Enhancing the Character Segmentation Accuracy of Bangla OCR using BPNN

Shamim Ahmed[1], Mohammod Abul Kashem[2]

[1]M.S. Student, Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur-1700, Bangladesh

[2]Professor, Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur-1700, Bangladesh

Abstract: ***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. We study and examine the various kinds of problems or limitations that arise during the segmentation of Bangla characters and try to minimize that problems or limitations. The main aim of this proposed research is to improving the character segmentation (line, word, and character segmentations) accuracy of Bangla Optical Character Recognition (OCR) system by using Artificial Neural Network as a Back Propagation Neural Network (BPNN) algorithm.***

Keywords: Artificial Neural Network, character segmentation, line segmentation, optical character recognition, pre-processing, word segmentation.

1. Introduction

Optical Character Recognition (OCR) is one of the most important fields of pattern recognition in Computer Science (CS) and has been the center of attention for researchers in the last four decades [1]. The modern version of OCR appeared in the middle of the 1940's with the development of the digital computers [1]. Since then several character recognition systems for English, Chinese and Japanese characters have been proposed [2]. However, developing OCR systems for other languages such as Bangla didn't receive the same amount of attention [3]. The most common applications of OCR are reading postal address of envelopes, reading customer filled forms, archiving and retrieving text, digitizing libraries, automated vehicle number plate recognition, the handwritten and typewritten text could be stored into computers to generate databases of existing texts without using the keyboard, etc. Various designers have been actively involved in developing perfect optical character recognition (OCR) systems; still the state-of-the-art accuracy levels have room for improvement of OCR system for Bangla characters [4].

A lot of work has already been done on character recognition and font identification by many researchers around the globe but those researches are mainly on English, Korean, Japanese, Chinese languages. The first complete OCR on printed Bangla documents was proposed by Chaudhuri& Pal [5]. In this method, text digitization, noise removal, skew detection, and correction are done as part of pre-processing. The text documents are segmented into lines, words, and characters using horizontal–vertical projection profile analysis and headline removal techniques. Character Segmentation is one of the most challenging task of Bangla and any other OCR and overall recognition performance mostly depends on character segmentation. However, in this research paper, we study and analysis a better approach to improving the character segmentation performance or accuracy of Bangla Optical Character Recognition (OCR) by using Artificial Neural Network.

2. Outline of the Methodology

In this proposed research, we developed a system to enhancing the character segmentation performance or accuracy for typewritten and offline printed Bangla character recognition using Artificial Neural Network. The proposed system consists into two main parts: pre-processing and character segmentation. The input of the system is the digitally scanned image of Bangla characters and the output is the corresponding character segmented recognition result. The two main sections are discussed in below.

2.1 Pre-processing

The goal of this pre-processing part is to extract numeral images of printed text from a letter image for the subsequent recognition. Pre-processing consists of a number of preliminary processing steps to make the raw data usable for the recognizer. The typical pre-processing steps are included Image Acquisition, Background Removal, Binarization, Noise reduction or Soothing, Skew Detection and Correction, and Page layout analysis.

2.2 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 colour (RGB image) and this has to be converted into a binary image, based on a threshold value.

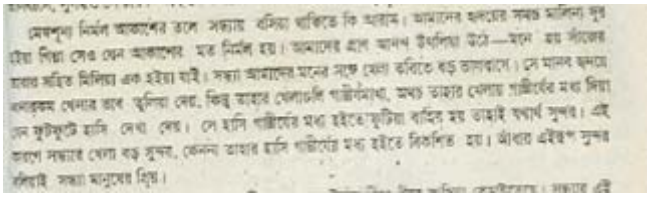

Figure 1: An example of digital image of Bangla text.

Y = Red * 0.2989 + Green * 0.5870 + Blue * 0.1140

The grayscale version of the image (Figure 1) is shown in Figure 2.

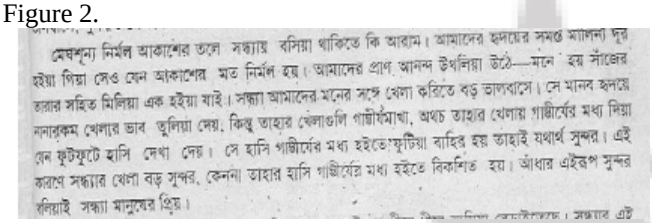

Figure 2: The grayscale image.

2.3 Thresholding Techniques

The algorithm but Otsu does not belong to the class of algorithms based on entropy. It is one of the most often used algorithms in image segmentation. Otsu's algorithm makes a discriminator analysis for defining whether a gray level t will be mapped into object or background information. The mean and variance of the object and background in relation to the threshold t are defined as follows.

$$m_b(t) = \sum_{i=0}^{t} i \cdot p_i$$

$$\sigma_b{}^2(t) = \sum_{i=0}^{t} [i - m_b(t)^2] p_i$$

$$m_w(t) = \sum_{i=t+1}^{255} i \cdot p_i$$

$$\sigma_w{}^2(t) = \sum_{i=t+1}^{255} [i - m_w(t)^2] p_i$$

The "optimal" value for this limit is the argument that maximizes the following expression.

$$\eta(t) = \frac{P_t(1-P_t)[m_b(t) - m_w(t)]^2}{P_t\sigma_b^2(t) + (1-P_t)\sigma_w^2(t)}$$

2.4 Background Removal

Threshold is the most trivial and easily applicable method for the differentiation of objects from the image background. It is widely used in image segmentation (Yahagi T. and Takano H., 1994) (Lawrence S., Giles C.L., Tsoi A.C. and Back A.d., 1993). We used threshold technique for differentiating the Bangla script pixels from the background pixels.

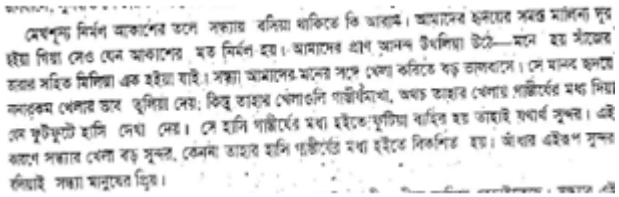

Figure 3: The binary image

2.5 Noise reduction or Soothing

Noise reduction or Soothing is one of the most important processes in image processing. Images are often corrupted due to positive and negative impulses stemming from decoding errors or noisy channels. Median filter is widely used for smoothing and restoring images corrupted by noise. It is a non-linear process useful especially in reducing impulsive or salt-and-pepper type noise. Median Filter is used in this study due to its edge preserving feature (Bishop C.M., 1995) (Kailash J., Karande Sanjay, Talbar N.) (FernandoDe La Torre, Michael J.Black 2003) (Douglas Lyon, 1998).

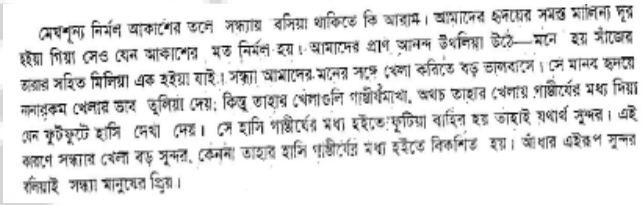

Figure 4: Noise free Image.

2.6 Skew Detection and Correction

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 labelling. As mentioned in (Schneiderman H., 2003), for skew angle detection, at first the connected component labelling 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 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.

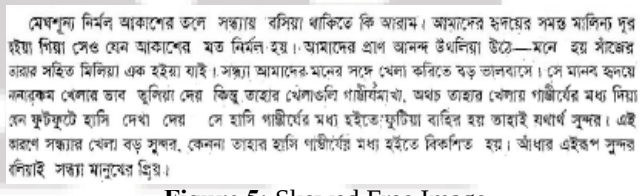

Figure 5: Skewed Free Image

3. Segmentation

Segmentation of binary image is performed in different levels includes lines segmentation, words segmentation, characters 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).

3.1 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.

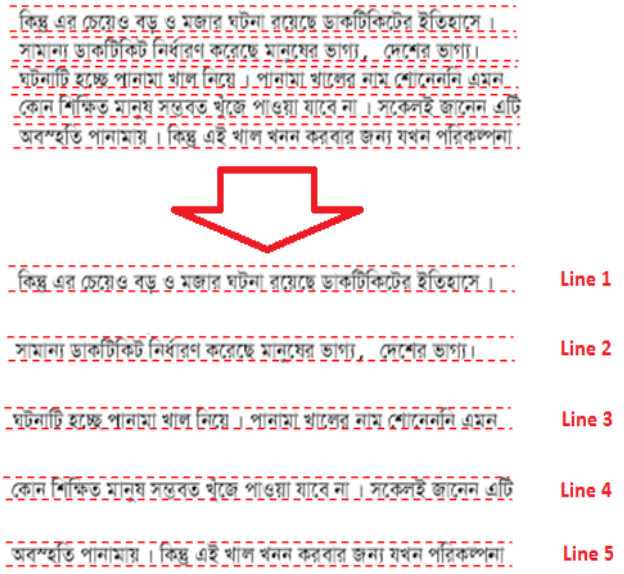

Figure 6: Line Segmentations.

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.

3.2 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. Figure 7 shows the word segmentation process.

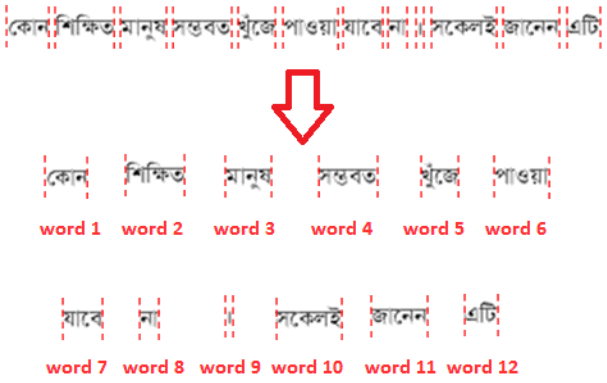

Figure 7: Word Segmentations.

3.3 Character Segmentation

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. 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.

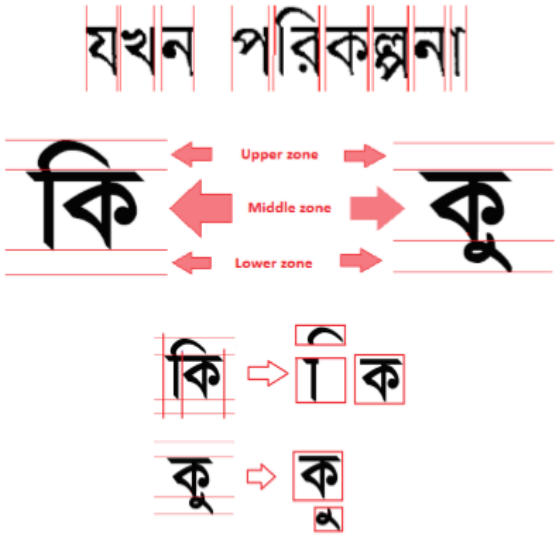

Figure 8: Character Segmentations.

3.4 Separation of Matra

Because of the existence of matraon most of the Bangla characters, in a horizontal scan over a line of text, the frequency of black dots in the head line will be the highest. Considering this property, the head line of any Bangla script can be identified. On removal of matra, the characters in a word are isolated and can easily be separated. But there are some Bangla characters (like M, Y, c, G, H etc) having no

matra. But some part of those characters span over the headline.

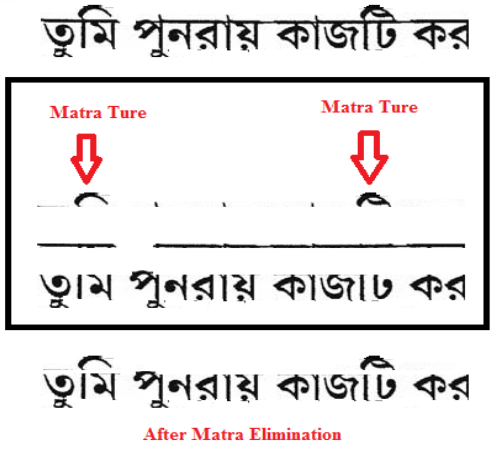

Figure 9: Matra Elimination.

Adding to the complexity, some characters (such as U, V etc) including some vowel modifiers (such as w, x) possess matra as well as some connecting portion over it. So a straightforward deletion of the matra splits the character into sub-segments.

4. Implementation and Results

In this research, for implementation we used Matlab simulator. In the figure 10, shows the output after successfully line and word segmentation of Bangla character as an input sample of figure 5. The success character segmentation rate is 89.3% and the failure rate is 10.7% shown in Fig. 11.

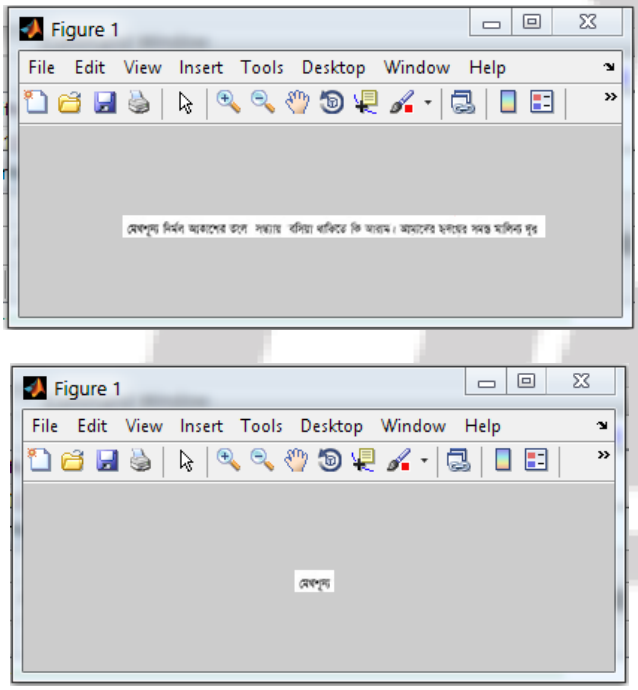

Figure 10: Output result of line and word segmentation of input figure 5.

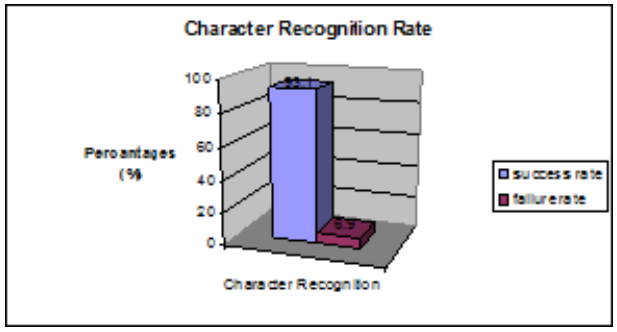

Figure 11: character segmentation rate.

5. Conclusion

This study aims at ascertaining to improve the overall performance of Bangla character segmentation including pre-processing, line segmentations, word segmentations, and character segmentations. To meet the requirements of a practical usage of Bangla characters, we consider not only the recognition rate and the error rate but also the response time. In this research paper, an efficient and better segmentation approach is successfully developed and tested by using artificial neural network back propagation algorithm. 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. 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.

References

[1] V. K. Govindan, and A. P. Shivaprasad, “Character Recognition – A Review,” Pattern Recognition (PA), Vol. 23, No. 7, pp. 671-683, 1990.
[2] K. Sekita, R. Toraichi, K. Mori, Yamamoto and H. Yamada, “Feature extraction of hand printed Japanese characters by spline function or relaxation matching,” Pattern Recognition (PA), Vol. 21, No. 14, pp. 9-17, 1988.
[3] X. L. Xie, and M. Suk, “On machine recognition of hand printed Chinese characters by feature relaxation,” Pattern Recognition (PA), Vol. 21, No. 14, pp. 1-7, 1988.
[4] H. Matsumura, K. Aoki, T. Iwahara, H. Oohama and K. Kogura, “Desktop optical handwritten character reader,” Sanyo tech, Vol. 18, pp. 3-12, 1986.
[5] B. B. Chaudhuri and U. Pal, “A complete Printed Bangla OCR System,” Pattern Recognition (PA), Vol. 31, No. 5, pp. 531-549, 1998.
[6] T. K. Bhowmik, A. Roy and U. Roy “Character Segmentation for Handwritten Bangla Words using Artificial Neural Network,” Neural Computation (NC), Vol. 28, No. 4, pp. 243-254, 2001.
[7] B. Scholkopf, A. Smola, K. R. Muller, “Nonlinear component analysis as a kernel eigen value problem,”

Neural Computation (NC), Vol. 10, No. 5, pp. 1299–1319, 1998.

[8] Z. Liang and P. F. Shi, "Kernel direct discriminant analysis and its theoretical foundation," Pattern Recognition (PA), Vol. 38, No. 1, pp. 445–447, 2005.

Author Profile

Shamim Ahmed he was born in Dhaka, Bangladesh on February, 1986. Now he has been serving as a lecturer, Department of Computer Science and Engineering (CSE), Bangladesh University of Business and Technology (BUBT), Dhaka, Bangladesh. He also has been studying as a M.Sc. in Engineering Student, Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh. He got B.Sc. in engineering degree in CSE in the year of 2010 from DUET, Gazipur, Bangladesh. He has been serving as an editorial board members and reviewers of some national and International Journals all around the world. Field of interest: Artificial Intelligence, Machine learning, Computer Vision, Wireless Sensor Networks, Systems and Networking, Distributed Computing System.

Mohammad Abul Kashem has been serving as a Professor, Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh. Field of interest: Speech Signal Processing.