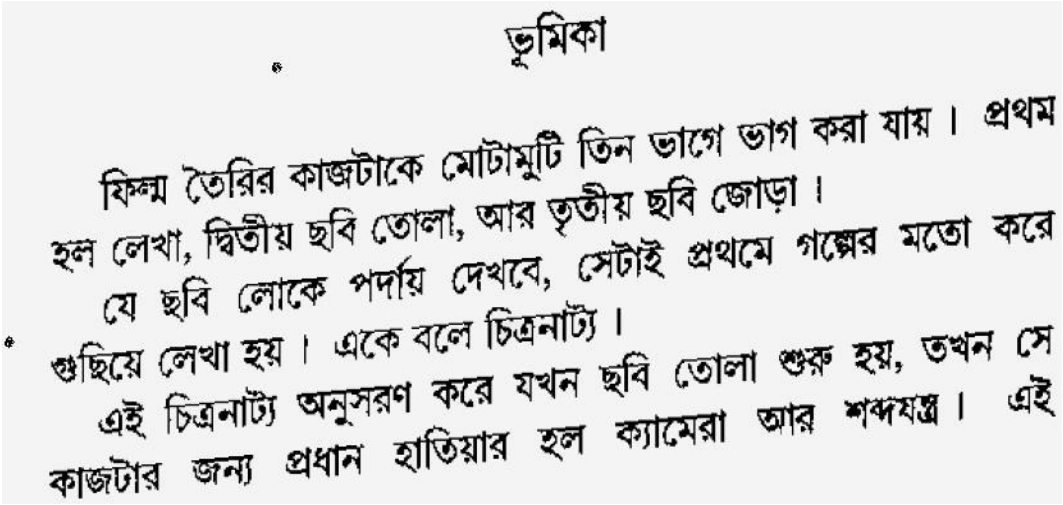


Figure -7. Common steps of an OCR system

The segmentation of character is very crucial for designing an efficient OCR system. So my present work has focused on this segmentation step of OCR system. Some existing procedures have been used for others steps. The various steps and my present work are discussed below.

3.1 Scanning

To recognize a character from a text document it is necessary to convert the document into a digital image. This task can be performed either by a Flat-bed scanner or by a hand-held scanner.


ভূমিকা

ফিল্ম তৈরির কাজটাকে মোটামুটি তিন ভাগে ভাগ করা যায়। প্রথম হল লেখা, দ্বিতীয় ছবি তোলা, আর তৃতীয় ছবি জোড়া।

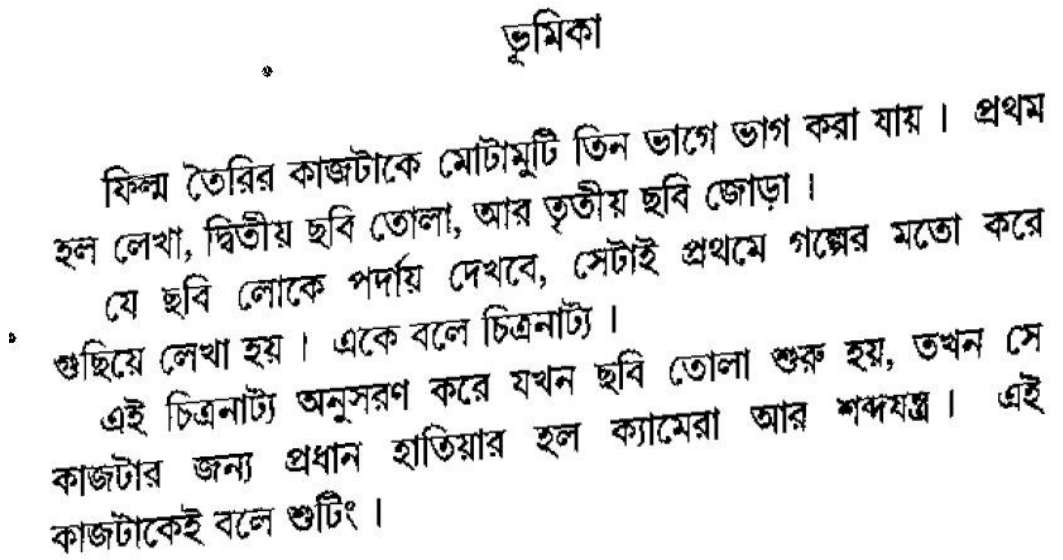
যে ছবি লোকে পর্দায় দেখবে, সেটাই প্রথমে গল্পের মতো করে গুছিয়ে লেখা হয়। একে বলে চিত্রনাট্য।

এই চিত্রনাট্য অনুসরণ করে যখন ছবি তোলা শুরু হয়, তখন সে কাজটার জন্য প্রধান হাতিয়ার হল ক্যামেরা আর শব্দযন্ত্র। এই

Figure-8. A scanned bangla document

3.2. Binarization

Binarization converts the grayscale image into a binary image. It separates the text from the background i.e. we can identify the character of the text. Binarization can happen in two ways either globally or locally. In both cases threshold intensity value is used. If the intensity value of the pixel is greater than the threshold value then it is set to white otherwise it is black. One intensity value is used for global method on the other hand multiple intensity values are used in local method. Several binarization methods are discussed in [3, 4].


ভূমিকা

ফিল্ম তৈরির কাজটাকে মোটামুটি তিন ভাগে ভাগ করা যায়। প্রথম হল লেখা, দ্বিতীয় ছবি তোলা, আর তৃতীয় ছবি জোড়া।

যে ছবি লোকে পর্দায় দেখবে, সেটাই প্রথমে গল্পের মতো করে গুছিয়ে লেখা হয়। একে বলে চিত্রনাট্য।

এই চিত্রনাট্য অনুসরণ করে যখন ছবি তোলা শুরু হয়, তখন সে কাজটার জন্য প্রধান হাতিয়ার হল ক্যামেরা আর শব্দযন্ত্র। এই কাজটাকেই বলে শুটিং।

Figure-9. The text document after binarization

3.3 Noise Detection and Removal

Noise can be produced due to printer, scanner, print quality, age of the document, etc. There are various algorithms for noise removal. But commonly used technique is low-pass filter. This filter removes as much of the noise as possible retaining the entire signal [5].

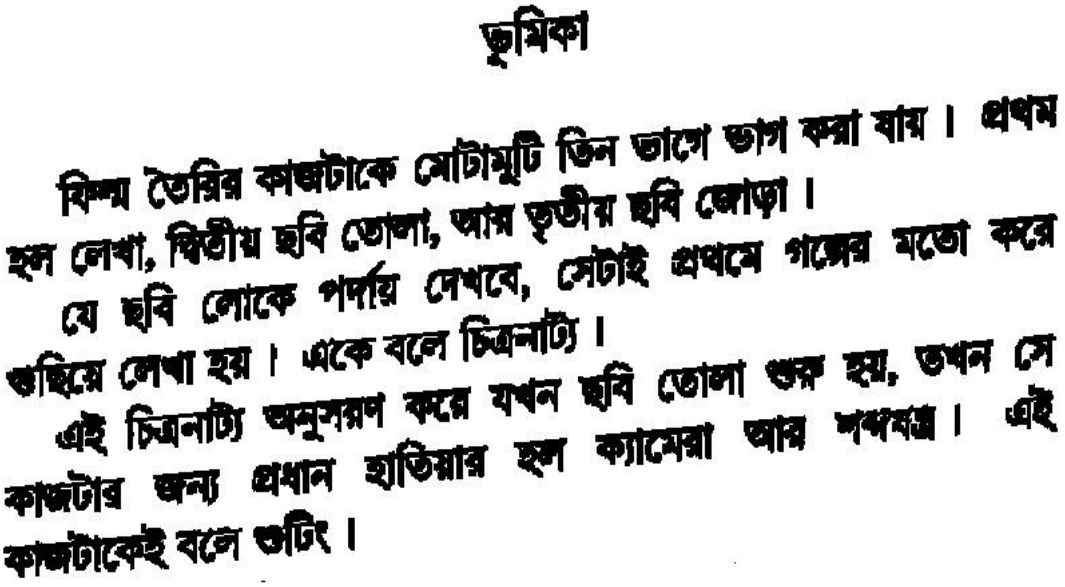


Figure-10. The text document after noise removal

3.4 Skew Detection and Correction

Printed or handwritten document may be skewed unintentionally while it is fed to the scanner. This skewness is measured by the skew angle. The skew angle is the angle of the text line with horizontal direction. Methods based on the Projection Profile, Nearest Neighbor Clustering of connected components, Hough transform and Fourier transform are used to estimate the skewed angle. In [6], different skew correction techniques have been discussed.

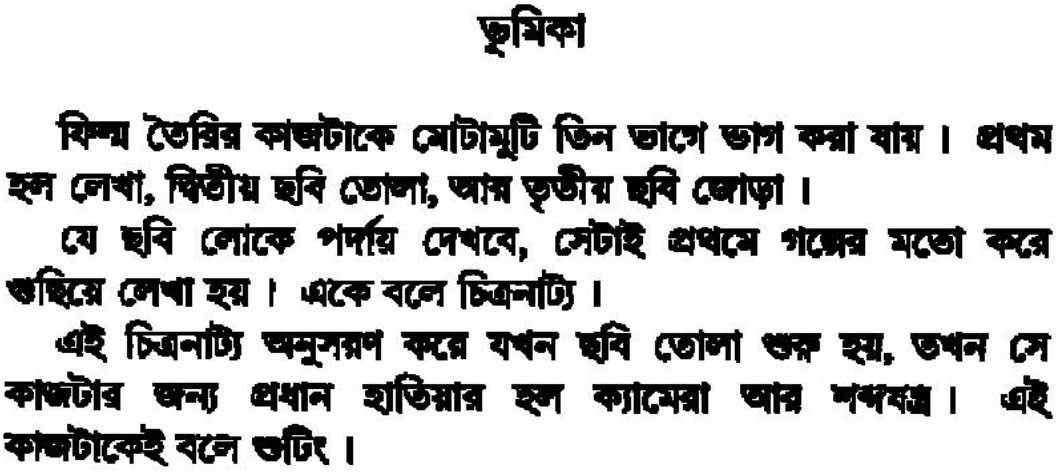


Figure-11. The text document after skew correction

3.5 Segmentation

Segmentations of line, word and character are needed for finding the individual characters. The order of these segmentations is shown below:

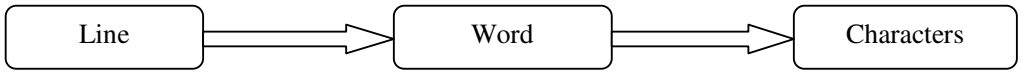


Figure-12.Order of segmentation

3.5.1. Line segmentation

Text line segmentation has been performed by scanning the input image horizontally and by keeping record of the number of black pixels in each row. Upper boundary of a line is the first row where the first black pixel is found. After finding the upper boundary, it continues scanning until a row whose next two consecutive rows have no black pixels, which is the lower boundary of the text line. It is noted that there exist more than two blank rows between two lines. The line detection process is shown in the Figure-13. And the various boundaries of the text lines are shown in Figure-14.

Algorithm: LineSegment

//This algorithm finds the lower and upper boundaries of all the lines of a printed bangla text and stores this in one-dimensional array UB and LB. The pixel values of the input image file are stored in two-dimensional array A of size HT x WD where HT and WD are the height and width of the input file.

```
Begin
        Set K=1
        For I=1 to HT by 1 do
                Set M=0
                For J=1 to WD by 1 do
                        If (A_IJ=0)
                                Set M=M+1
                        EndIf
                EndFor
                If (M=WD)
                        Set L_K = I        // L is an one-dimensional array
                        Set K = K+1
                EndIf
        EndFor
        Set B1=1
        Set B2=1
        For I=1 to K by 1 do
                If ((L_I+1-L_I) ≠1)
                        Set UB_B1=L_I
                        Set LB_B2=L_I+1
                        Set B1=B1+1
                        Set B2=B2+1
                EndIf
        EndFor
End
```

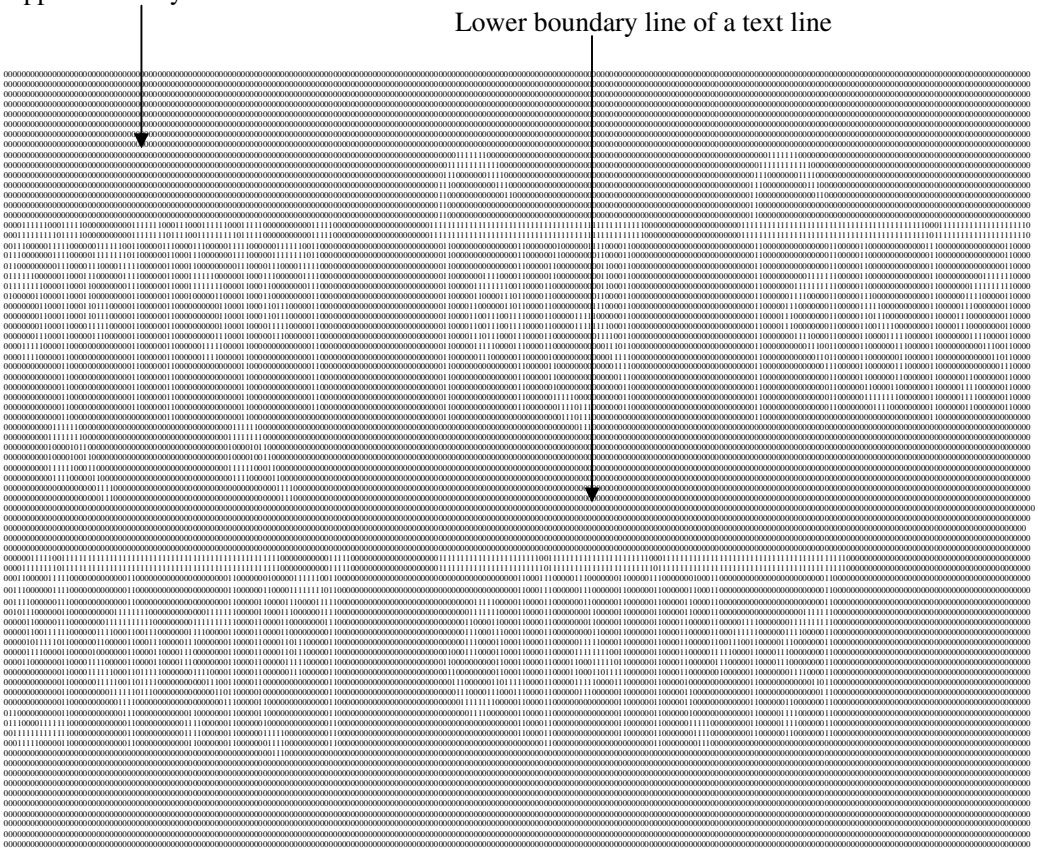


Figure-13. Boundary line detection of a text line

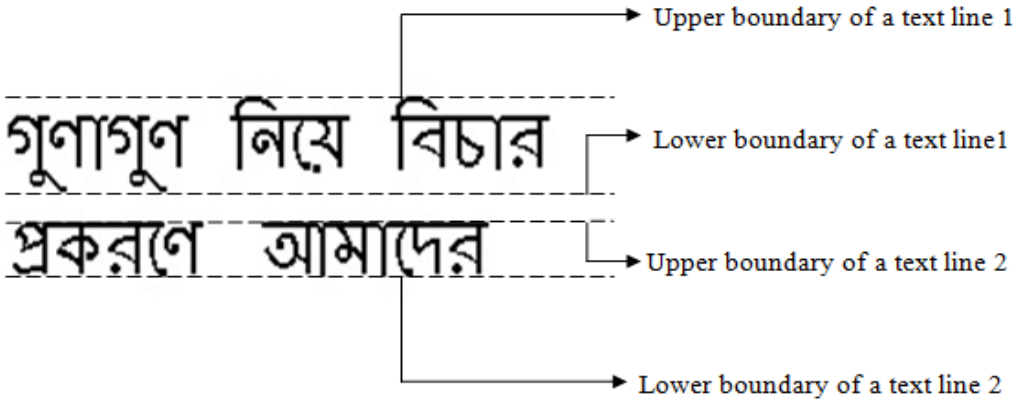


Figure-14. Boundaries of text line

3.5.2 Word segmentation

After detecting a line, the system scans the image vertically from the upper boundary line to the lower boundary line of a text line. The number of black pixels in each column is counted. Starting boundary of a word is the first column where the first black pixel is found. After finding the starting boundary, it continues scanning until a column whose next two consecutive columns have no black pixels, which is the ending boundary of the word being processed. It is noted that

there exist more than two blank columns between two words. Figiure-15 and 16 shows the word segmentation process.

Algorithm: WordSegment

// This algorithm finds the starting and ending boundaries of the words of a line. The starting and ending boundaries are stored in one-dimensional arrays SB and EB respectively.

```
Begin
        Set K=1
        For I=1 to WD by 1 do
                Set M=0
                For J=1 to (LB-UB) by 1 do
                        If (AJI=0)
                                Set M=M+1
                        EndIf
                EndFor
                If (M= (LB-UB))
                        Set WK = I        // W is an one-dimensional array
                        Set K = K+1
                EndIf
        EndFor
        Set B1=1
        Set B2=1
        For I=1 to K by 1 do
                If ((WI+1-WI) >7)
                        Set SBB1=WI
                        Set EBB2=WI+1
                        Set B1=B1+1
                        Set B2=B2+1
                EndIf
        EndFor
End
```

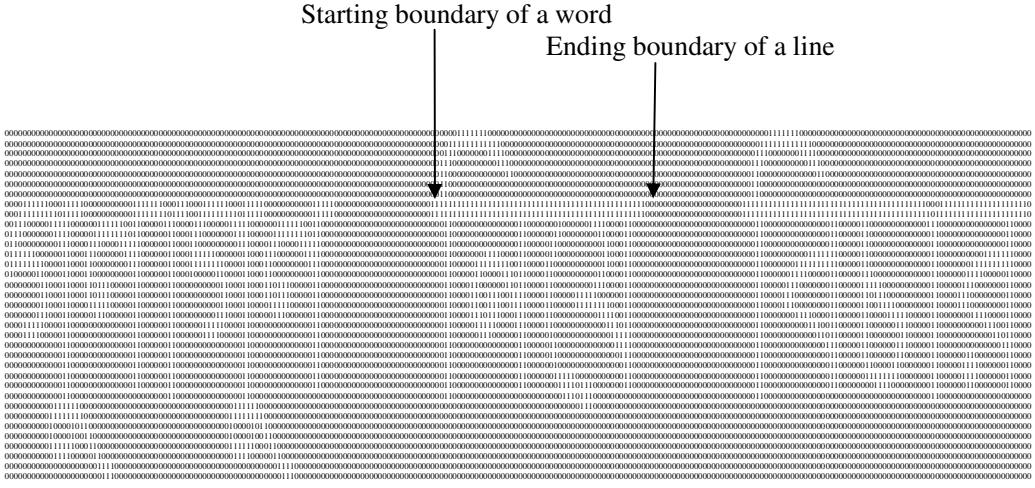


Figure-15 Boundary line detection of a word.

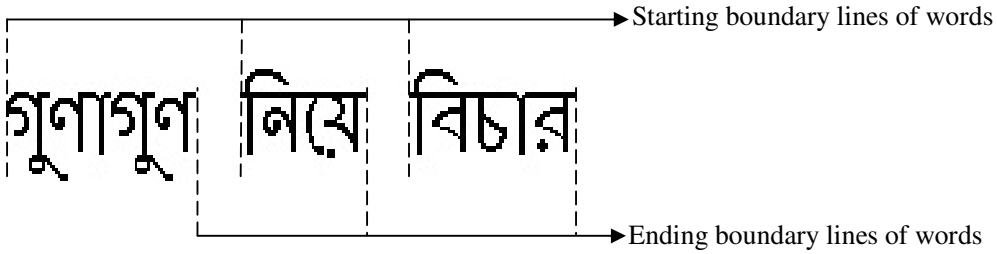


Figure-16 Boundary lines of words in a line

3.5.3 Character segmentation

To segment the individual characters in a word a vertical scan is performed from the upper boundary line of a word to the lower boundary line. If we reach the lower boundary line without facing any black pixel during scan then this column is assumed that the starting/ending boundary line of a character in the word. Vertical scanning is applicable if two consecutive characters are not connected by the Matra line. Characters in a word may be connected by a Matra line. Here Matra line is detected first then vertical scanning is applied from the row which is just below the Matra line to the lower boundary line. Both procedures are shown in the Figure-17 and 18 respectively.

Algorithm: CharacterSegment

// This algorithm finds the starting and ending boundaries of the characters of a word. The starting and ending boundaries are stored in one-dimensional arrays SBC and EBC respectively.

```
Begin
        Set ML=0
        For I=UB to LB by 1 do
                M=0
                For J=SB to EB by 1
                        If (W_IJ=1)
                                Set M=M+1
                        EndIf
                EndFor
                        If (M > (EB-SB)*0.70)
                        Set ML = ML+1
                        Set T=I+1
                EndIf
        EndFor
        If (ML=0)       // If Matra Line is not present
                Set K1=1
                For I=1 to (EB-SB) by 1 do
                        Set M=0
                        For J=1 to (LB-UB) by 1 do
                                If (A_JI=0)
                                        Set M=M+1
                                EndIf
                        EndFor
                        If (M= (LB-UB))
                                Set CB_K1 = J //CB is an one-dimensional array
                                Set K1 = K1+1
```

```
                EndIf
        EndFor
        Set B1=1
        Set B2=1
        For I=1 to K by 1 do
                If ((CB_{I+1}-CB_I) > 5)
                        Set SBC_{B1}=CB_I          // SBC is an one-dimensional array

                        Set EBC_{B2}=CB_{I+1}      // EBC is an one-dimensional array
                        Set B1=B1+1
                        Set B2=B2+1
                EndIf
        EndFor
Else
        Set K2=1

        For I=1 to EB-SB by 1 do
                Set M=0
                For J=T to (LB-T) by 1 do
                        If (A_{JI}=0)
                                Set M=M+1
                        EndIf
                EndFor
                If (M= (LB-T))
                        Set CB_{K2} = I
                        Set K2 = K2+1
                EndIf
        EndFor
        Set B3=1
        Set B4=1
        For I=1 to K2 by 1 do
                If ((CB_{I+1}-CB_I) > 5)
                        Set SBC_{B3}=CB_I
                        Set EBC_{B4}=CB_{I+1}
                        Set B3=B3+1
                        Set B4=B4+1
                EndIf
        EndFor
EndIF-Else
End
```

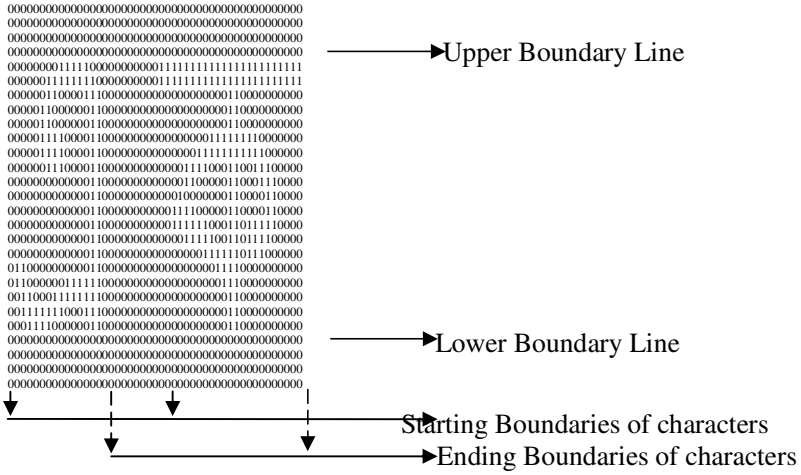


Figure-17

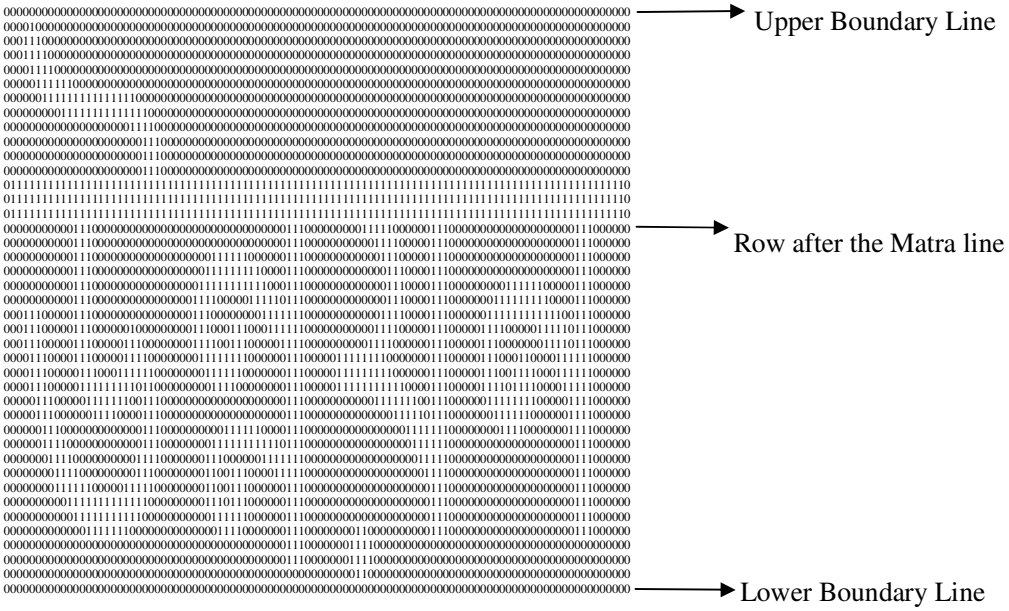
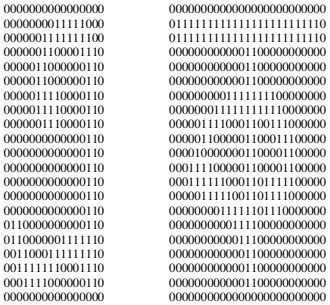
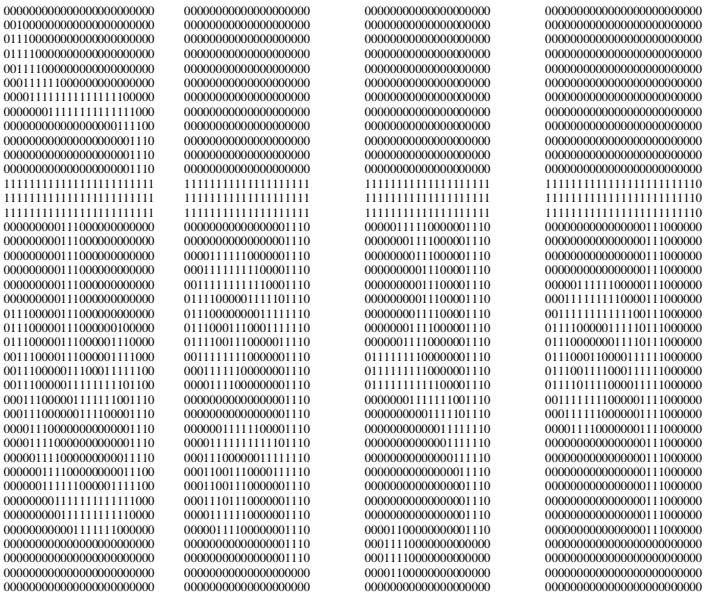


Figure-18

Now the segmented characters are:





3.6 Feature Extraction and Selection

The identification of attributes (features) that defines the shape of the characters and the choice of the right attributes for the given problem is called Feature Extraction and Selection. The following procedures are commonly used for OCR system [7]:

- Statistical
- Structural
- Hybrid

In statistical method, various samples of a pattern are used to collect the statistics during training phase. Statistical features such as zoning, crossings and projections are used for character representation. Structural method uses structural features such as strokes, holes, loops, concavities etc. In hybrid method, both statistical and structural methods are combined to represent the characters.

3.7 Classification

In this stage labels are assigned to classify the characters based on the relationship among the features extracted. The classification techniques used for this purpose may be the following types [8]:

- Neighborhood
- Statistical
- Neural Network

In neighborhood approach the neighbors of the current feature point in the features space are identified by defining a distance measure. This type of classifiers is very simple and the computation cost is less. Some classifiers of this type are Nearest Neighbor (NN), k-NN, condensed-NN, reduced-NN, edited-NN. Statistical classifiers depend on the statistics of the

ensemble of features of the reference inputs. Bayesian classifier, Support Vector Machine (SVM), Parzen Window based classifier are some examples of statistical approach. Current research on OCR focuses on Neural Network base classifier. A neural network is a computing architecture which can perform computations at a higher rate compared to the classical methods. The detailed comparison of various neural networks is in [9].

4. CONCLUSIONS AND FUTURE WORKS

In this paper the segmentation procedure of printed characters without modifiers in a bangla text has been discussed. These segmented characters are used in the recognition step of OCR development. There is a complex set of characters in the Bangla language. Sophisticated algorithms are needed for recognizing these characters. Segmentation procedure of characters with modifiers has not been discussed in this work. This work may be extended by segmenting the characters with modifiers.

REFERENCES

[1] R. Plamondon and S.N. Srihari, "Offline and Online handwritten character recognition: A comprehensive survey ," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000.

[2] N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Offline Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216-233.

[3] J. He, Q. D. M. Do*, A. C. Downton and J. H. Kim, "A Comparison of Binarization Methods for Historical Archive Documents".

[4] Tushar Patnaik, Shalu Gupta, Deepak Arya, "Comparison of Binarization Algorithm in Indian Language OCR".

[5] Rangachar Kasturi, Lawrence O'Gorman and Venu Govindaraju 2002 Document image analysis: A primer. Saadhanaa Vol. 27, Part 1, pp. 3–22.

[6] Chaudhuri B.B. and U. Pal 1997 Skew Angle Detection of Digitized Indian Script Documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 19, NO. 2,February 1997.

[7] B.Anuradha Srinibas, Arun Agarwal,C.Raghavendra Rao, " An Overview of OCR Reaserch in Indian Scripts," IJCSES, vol. 2, no.2, April 2008.

[8] R. O. Duda and P.E. Hart, Pattern classification and Scene analysis. John Wiley and Sons, 1973.

[9] M. Egmont-Peterson, D. de Ridder, H. Handels, "Image Processing with Neural Networks: A Review", Pattern Recognition, Vol 35, pp 2279-2301, 2002.

AUTHORS

Fakruddin Ali Ahmed received B.Tech.in Computer Science & Engineering from Murshidabad College of Engineering & Technology, WBUT, India in 2005 and M.E. in Software Engineering from Jadavpur University, Kolkata, India in 2009. He has more than 7 years of teaching and industry experience and currently working as an Assistant Professor in Global Institute of Management & Technology, West Bengal, India. His fields of interest are image processing and pattern recognition.