COPYRIGHT © 2010 JCIT, ISSN 2078-5828 (PRINT), ISSN 2218-5224 (ONLINE), VOLUME 01, ISSUE 01, MANUSCRIPT CODE: 100707

A Complete Bangla OCR System for Printed Chracters

Md. Mahbub Alam and Dr. M. Abul Kashem

Abstract—Bangla character recognition is very important field of research because Bangla is most popular language in the Indian subcontinent. Research on Bangla character recognition has been started since mid 1980's. Different types of techniques already applied and the performance is examined. This paper is a complete Optical Character Recognition (OCR) system for printed Bangla characters. Preprocessing steps includes binarization, noise removal, skew detection and correction, segmentation in various levels and scaling. Features are extracted from scaled character and Freeman chain code used for representing a character. Multilayer feed forward neural network is used to classify and recognize character.

Index terms: Preprocessing, segmentation, feature extraction, neural network, chain code.

1 INTRODUCTION

Optical Character Recognition began as field of research in pattern recognition, artificial intelligence and machine vision. Through academic research in the field continues, the focuses on OCR has shifted to implementation of proven techniques because of its applications potential in banks, post-offices, defense organization, license plate recognition, reading aid for the blind, library automation, language processing and multi-media system design.

Bangla is one of the most popular scripts in the world, the second most popular language in the Indian subcontinent. About 200 million people of eastern India and Bangladesh use this language, making it fourth most popular in the world. Therefore recognition of Bangla character is a special interest to us. Many works already done in this area and various strategies have been proposed by different authors.B.B. Chowdhury and U. Pal suggested "OCR in Bangla: an Indo-Bangladeshi language"[3] and also suggested a complete Bangla OCR system[9] eliciting the feature extraction process for recognition.A. Chowdury, E. Ahmmed, S. Hossain suggested a beeter approach[7] for "Optical Character Recognition of Bangla Characters using neural network", J. U. Mahmud, M.F. Raihan and C.M. Rahman provide a "A Complete OCR system[6] for Continuous characters". But we presented here a complete OCR system for printed characters.

Bangla script is in two formats, machine printed and handwritten. We are concerned with recognition of printed Bangla character. This paper helps developers to develop a complete OCR system for printed Bangla script. Section 2 we describe the properties of bangla script. In section 3 we shows the methodology of OCR sytem. In section 4 we briefly explain the techniques of preprocessing. Section 5 related with feature extraction process. In section 6 we discuss neural network for training and recognition. Result and conclusion drawn in section 7 and 8.

2 PROPERTIES OF BANGLA SCRIPT

Basic Bangla character set comprises 11 vowels, 39 consonant, 10 numerals. There are also compound characters being combination of consonant with consonant as well as consonant with vowel. A vowel following a consonant sometimes takes a modified shape and is called a vowel modifier.

Vowels A Av B C D E F G H I J
Vowel Modifiers ○v w○ ○x ○ ○ ○ †○ ‰○ ‡○v †○Š
Consonants K L M N O P Q R S T U V W X Y Z
_ ` a b c d e f g h i j k l m o
p q r s t ○
Numerals 0 1 2 3 4 5 6 7 8 9
Compound Characters ¶ ¼ ½ Á Ä ² » ß Ü ô

Many characters of Bangla script have a horizontal line at the upper part called '**matra**' or headline. A Bangla text may be partitioned into three zones.. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic characters or compound characters below the head-line and lower zone is the portion where some of the modifiers can reside. The imaginary line

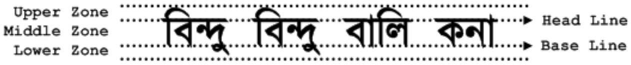


Figure 1: Illustrate the zones of Bangla Script.

- Md. Mahbub Alam is with the Department of Computer Science and Engineering, Dhaka University of Engineering and Technology[DUET] (www.duet.ac.bd),Gazipur,Bangladesh,.E-mail:emahbub.cse@gmail.com
- Dr. M. Abul Kashem is with the Department of Computer Science and Engineering, Dhaka University of Engineering and Technology[DUET] (www.duet.ac.bd),Gazipur,Bangladesh,.E-mail:drkashemll@duet.ac.bd

separating middle and lower zone is called base line. Figure 1 shows the zones of Bangla script. The concept of uppercase and lowercase is absent in Bangla script and writing style of Bangla is from left to right in a horizontal manner.

3 METHODOLOGY

Different fonts and sizes of Bangla characters used in Bangla printed document. Also a character is in bold or italic or in normal form. So it is difficult task to recognize Bangla characters and each step of the OCR system must be robust Several steps required to implement OCR system which shown in figure 2.

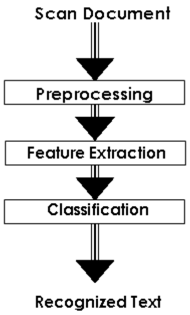


Figure 2: Illustrates OCR system

4 PREPROCESSING

Preprocessing consists of number of preliminary processing steps to make the raw data usable for the classifier. Preprocessing aims to produce data that are easy for the OCR to operate accurately. Typical preprocessing process includes the several steps shown in figure 3. namely text digitization, binarization, noise removal, skew detection and correction, segmentation in various levels and scaling.

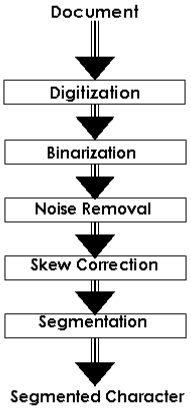


Figure 3: Illustrates preprocessing steps.

4.1 Text Digitization

The process of text digitization can be performed either by a Flat-bed scanner or a hand-held scanner. Hand held scanner typically has a low resolution range. We have used a Flatbed scanner. Scanner was manufactured by Hewlett-Packard, Model number hp-scanjet 3670 series. The digitized images are in gray tone and we have used a histogram-based threshold approach to convert them into two-tone images.

4.2 Binarization

Binarization is a technique by which the gray scale images are converted to binary images. Binarization separates the foreground (text) and background information. The most common method for binarization is to select a proper threshold for the intensity of the image and then convert all the intensity values above the threshold to one intensity value ("white"), and all intensity values below the threshold to the other chosen intensity ("black").

4.3 Noise Removal

Scanned documents often contain noise that arises due to printer, scanner, print quality, age of the document, etc. Therefore, it is necessary to filter this noise before we process the image. The commonly used approach is to low-pass filter the image and to use it for later processing. The objective in the design of a filter to reduce noise is that it should remove as much of the noise as possible while retaining the entire signal.

4.4 Skew Detection and Correction

When hard copy document is fed to the scanner carelessly, digitized image may be skewed and skew correction necessary to make text lines horizontal. Skew angle is the angle that the text lines of the document image makes with the horizontal direction. Skew correction can be achieved in two steps. First, we estimate the skew angle θ_t and second, we will rotate the image by θ_t, in the opposite direction. An approach based on the observation of head line of Bangla script used for skew detection and correction. The connected components whose bounding box width is greater than the average width of the components are selected for skew angle detection. For each such component we chose the uppermost pixels of each column of the component. Among the uppermost pixels of each component we note the pixels which satisfy the properties of digital straight line (DSL) and select those pixels which lie on the longest digital straight line. Using Hough transform on the pixels of longest digital straight line of each selected component we get the actual skew angle θ_t. Skew correction is done by rotating the document image by θ_t in the opposite direction so that the script line becomes horizontal. This approach is accurate, robust and computationally efficient.

4.5 Segmentation

Segmentation of binary image is performed in different levels includes line segmentation, word segmentation, character segmentation.

4.5.1 Line Segmentation

Text line detection has been performed by scanning the input image horizontally which. Frequency of black pixels in each row is counted in order to construct the row histogram .The position between two consecutive lines, where the number of pixels in a row is zero denotes a boundary between the lines. Line segmentation process shown in figure4.

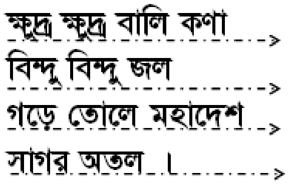


Figure 4: Line Segmentation

4.5.2 Word Segmentation

After a line has been detected, each line is scanned vertically for word segmentation. Number of black pixels in each column is calculated to construct column histogram. The portion of the line with continuous black pixels is considered to be a word in that line. If no black pixel is found in some vertical scan that is considered as the spacing between words. Thus different words in different lines are separated. So the image file can now be considered as a collection of words. Figure5 shows the word segmentation process.

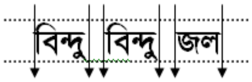


Figure 5: Word Segmentation

4.5.3 Character Segmentation

Zones of Bangla script: From figure 1 we see that Bangla text may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic characters or compound below the head-line and lower zone is the portion where some of the modifiers can reside. The imaginary line separating middle and lower zone is called base line.

Detection of Matra: To segment the individual character from the segmented word, we first need to find out the headline of the word which is called 'Matra'. From the word, a row histogram is constructed by counting frequency of each row in the word. The row with highest frequency value indicates the headline. Sometimes there are consecutive two or more rows with almost same frequency value. In that case, 'Matra' row is not a single row. Rather all rows that are consecutive to the highest frequency row and have frequency very close to that row constitute the matra which is now thick headline.

Detection of character between baseline and headline: At first we remove the headline, so the characters in a word are isolated and can easily be separated.

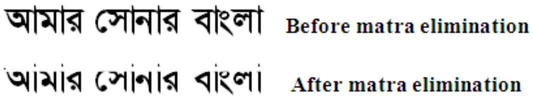


Figure 6: Matra Elimination

To find the demarcation line between characters a vertical scan is initiated from the row that is just beneath the 'Matra' row. If during scan, one can reach the base line without touching any black pixel then this scan successfully found a demarcation line between characters.But only linear vertical scan fails to find the demarcation line between characters, which illustrates below:

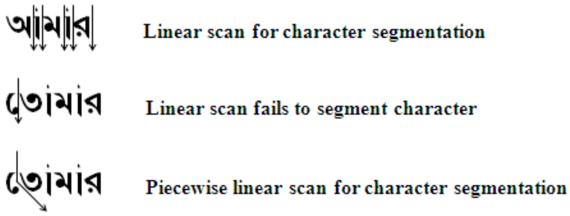

আমার
Linear scan for character segmentation
তোমার
Linear scan fails to segment character
তোমার
Piecewise linear scan for character segmentation

Figure 7: Various Scans for character segmentation.

To overcome the problem of vertical scan, we have used an approach called piecewise linear [6] in the sense that the scan takes turns whenever it sees black pixel and tries to reach the baseline.

Detection above 'Matra': To find the portion of any character above the 'Matra', then we can move upward from the 'Matra' row from a point just adjacent to the 'Matra' row and between the two demarcation lines. If it is, then a greedy search is initiated from that point and the whole character is found.

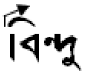


Figure 8: Greedy search for finding the portion of the character above the Matra.

Detection below the baseline: To segment the characters below another character, baseline of the segmented word has been calculated. Each word can be considered to have an imaginary line that crosses at the middle of the word called 'baseline'. A greedy search is initiated for the presence of black pixels below the baseline, which will result some connected components below that baseline. All the components below the baseline contain lowest point called 'Base point'. Baseline is highest frequency row of those points. After determining the baseline, a depth first search (DFS) easily extracts the characters below the baseline.

Figure 9: Extracting characters below the baseline

4.6 Scaling

Different size character is used in Bangla script. Before extracting the feature of the character, we need to uniform the size of each character for better result. i.e. we need to perform scaling so that size invariant character recognition can be possible. Our system has been taken character ranges from 8 pt to 24 pt. we convert the character into standard size which is 44×44 size of character for our system.

5 Feature Extraction

Feature extraction provides an important role in character recognition. It is the most challenging task to recognition a character but choice of good features significantly improves the recognition rate and minimizes the error in case of noise. In feature extraction stage each character is represented as a feature vector. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. Various steps[6] required to extract features are discussed below:

5.1 Connected Componetn Extraction

First we extract all the connected component in a character. Because in Bangla language, a character can have more than one connected component. Recognition of connected component is important to achieve desired result. Therefore all the connected components are detected from the isolated character. Depth First Search (DFS) approach is used for the detection of connected component

5.2 Center of Mass for Each Component (Centroid)

Center of mass has been calculated for each connected component. Center of mass for i^{th} connected component is (X_i, Yi).

Here,

Where,

$$X_i = \sum_{j=1}^{Ni} p_{ij} / N_i \quad \text{.......................................} \quad (1)$$

$$Y_i = \sum_{j=1}^{Ni} Q_{ij} / N_i \quad \text{.......................................} \quad (2)$$

N_i = Number of black pixels in connected component i.

P_{ij} = x coordinate of the j^{th} black pixel in the i^{th} connected component

Q_{ij} = y coordinate of the j^{th} black pixel in the i^{th} connected component

5.3 Bounded Rectangle Calculation

To divide the connected component into regions, we need to calculate the bounded rectangle [6] of the connected component. If the grid is searched in a row wise manner and connected component is found using Depth First Search (DFS) strategy, top left and bottom right coordinate of a component can be found. Besides its minimum and maximum span in its x direction as well as y direction can be found which results a bounded rectangle of the component.

5.4 Division of the Component into Regions

After calculating the boundary rectangle of the connected component, each connected component has been divided into four regions indicating four quadrants in 2-D geometric system. The origin is the center of mass of that connected component. With the origin and the bounded rectangle of the connected component, four regions can be established.

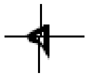


Figure 10: Four regions for a connected component.

5.5 Chain Code Generation

Chain code was introduced by Freeman [5] as a means to represent lines or boundaries of shapes by a connected sequence of straight line segments of specified length and direction. A chain code has two components: the coordinates of the starting point and a chain of codes representing the relative position of the starting pixel and its followers. The chain code is generated by using the changing direction of the connected pixels contained in a boundary. The representation is based on 4-connectivity or 8- connectivity of the segments. After the character has been divided into connected components and boundary of the connected components are established, chain code is to be calculated. Freeman chain code is based on the observation that each pixel has eight neibourhood pixel is shown in figure 11.

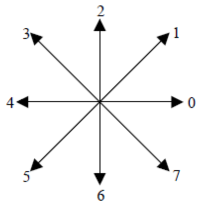
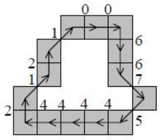


Figure 11: Slope Convention for Freeman Chain Code

For a closed boundary, its chain code obviously depends on the starting point of the boundary. To make it invariant to the starting point, the chain code can be normalized according to the following method [1]: A chain code can be treated as a circular sequence of direction numbers. Therefore, the starting point can be redefined so that the resulting sequence of numbers forms a minimum integer (refer to the example shown by Figure 12).



Chain Code: 21210066754444
Normalized Chain Code: 00667544442121

Figure 12: 8-directional chain code example.

5.6 Slope Distribution Generation

When searching for a closed contour continues, there is a variation of slop in each region. The frequency of each directional slope at each region is recorded and updated during traversal. There are eight directional slopes in a region, therefore total 32 directional slope[6] for the whole component. The frequency of jth directional slope at ith region is local feature S_{ij}, where j = 0 1 . . . , 7 and i = 0,1,2,3.

5.7 Normalized Slope Calculation

In order to obtain fractional value, feature values must be normalized [6] to (0-1) scale. The rule for normalization is: If $\overline{a_1}, \overline{a_2}, \overline{a_3}, \ldots \overline{a_n}$ are n feature vectors in n dimensional feature space, then their normalized values are $a_1, a_2, a_3, \ldots a_n$.

Here,

$$\overline{a_1} = a_1 / N$$
$$\overline{a_2} = a_2 / N$$
$$\ldots\ldots\ldots$$
$$\overline{a_n} = a_n / N$$

and $N=\sqrt{(a_1^2 + a_1^2 + \ldots + a_n^2)}$

5.8 Conversion to Character Slope Distribution

If there is a more than connected component in the character, then 32 normalized slopes for each connected component will be found after the previous step. But recognition step recognizes the whole character, not its individual connected component therefore normalized feature for each connected components are averaged to get the total features for the character[6].

6 Classification and Recognition

A neural network [12] is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. First knowledge is acquired by the network through a learning process and then storing knowledge is for recognition of character. We use feed forward neural network for the classification and recognition Bangla characters.

6.1 Training

We trained the neural network by normalized feature vector obtained for each character in the training set. Four layer neural networks has been used with two hidden layers for improving the classification capability. For 32 dimensional feature vectors and 4 layers, number of neuron used in the first hidden layer is 80 and that in the second hidden layer is 65.Output of the neuron is 50 for each character**.**

6.2 Recognition

Character is recognized using stored knowledge of the network. Use of two hidden layers increases the recognition rate significantly.

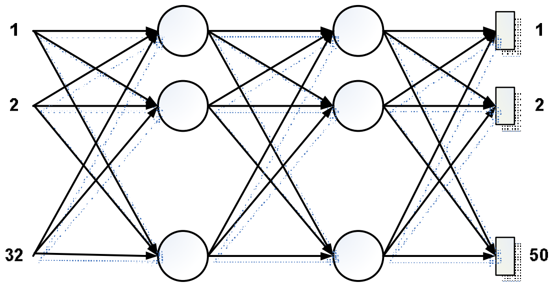


Figure 13: A neural network with 4 layers, 80 neurons and 65 neurons in hidden layers.

7 EXPERIMENTAL RESULT

We use many samples with different fonts and sizes for training and recognition purpose. We use Matlab 7.0 for our experiment. The fonts used in our experiment include sutonny, sulekha, sunetra. If the font size is large then we get good result. From our experiment we see that an accuracy about 97% possible to obtain.

8 CONCLUSION

An approach for the recognition of Bangla scripts presented in this paper. Approach suggested from the beginning of scanning a document to converting it to binary image, noise removal, skew detection and correction, line segmentation and word segmentation and character segmentation and scaling. It is the challenging task in the character segmentation part when two characters are sometimes joined together. Good Performance of the OCR system depends on good feature extraction of character which is more challenging task. We propose chain code for image representation; if we implement this properly then we think we get good result. We use multilayer feed forward neural network for classify and recognition of character. We think that if we train the network using different fonts and images with distorted character and with good shaped character then recognition performance will increases.

References

[1] R. Gonzalez and R. E. Woods, Digital Image Processing, Prentice Hall, 2002.

[2] S. Rajasekaran and G.A. Pai, Neural Networks, Fuzzy Logic, and Genetic Algorithms, 2003

[3] U. Pal, B. B. Chaudhuri, "OCR in Bangla: an Indo-Bangladeshi Language", Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image Processing, Proceedings of the 12th IAPR International, 1994.

[4] B. B. Chaudhuri, U. Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari(Hindi)", IEEE Computer Society, 1997.

[5] H. Freeman, “Computer processing of line drawing images,” Computer Survey, Vol.6, pp.57-97, 1974.

[6] 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.

[7] Ahmed Asif Chowdhury, Ejaj Ahmed, Shameem Ahmed, Shohrab Hossain and Chowdhury Mofizur Rahman"Optical Character Recognition of Bangla Characters using neural network: A better approach". 2nd International Conference on Electrical Engineering (ICEE 2002), Khulna, Bangladesh

[8] Md. Abul Hasnat, S. M. Murtoza Habib, and Mumit Khan, Segmentation free Bangla OCR using HMM: Training and Recognition, Proc. of 1st DCCA2007, Irbid, Jordan, 2007

[9] B.B Chaodhuri and U. Pal, “A Complete Printed Bangla OCR System”, Pattern Recognition Vol-31, 531-549, 1997.

[10] Ahsan Amin, Monower Syed, Optical Bangla Digit Recognition, Dhaka University, 2003

[11] Rudra Pratap, Matlab 7, 2006

[12] S. Haykin, Neural Networks, A Comprehensive Foundation (2nd Edition)

[13] http://en.wikipedia.org

[14] http://www.mathworks.com

Md. Mahbub Alam has been serving as a Lecturer of Department of Computer Science and Engineering (CSE), Dhaka University of Enigineering and Technology (DUET).Gazipur, Bangladesh.

Dr. M. Abul Kashem has been serving as an Associate Professor and Head of Department of Computer Science and Engineering (CSE), Dhaka University of Enigineering and Technology (DUET).Gazipur, Bangladesh.