The Anatomy of Bangla OCR System for Printed Texts using Back Propagation Neural Network

Shamim Ahmed[1] and A.K.M. Najmus Sakib[2]

[1] Dhaka International University, Dhaka, Bangladesh

Received: 10 April 2012 Accepted: 2 May 2012 Published: 15 May 2012

Abstract

This paper is based on Bangla (National Language of Bangladesh) Optical Character Recognition process for printed texts and its steps using Back Propagation Neural Network. Bangla character recognition is very important field of research because Bangla is most popular language in the Indian subcontinent. Pre-processing steps that follows are Image Acquisition, binarization, background removal, noise elimination, skew angle detection and correction, noise removal, line, word and character segmentations. In the post processing steps various features are extracted by applying DCT (Discrete Cosine Transform) from segmented characters. The segmented characters are then fed into a three layer feed forward Back Propagation Neural Network for training. Finally this network is used to recognize printed Bangla scripts.

Index terms— Optical Character Recognition, Binarization, Skew Angle Detection, Segmentation, Artificial Neural Network.

utomatic processing and analysis of document images is rapidly becoming one of the most important fields in pattern recognition and machine vision applications. In recent years there has been a trend to the formalization of a methodology for recognizing the structures of various types of documents in the framework of document understanding_since the whole process of document understanding is too complex to be covered by a single specialized approach. Other fields that are closely related to this relatively new applied field are the development of standard databases, compression and decompression techniques, cross validation, image filtering and noise removal, fast information retrieval systems, document segmentation and, above all, recognition of alpha, numeric characters.

All these fields are closely interrelated. Numerous research works has been done on the Roman character set and very efficient character recognition systems are now commercially available. Much effort has also been made to recognize Chinese characters because of the fact that scientist visualized the task of Chinese character recognition as the ultimate goal in character recognition. Unfortunately, very few efforts have been made so far to recognize the characters commonly found in the Indian sub-continent. This paper presents an approach to the formation of a complete character recognition system to recognize hand-printed Bengali characters.

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" (Pal U., Chaudhuri B. B., 1994) and also suggested a complete Bangla OCR system (Rowley H., Baluja S. and Kanade T., 1998) eliciting the feature extraction process for recognition.A. Chowdury, E. Ahmmed, S. Hossain suggested a beeter approach

(Ahmed Asif Chowdhury, Ejaj Ahmed, Shameem Ahmed, Shohrab Hossain and Chowdhury Mofizur Rahman, ICEE 2002) 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 for Continuous characters".

Optical Character Recognition (often abbreviated as OCR) involves reading text from paper and translating the images into a form (say ASCII codes) that the computer can manipulate. Although there has been a significant number of improvements in languages such as English, but recognition of Bengali scripts is still in its preliminary level. This thesis tries to analyze the neural network approach for Bangla Optical Character Recognition. A feed forward network has been used for the recognition process and a back propagation algorithm had been used for training the net. Before the training, some preprocessing steps were involved of course. Preprocessing includes translating scanned image into binary image, skew detection & correction, noise removal, followed by line, word and character separation. Translation of scanned image into binary image, skew detection & correction, noise removal, line and word separation of the pre-processing steps and feature extraction, recognition and classification, and, various post processing steps and learning sections were analyzed in this paper.

Bangla is an eastern Indo-Aryan language and evolved from Sanskrit (Barbara F. Grimes, 1997). The direction of the writing policy is left to right. Bangla language consists of 50 basic characters including 11 vowels and 39 consonant characters and 10 numerals. In Bangla, the concept of upper case or lower case letter is not present. Bangla basic characters have characteristics that differ from other languages. Bangla character has headline which is called matraline or matra in Bangla. It is a horizontal line and always situated at the upper portion of the character. Among basic characters, there are 8 characters which are with half matra, 10 characters with no matra and rest of them with full matra. Most of consonants are used as the starting character of a word whereas, vowels are used everywhere. Vowels and consonants have their modified shapes called vowel modifiers and consonant modifiers respectively. Both types of modifiers are used only with consonant characters. There are 10 vowels and 3 consonant modifiers which are used before or after a consonant character, or at the upper or lower portion of a consonant character or on the both sides of a consonant character. In Bangla, some special characters are there which are formed by combining two or more consonants and acts as an individual character. These types of characters are known as compound characters. The compound characters may further be classified as touching characters and fused characters. Two characters placed adjacent contact to each other produce a touching character. Touches occur due to horizontal placement of only two characters and/or vertical placement of two or more characters.

About 10 touching characters are there in Bangla. Fused characters are formed with more than one basic character. Unlike touching characters, the basic characters lose their original shapes fully or partly. A new shape is used for the fused characters. In sum, there are about 250 special characters in Bangla except basic and modified characters. The occurrence of vowels and consonants are larger compared to special characters in most of the Bangla documents. A statistical analysis, we took 2 sets of data populated with 100,000 words from Bangla books, newspapers and 60,000 words from Bangla dictionary respectively (B.B. Chaudhuri, 1998).

1 a) Image Acquisition

Image Acquisition is the first steps of digital processing. Image Acquisition is the process of capture the digital image of Bangla script through scanning a paper or book containing Bangla script. Generally the scanning image is true color (RGB image) and this has to be converted into a binary image, based on a threshold value. **??**d., 1993). We used thresholding technique for differentiating the Bangla script pixels from the background pixels. Most of the Bangla character has headline (matra) and so the skew angle can be detected using this matra. In Bangla, head line connects almost all characters in a word; therefore we can detect a word by the method of connected component labeling. As mentioned in (Schneiderman H., 2003), for skew angle detection, at first the connected component labeling is done. 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 ?t and second, we will rotate the image by ?t, in the opposite direction. An approach based on the observation of head line of Bangla script used for skew detection and correction.

2 March e) Segmentation

Segmentation of binary image is performed in different levels includes line segmentation, word segmentation, character segmentation. We have studied several segmentation approaches. From implementation perspective we observed that, most of the errors occurred at character level segmentation. Line and word level segmentation failed due to the presence of noise which gives wrong estimation of the histogram projection profile. However character level segmentation mostly suffers from joining error (fail to establish a boundary where there should be one) and splitting error (mistakenly introduce a boundary where there should not be one). Considering all these we made our effort up to a minimal segmentation and we resolved these issues during classification. Finally we used a simple technique similar to (Yang and Huang 1994).

3 f) 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 figure 6. 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. Figure 7 shows the word segmentation process. 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. a) Segmented Image to Feature Calculation Here I assume that I have already got the segmented image that can be either a character or a word and the image is already converted to binary image. Let take a segmented character and a segmented word which is shown in Figure 9. Now from these images number of frame will be calculated. In our approach we choose the frame width to be 8 and the frame height to be 90. The frame width and height is chosen according to our statistical analysis. Based on the frame width and height we divide the segmented image into several frames. The size of mean and variance vector is also determined from the frame width and height. For example the number of frame of the segmented character tao is 3 and segmented word mitu has 6 frames. Number of frame is most important because it determines the number of states for learning model. So we can say that the March number of states for learning model tao is 3 and mitu has 6 states. The above discussion is illustrated in Figure 10 and Figure 11. 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.

4 g) Pixel Grabbing from Image

As we are considering binary image and we also fixed the image size, so we can easily get 250 X 250 pixels from a particular image containing Bangla character or word. One thing is clear that we can grab and separate only character portion from the digital image. In specific, we took a Bangla character contained image. And obviously it's a binary image. As we specified that the pixel containing value 1 is a white spot and 0 for a black one, so naturally the 0 portioned spots are the original character. h) Finding Probability of Making Square Now we are going to sample the entire image into a specified portion so that we can getthe vector easily. We specified an area of 25 X 25 pixels. For this we need to convert the 250 X 250 image into the 25 X 25 area. So for each sampled area we need to take 10 X 10 pixels from binary image.

5 March

The presence of a matra is manifested by a horizontal line on the upper part of the character symbol. It is stipulated that the presence of a horizontal or nearly horizontal line with a continuous or almost continuous pixel proximity would be an ideal candidate to be identified as a matra. But this is not the only consideration. Depending on the writing style, the position of the matra within the symbol with respect to the base line may vary a lot. It is assumed that to be a candidate for a matra, it must be found in the upper portion of the symbol. More specifically, while developing the matra detection algorithm, it has been assumed that it should be found within one third of the total height from top most row of pixels containing a valid symbol presence. In the actual implementation, the total number of pixels were calculated and the rows having a valid "ON" pixel were detected. Dividing the total number of pixels present within the image by the total number of rows containing those pixels, the statistical average of the number of pixels per line was calculated. It has been further assumed that the matra should contain at least twice the number of valid pixels with respect to the statistical average number of pixels calculated on the whole image.

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.

6 i) 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. As we are considering binary image and we also fixed the image size, so we can easily get 250 X 250 pixels from a particular image containing Bangla character or word. One thing is clear that we can grab and separate only character portion from the digital image. In specific, we took a Bangla character contained image. And obviously it's a binary image. As we specified that

the pixel containing value 1 is a white spot and 0 for a black one, so naturally the 0 portioned spots are the original character. b) Finding Probability of Making Square Now we are going to sample the entire image into a specified portion so that we can getthe vector easily. We specified an area of 25 X 25 pixels. For this we need to convert the 250 X 250 image into the 25 X 25 area. So for each sampled area we need to take 10 X 10 pixels from binary image.

7 c) Mapped To Sampled Area

The same sample pixel from binary image after separating, we will find out for each 5 X 5 pixel from the separated pixel portion and give an unique number for each separated pixel class. And this number will be equal to the 5 X 3 sampled areas. Now we need no consider whether 5 X 5 pixels will make a black area or square or a white area or square. We will take the priority of 0s or 1s from 5 X 5 pixels. And from there we can say, if the 0s get the priority from 5X5 in ith location then we will make a black square on ith position of sample area.

8 March

Here is an example of how a 250 X 250 pixels of Bangla character is sampled into 25 X 25 sampled area. This stage describes the training and recognition methodology. The extracted features for each segmented character are considered as the input for this stage. However we did not limit ourselves on several issues like training from multiple samples and also the trained data representation using a fixed prototype model. We introduced the concept of dynamic training at any level of recognition and dynamic prototyping as well. For the recognition process we create a temporary model from the feature file of each character image and simply pass the model to the recognizer (Back Propagation Neural Network) for classification. For classification purpose we use multilayer feed forward neural network. This class of networks consists of multiple layers of computational units, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. In many applications the units of these networks apply a sigmoid function as an activation function. Multi-layer networks use a variety of learning techniques, the most popular being back propagation. In our training data set initially we considered only the alphabets of Bangla character set with the traditional segmentation method, but the recognition performance was not considerable. Then we added the compound characters into the training set and we obtain a good performance. However with this database the system was yet suffering from segmentation error occurred at the places of the vowel and consonant modifiers. So, finally we have taken the minimal March segmentation approach (Angela Jarvis) and added the characters with the vowel and consonant modifiers into the training set. During training, we must associate the appropriate Unicode character in the same order as they appear in the image.

9 c) Back Propagation Neural Networks Algorithm

A typical back propagation network with Multilayer, feed-forward supervised learning network. Here learning process in Back propagation requires pairs of input and target vectors. The output vector 'is compared with target vector. In case of difference of output vector and target vector, the weights are adjusted to minimize the difference. Initially random weights and thresholds are assigned to the network. These weights are updated every iteration in order to minimize the mean square error between the output vector and the target vector.

10 Weight Initialization

Set all weights and node threshold to small random numbers. Note that the node threshold is the negative of the weight from the bias unit (whose activation level is fixed at 1).

Calculation of Activation 1. The activation level of an input unit is determined by the instance presented to the network. 2. The activation level O j of a hidden layer and output unit is determined by theO j = F (? W ji O i -?"¨ j)

Where W ji is the weight from an input O i , ?"¨ j is the node threshold, and F is a sigmoid function : Where W ji (t) is the weight from unit I to unit j at time t (or t th iteration) and Î?"W ji is the weight adjustment.

11 The weight change is computed by

Î?"W ji = ?? j O i

Where ? is a trial independent learning rate ($0 < ? < 1$) and ?j is the error gradient at unit j. Convergence is sometimes faster by adding a momentum term:W ji (t + 1) = W ji (t) + ?? j O i + ?[W ji (t) -W ji (t -1)]

Where $0 < ? < 1$.

1. The error gradient is given by : For the output units:Î?"j = O j (1-O j) (T j -O j)

Where Tj is the desired (target) output activation and Oj is the actual output activation at output unit j. For the hidden units:Î?"j = O j (1 -O j) ? ? k W k

Where ? k is the error gradient at unit k to which a connection points rom hidden unit j. 1. Repeat iterations until convergence in terms of the selected error criterion. 2. An iteration includes presenting an instance, calculating activations, and modifying weights.

12 d) Performance Analysis

In our approach the performance of the recognizer depends on the number of trained characters and words. Usually the recognizer does not give any transcription as output if the ANN model for the character or word to be recognized not likely to the trained models of the system. In some cases the recognizer give wrong output when the ANN model to be recognized not trained previously and there exists a similar type model in the system. In such case ANN output a transcription to which the model is most likely that means when the score of the model exceeds the threshold value. So we can say that the recognizer produce maximum performance when the system is trained with a large training corpus. Here we start with an example that shows the performance measurement of the recognizer. The test image to be recognized is shown in Figure **??** 15. Approaches suggested from the beginning of scanning a document to converting it to binary image, skew detection and correction, line separation, word segmentation, and character segmentation has been successfully stated. One of the challenges faced in the character segmentation part is that two characters are March sometimes joined together. There are even cases where a single character breaks apart. Solutions to these challenges are likely to be presented in future. Good Performance of the OCR system depends on good feature extraction of character which is more challenging task. In our current approach, the whole character itself was used as a feature. In future implementation feature extraction will be more comprehensive. As I said we are at the preliminary level of the Bangla Character Recognition so the main drawback we can consider is we need to modify and make it more accurate. Again like all other Neural Network training time increase with the increase in number of characters or words in Back Propagation Neural Network.

Extracting high level information in the form of a priori knowledge is now considered to be a very important aspect of practical character recognizer design. It is hoped that successful application of the information extracted from the database in the form of high level feature detection will help in future recognizer design especially in the case of printed Bengali character recognition. The results obtained should be considered to be indicative rather than conclusive because of the very small size of the character database. When tested on the train dataset, the system produces a 100% recognition rate, but as completely unseen samples are tested, the recognition was up to 97.5%. Discussions about the possible improvement of the system in future have also been incorporated.

The efficiency can be increased by using better scanner and camera, better technique of scaling, efficient technique of matra detection and feature extraction of the Bangla character image. Future work includes the expansion of the system to include a wider range of rotations and illumination conditions. Extension of segmented frame and illumination invariance would involve training on synthetic images over a larger range of views and conditions. Another area of improvement is the accuracy in character detection, which was not explored in depth in this thesis. Bangla character detection accuracy was improved by using a more sophisticated geometrical model for the positions of the components along with more carefylly selected negative training data. [1]

1

Figure 1: Fig. 1 .

2 I.

Figure 2: Fig. 2 .

3 I

Figure 3: Fig. 3 .

4 NTRODUCTION

Figure 4: Fig. 4 .

2

[1] © 2012 Global Journals Inc. (US)
[2] © 2012 Global Journals Inc. (US)

5 

Figure 5: Fig. 5 .

6 

Figure 6: Fig. 6 .

7 

Figure 7: Fig. 7 .

9 

Figure 8: Fig. 9 .

10 

Figure 9: Fig. 10 .

12 

Figure 10: Fig. 12 .

13 

Figure 11: Fig. 13 .

14 

Figure 12: Fig. 14 .

15 

Figure 13: Fig. 15 .

16 

Figure 14: Fig. 16 .

Figure 15: F

Figure 16: Fig.

1

Vowels
Consonants
Vowel
Modifiers
Vowel
Modifiers
attached with
consonants
Consonant
Modifiers
Consonant
Modifiers
attached
with
consonants
Compound
Characters:
Horizontal
Touching
Characters
Compound
Characters:
Vertical
Touching
Characters
Compound
Characters:
Fused
Characters

[Note: NumeralsTable.1. Different types of Bangla characters. A subset of 112 compound characters out of about 250 characters (B.B. Chaudhuri, 1998) is shown here]

Figure 17: Table 1 :

Word	the recognizer Model Name	Unicode Sequence
????	h0800	0986, 09AE, 09BE, 09B0
??????	h0801	09B8, 09CB, 09A8, 09BE, 09B0
??	h0802	09AC, 09BE
?????	h0803	0982, 09B2, 09BE

Figure 18:

.

Figure 19: Table . 2

[Lawrence et al. ()] , S Lawrence , C L Giles , A C Tsoi , A D Back . *IEEE Transactions of Neural Networks* 1993. 8 (1) p. .

[Duda et al. ()] , R O Duda , P E Hart , D G Stork . 2001. New York: Pattern Classification. Wiley.

[Jalal Uddin Mahmud Feroz Raihan Rahman ()] 'A Complete OCR System for Continuous Bangla Characters'. *Proceedings of the Conference on Convergent Technologies for the Asia Pacific*, Mohammed Jalal Uddin Mahmud, Chowdhury Mofizur Feroz Raihan, Rahman (ed.) (the Conference on Convergent Technologies for the Asia Pacific) 2003.

[Lu et al. ()] 'A Robust, Language-Independent OCR System'. Zhidong Lu , Issam Bazzi , Andras Kornai , John Makhoul , Premkumar Natarajan , Richard Schwartz . *27th AIPR Workshop: Advances in Computer-Assisted Recognition*, 1999. 3584 p. . (Proc. SPIE)

[Gunturk et al. ()] 'Eigenface-domain super-resolution for face recognition'. B K Gunturk , A U Batur , Y Altunbasak . *IEEE Transactions of . Image Processing*, 2003. 12 p. .

[Yahagi and Takano ()] 'Face Recognition using neural networks with multiple combinations of categories'. T Yahagi , H Takano . *International Journal of Electronics Information and Communication Engineering* 1994. 77 (11) p. .

[Fernandode La Torre and Black ()] Michael J Fernandode La Torre , Black . *Internatioal Conference on Computer Vision (ICCV'2001)*, (Vancouver, Canada) 2003. July 2001. IEEE 2001.

[Wang et al. ()] 'HMM Based High accuracy off-line cursive handwriting recognition by a baseline detection error tolerant feature extraction approach'. Wenwei Wang , Anja Brakensiek , Andreas Kosmala , Gerhard Rigoll . *7th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2000.

[Lyon ()] *Image Processing in Java*, Douglas Lyon . 1998. Upper Saddle River, NJ: Prentice Hall.

[Kailash et al.] 'Independent Component Analysis of Edge Information for Face Recognition'. J Kailash , Karande Sanjay , N Talbar . *International Journal of Image Processing* (3) p. .

[Rowley et al. (1998)] 'Neural Network-Based Face Detection'. H Rowley , S Baluja , T Kanade . `http://www.ri.cmu.edu/pubs/pub_926_text.html` *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998. January. 20 (1) p. .

[Bishop ()] *Neural Networks for Pattern Recognition*, C M Bishop . 1995. London, U.K.: Oxford University Press.

[Pal and Chaudhuri ()] 'OCR in Bangla: an Indo-Bangladeshi Language'. U Pal , B B Chaudhuri . *Proceedings of the 12th IAPR International*, (the 12th IAPR International) 1994. 2. (Computer Vision & Image Processing)

[Pal and Chaudhuri ()] 'OCR in Bangla: an Indo-Bangladeshi Language'. U Pal , B B Chaudhuri . *Proceedings of the 12th IAPR International*, (the 12th IAPR International) 1994. 2. (Computer Vision & Image Processing)

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