Extraction and Recognition of Bangla Texts from Natural Scene Images Using CNN

Rashedul Islam^(✉) , Md Rafiqul Islam, and Kamrul Hasan Talukder

Computer Science and Engineering Discipline, Khulna University,
Khulna 9208, Bangladesh
rashedcse98@yahoo.com, dmri1978@yahoo.com, k.h.t@alumni.nus.edu.sg

Abstract. The semantic information presents in the scene images may be the useful information for the viewers who is searching for a specific location or any specific shop and address. This type of information can also be useful in licenseplate detection, controlling the vehicle on the road, robot navigation, and assisting visually impaired persons. An efficient method is presented in this paper to detect and extract Bangla texts from scene images based on a connected component approach along with rule-based filtering and vertical scanning scheme. Next, extracted characters are recognized by using Convolutional Neural Network (CNN). The method consists of the four basic consecutive steps such as detection and extraction of the Region of Interest (ROI), segmentation of the words, extraction of characters, and recognition of the extracted characters. After extracting the ROI from the input image, connected component(CC) analysis and bounding box technology are used for segmentation of Bangla words. To separate and extract Bangla characters from the segmented Bangla words, vertical scanning based method along with a dynamic threshold value has been applied. Finally, character recognition is carried out using CNN. The proposed algorithm is applied to 600 scene images of different writing styles and colors, and we have obtained 89.25% accuracy in text detection and 94.50% accuracy in the extraction of characters. We have achieved an accuracy of 99.30% and 95.76% in recognition of Bangla digits and characters respectively. By combining both the digits and characters, obtained recognition accuracy is 95.39%.

Keywords: Bangla text · Bounding box · Connected component · CNN · Vertical projection

1 Introduction

It is always challenging as well as an important task to extract and recognize texts from natural scene images. These types of images include banners, posters, billboards, license plates, etc. which may contain valuable information. This type of information can be used in many applications like the text to speech

conversion, text based image indexing, text mining [1], robot navigation, license plate recognition [2,3] etc. The variation in font size, color, style, alignment, light intensity, blurry image, noise, etc. makes it a difficult issue to design a standard Text Information Extraction (TIE) system. The extraction of Bangla text is another challenging issue as headline or ‘matra’ presents in this type of text. A ‘matra’ is a horizontal line located at the upper portion of a character. A Bangla text may be partitioned into three zones as shown in Fig. 1.

Fig. 1. Different zones of Bangla text.

As Bangla characters are connected by a headline or ‘matra’, we have proposed and applied a new algorithm to separate characters from each of the Bangla words by the method of vertical projection along with dynamic threshold values. The whole process of character detection, extraction, and recognition has been described in Sect. 3.

There is no benchmark database of scene images containing Bangla texts to perform research on extraction and recognition of Bangla characters. From this point of view, we have contributed to this field by providing a database of scene images consisting of 600 images. Another contribution of this paper is that we have a rich collection of Bangla characters which can be used by other researchers in developing a system of searching and recognition of office documents, text in scene images, etc.

2 Related Work

Text detection is a very challenging task for researchers who work with natural scene images. Various methods have been introduced earlier for the detection and localization of texts from scene images. In [2–4], text detection and localization techniques have been discussed based on the edge, texture, CC, stroke, and different combination of these methods. An edge detector is used in edge based method [5–7] for detecting the edges followed by morphological operation.

Bangla text extraction from the natural scene images is still now an ongoing research [8]. In the early stage, most of the researchers were concerned only with the images of printed documents, where the text was written in black color with white background [9]. Another method proposed by A. Asaduzzaman et al. [10] to detect and recognize Bangla text from printed documents using the heuristic method and Artificial Neural Network (ANN).

U. Bhattacharya et al. [11], proposed a method for the recognition of Bangla characters from scene images. The method can separate the CCs from scene

images using morphological operation by calculating height and standard deviation of the CCs. Their achieved precision and recall values were 68.8% and 71.2% respectively considering a set of 100 images. R. Ghoshal et al. [12] proposed a morphological approach for Bangla text extraction from images. Their approach was limited to highlighted texts only. The algorithm can perform detection of text area and segmentation of CCs.

In [13], a texture based method was proposed to detect text at gray level natural scene images. A probabilistic model with ANN based classifier is used here to separate text from non-text objects. They achieved text detection and false alarm rate of 64% and 25% respectively.

The detail description of the originality and other contributions of the work are given below.

1. The system architecture has been proposed for the extraction and recognition of Bangla characters from scene images. This proposed architecture is a new one and got very good results both in extraction and recognition of Bangla characters.
2. The proposed method of character extraction is a new one and can effectively perform the task and gives better results than the existing methods.
3. The experimental results of character extraction and recognition have been compared with other related methods. The results show that our proposed method gives better results both in extraction and recognition of Bangla characters.

3 Proposed Method

In this paper, the proposed method is executed in two phases. Such as: Character Extraction and Character Recognition. In the first phase, the main emphasis is given on text localization and extraction that lead to better accuracy of character extraction. In the second phase, character recognition is performed by using CNN. At first, the text area is selected then each of the text regions is marked by a rectangular bounding box and finally, individual characters are extracted from each text region using the newly proposed vertical scanning algorithm.

3.1 Character Extraction

In this phase, a database of scene image is prepared. Then some pre-processing measures are taken to resize the images into 500×500 pixels. Then the images are converted to Binary image. Some other necessary steps are taken to extract the characters from the scene images. The detail description of each of the steps of this phase is stated below.

Preparing a Database of Scene Images: Since there is no benchmark database of scene images with Bangla text, we have collected scene images from different locations of Bangladesh using the digital camera and the camera of the

smartphone. Then the captured images were renamed as 1.jpg, 2.jpg, 3.jpg, etc. and resized to a resolution of 500×500 and stored in a folder. The images were collected from different types of sources like banners, posters, billboards, vehicle license plates, etc. In this way we have created a database of scene images.

Pre-processing: This step involves two subsections as mentioned below.

Convert to the Grayscale Image: The captured images are the RGB image. So, to prepare them for the next step, we have to convert them into grayscale images. We have done it by using the National Television Standard Committee (NTSC) standard as shown in (1).

$$Gray_Image = 0.2999 * R + 0.587 * G + 0.114 * B \quad (1)$$

Convert to the Binary Image: To convert the grayscale image to a binary image, a threshold value is selected and all the gray level pixels below the threshold value are classified as 0 (black or background) and all the gray level pixels, equal to or greater than the threshold value are classified as 1 (white or foreground) as shown in (2).

$$g(x, y) = \begin{cases} 1, & \text{if } f(x, y) \geq T \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Here, $g(x, y)$ represents the threshold image pixel at (x, y) and $f(x, y)$ represents grayscale image pixel at (x, y) .

Detection and Extraction of ROI: In this process, the best possible regions are selected as ROI by the users. It helps to decrease false positive and also helps to collect more Bangla characters for preparing the training and test set.

Word Segmentation: CC based approach along with bounding box technology is applied to select each of the Bangla words as CCs. For this purpose, we have used the labeling of CCs of the binary image. Here all the CCs are marked by the red color rectangular bounding boxes as shown in Fig. 2(c).

Character Extraction: Bangla character extraction is one of the challenging tasks of the character recognition system. As the words are connected by a headline or 'matra', it is difficult to segment out individual characters. The technique to remove headlines to separate characters from Bangla words has been followed by the existing methods. But the main problem of removing 'matra' is that after removing the 'matra' some characters will be changed to another character. Some examples of such characters are shown in Fig. 3. We can solve this problem by the following way. At first, we take all the CCs as input and

Fig. 2. (a) Original color image (b) Binary image (c) Localization of Bangla words.

count the number of white pixels in every column to determine minimum value among all the columns. The column that contains minimum value will be treated as a separating zone among the characters of a word. To separate two characters vertically, we have set the pixel values of all the pixels of a specific column to 0 where the number of white pixels of the column is less than (minimum+5). Then the separated characters are resized to 16×16 pixels and store them into a specific folder as Bangla characters.

Original Character	After removing 'matra'	Look like the following character
৬	৬	৬ (Six)
৮	৮	৮ (Ha)
৯	৯	৯ (Dha)

Fig. 3. Effects of removing matra.

3.2 Character Recognition

This is the final stage of the proposed method. In this stage, experiment is performed based on the two consecutive phases such as the training and the testing phase. The brief description of each of the phases is stated below:

Training Phase

Prepare Training and Test Dataset: To prepare the data sets, at first it is required to load the database named 'banglacharacter' as an imagedatastore. The main function of the imagedatastore object is to automatically labels the images based on folder names. Finally data are stored as an imagedatastore object. To prepare the training data set, the system will randomly select a fixed number of images as mentioned by the user from each of the folders containing Bangla characters. In this experiment, we have assigned 250 as the number of images to be selected from each folder for training. The remaining characters of the folder will be treated as a test data set.

Initialize the CNN Layers: CNN is designed with many layers. To work with CNN, at first we have to define each of the layers by specific parameter values. Brief description of the layers of our designed CNN is stated below:

- *Image Input Layer:* In this layer, the image size is specified for our database. We have specified the said size as 16-by-16-by-1. Here, height and width of the image is 16 and the channel size is 1. The ‘banglacharacter’ data consists of binary images, so the channel size is 1. For a color image, the channel size 3 is recommended.
- *Convolutional Layer:* This layer contains three parameters. The first parameter is the filter size. The second parameter is the number of filters, which represents the number of neurons that connect to the same region of the input. ‘Padding’ name-value pair is used to add padding to the input feature map. We have used the following hyperparameters for the function `convolution2dLayer(3, 16, ‘Padding’, 1)`. FilterSize: [3 3], NumChannels: ‘auto’, NumFilters: 16, Stride: [1 1], PaddingMode: ‘manual’, and PaddingSize: [1 1 1 1].
- *Batch Normalization Layer:* This layer of CNN helps to speed up network training and reduces the sensitivity to network initialization.
- *ReLU Layer:* A ReLU layer performs a threshold operation to each element of the input, where any value less than zero is set to zero as shown in 3.

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x < 0 \end{cases} \quad (3)$$

- *Max Pooling Layer:* The max pooling layer returns the maximum values of rectangular regions of inputs. the hyperparameters that we have used for this layer are as follows: PoolSize: [2 2], Stride: [2 2], PaddingMode: ‘manual’, PaddingSize: [0 0 0 0]
- *Fully Connected Layer:* The last fully connected layer combines the features to classify the images. This layer is equal to the number of classes in the target data. For the classification of Bangla digits, we have used the function `fullyConnectedLayer(10)` with the two following hyperparameters. InputSize: ‘auto’ and OutputSize: 10

Figure 4 shows the working process of a CNN with input image ‘zero’.

Fig. 4. Classification process of CNN.

Set the Training Options: Before train the CNN classifier, it is required to specify the training options for classification of Bangla digits and characters. The Following training options have been used in this experiment.

options = trainingOptions('sgdm', 'InitialLearnRate', 0.01, 'MaxEpochs', 4, 'ValidationData', imdsValidation, 'ValidationFrequency', 30, 'Verbose', false, 'Plots', 'training-progress').

Train the Network: The Main purpose of training is to perform the task of recognition successfully. For this, the training data set is used along with predefined values of CNN layers and training options. These three parameters help to train the CNN successfully.

Testing Phase

Classify Using the Trained CNN: In this step, all the characters under the test data set are classified using the trained CNN. In this process labels of test data set are matched with the labels of the training data and obtained result is stored as predicted_data.

Calculation of Accuracy: At first, labels of test data set are stored as test_validation. Then Recognition accuracy is calculated by making a one-to-one comparison between the predicted_data and the test_validation. Figure 5 shows the system architecture of the proposed method.

Fig. 5. System architecture of the proposed method.

4 Experimental Results

The experiments were conducted in the following two phases. Such as

- a) Character extraction and
- b) Calculation of the accuracy of recognition.

In the first phase, Bangla characters were extracted from the natural scene images and in the second phase training and testing were performed on the extracted characters and the accuracy of recognition was calculated. All the experiments were performed in MATLAB environment using the images of our image database. The proposed method was applied to 600 scene images. The algorithm will not work properly or fail in the case of character extraction if all the characters are connected with each other by any way other than “matra”. The Algorithm will fail in another case where the texts are written in a curved or round shape. A few such images are shown in Fig. 6 where the proposed algorithm will fail to extract Bangla characters.

Fig. 6. Images where the proposed algorithm will fail.

Though there are some limitations of the proposed method, it is better in comparison with the existing methods regarding the results of the accuracy of extraction and the results of the accuracy of the character recognition. Detail description of the major two phases of the experimental results is given below.

4.1 Calculation of the Accuracy of Character Extraction

To analyze the results of character extraction, we have used four metrics, such as precision, recall, f1-score, and accuracy based on the following parameters [14]. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The accuracy of character extraction is calculated by the way as shown in (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Table 1 shows the percentage of precision, recall, f1-score, and the accuracy of character extractions from different types of scene images like banners, posters and license plates.

Table 1. Results of character extraction.

Image type	No. of images	Pr(%)	RR(%)	F1 score(%)	Accuracy(%)
Banner	240	92.73	94.08	93.00	92.92
Poster	260	95.30	94.13	94.45	95.09
License plate	100	95.43	97.58	96.32	95.48

4.2 Calculation of the Accuracy of Recognition

To calculate the accuracy of recognition, CNN predicts the labels of the test data using the trained network, and calculate the final validation accuracy. Accuracy is the fraction of labels that the network predicts correctly. Recognition accuracy is calculated by the following equation as shown in (5).

$$Accuracy = \frac{\text{total number of matching labels}}{\text{total number of elements in the test data}} \times 100\% \quad (5)$$

The comparison of recognition accuracy is shown in Table 2. The cited approaches mentioned in Table 2 do not use the same database as ours. In the table, ‘—’ indicates that the result was not found in the respective paper. From Table 2, it is clear that the proposed method outperforms the existing methods.

Table 2. Comparison of the recognition accuracy with existing methods.

Methods	No. of images	No. of characters	Recognition accuracy(%)
Proposed	600	21108	95.39
Moyeen, M.A., et al. [8]	400	—	73.25
Ghoshal, R., et al. [15]	100	7500	85.93
Ghoshal, R., et al. [16]	250	7100	92.00

5 Conclusions

The proposed method of Bangla character recognition has been tested on Bangla digits and letters extracted from the varied sorts of scene images and achieved smart ends up in comparison with the present strategies. To separate Bangla characters from the words, we have applied the vertical scanning algorithm. In the case of the extraction of Bangla characters, we’ve achieved 94.50% accuracy from 600 natural scene images. Within the recognition phase, character recognition is performed exploitation CNN classifier. We’ve used the CNN for the popularity due to its high accuracy. A hierarchical model is followed in the CNN that works on building a network, like a funnel, and at last offers out a

fully-connected layer wherever all the neurons are connected and the output is processed. The achieved recognition accuracy for Bangla digits is 99.30% and for Bangla characters, it is 95.76% and their combined result is 95.39% that is best than the results of the present strategies. Our future set up is to counterpoint our information with all the essential characters and joined letters of the Bangla alphabet and to represent the recognized characters in the editable form.

References

1. Bouakkaz, M., Ouinten, Y., Loudcher, S., Fournier-Viger, P.: Efficiently mining frequent itemsets applied for textual aggregation. *Appl. Intell.* **48**(4), 1013–1019 (2017). <https://doi.org/10.1007/s10489-017-1050-9>
2. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. *Front. Comput. Sci.* **10**(1), 19–36 (2016)
3. Zhang, H., Zhao, K., Song, YZ., Guo, J.: Text extraction from natural scene image: a survey. *Neurocomputing* **122**, 310–323 (2013)
4. Unar, S., Hussain, A., Shaikh, M., Memon, K.H., Ansari, M.A., Memon, Z.: A study on text detection and localization techniques for natural scene images. *IJCSNS* **18**(1) (2018)
5. Yu, C., Song, Y., Zhang, Y.: Scene text localization using edge analysis and feature pool. *Neurocomputing* **175**, 652–661 (2016)
6. Silva, B.L.S., Ciarelli, P.M.: Edge detection and confidence map applied to identify textual elements in the image (2016)
7. Lee, S., Cho, M.S., Jung, K., Kim, J.H.: Scene text extraction with edge constraint and text collinearity. In: 20th International Conference on Pattern Recognition (ICPR), pp. 3983–3986. IEEE (2010)
8. Moyeen, M.A., Alam, K.M.R., Awal, M.A.: Bangla text extraction from natural scene images for mobile applications. *J. Electr. Eng. Inst. Eng.* **EE 39**(I & II) (2013)
9. Aurich, V., Weule, J.: Non linear Gaussian filters performing edge preserving diffusion. In: 17 DAGM Symposium, pp. 538–545 (1995)
10. Asaduzzaman, A., Molla, M.K.I., Ali, M.G.: Printed Bangla text recognition using artificial neural network with heuristic method. In: Proceedings of ICCIT, Dhaka, Bangladesh (2002)
11. Bhattacharya, U., Parui, S.K., Mondal, S.: Devanagari and Bangla text extraction from natural scene images. In: Proceedings of International Conference on Document Analysis and Recognition, pp. 26–29 (2009)
12. Ghoshal, R., Roy, A., Bhowmik, T.K., Parui, S.K.: Headline based text extraction from outdoor images. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) *PRMI 2011*. LNCS, vol. 6744, pp. 446–451. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21786-9_72
13. Hanif, S.M., Prevost, L.: Texture based text detection in natural scene images: a help to blind and visually impaired persons. In: Conference Workshop on Assistive Technologies for People with Vision Hearing Impairments Assistive Technology for All Ages CVHI
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)

15. Ghoshal, R., Roy, A., Parui, S.K.: Recognition of Bangla text from scene images through perspective correction. In: 2011 International Conference on Image Information Processing (ICIIP), pp. 1–6 (2011)
16. Ghoshal, R., Roy, A., Dhara, B.C., Parui, S.K.: Recognition of Bangla text from outdoor images using decision tree model. *Int. J. Knowl.-Based Intell. Eng. Syst.* **21**(1), 29–38 (2017)