A TWO STAGE METHOD FOR BENGALI TEXT EXTRACTION FROM STILL IMAGES CONTAINING TEXT

Ankita Sikdar[1], Payal Roy[1], Somdeep Mukherjee[1], Moumita Das[1] and Sreeparna Banerjee[2]

[1]Department of Computer Science and Engineering, West Bengal University of Technology, Kolkata, West Bengal, India
ankita.sikdar@gmail.com
payalroys@gmail.com
somdeep.mukherjee@gmail.com
moumitadas8484@gmail.com
[2]Department of Natural Sciences, West Bengal University of Technology, Kolkata, West Bengal, India
sreeparnab@hotmail.com

ABSTRACT

Bengali text data present in multimedia images having multiple content forms, such as still images and text, contain information that when extracted finds a lot of applications. The images can be of different types, where objects and text may be completely separated or overlapped or embedded in each other. The Bengali text can be of different shapes and sizes. Extraction of text from these types of images becomes challenging because the textual portion has to be correctly separated from the rest of the background. The input image passes through two stages. The first step tries to locate the different components in the image using entropy filtering and the second stage distinguishes the components representing text from the non-textual components based on several features of Bengali text. The text thus obtained from the image can then be used in software such as Bengali OCR for character recognition.

KEYWORDS

Bengali character feature identification, Connected components, Entropy filtering & Text extraction

1. INTRODUCTION

Bengali text which appears in the images may have different shapes, sizes, color, fonts and orientations. Extracting this textual portion from an image poses a challenging problem as we have to carefully identify the Bengali text from the surrounding objects. In this paper, we have focused on still images having horizontally aligned text. Our proposed work is to correctly separate the connected components in the multimedia image and then identify the component as text or non-text. Our method is divided into two stages. In the first stage, the image containing text is first converted to a logical image, followed by histogram thresholding and entropy filtering, after which the connected components in the image have been extracted. In the second stage, each of the components is classified as text or non-text based on an exhaustive study of the features of Bengali characters and Bengali text.

Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 47–55, 2012.
 DOI : 10.5121/csit.2012.2306

Bengali text extraction and identification finds its application in banks, post offices, library automation and publishing industry where Bengali language is used [1]. If we are able to automatically do the tasks of text identification, then that would save a lot of manual labor. Moreover, this also helps in preparing digital versions of Bengali books.

The paper is outlined as follows: Section 2 discusses the past work. Section 3 presents the stepwise algorithm for the two stages of our method. Section 4 describes the procedure in details along with an illustration. Section 5 shows the different types of test inputs followed by the results and discussions. Section 6 gives a conclusion and future research scope on our work.

2. PAST WORK

Bengali text extraction from multimedia images is a new research area and a large number of researchers have been working in this field. Some of the earlier works in this field were first done on scanned images which contained only textual portions where the background was in white and the text was in black [2]. The pre-processing phase becomes more challenging when the text has to be extracted from images containing non-connected components. Pre-processing of a picture including character string with multi-coloured background is followed by text recognition phase. Bhattacharya et al. [3] proposed a morphological approach of text area extraction obtaining connected components by applying horizontal line segmentation, then bridging them and calculating features like height, mean and standard deviation. Based on that, they performed image operations such as dilation, erosion and opening. The precision and recall values of their algorithm obtained on the basis of the present set of 100 images were respectively 68.8% and 71.2%. Captured images of head lighted text in Bengali suffer from proper character segmentation also. Another morphological approach was proposed by Ghoshal et al. [4] based on detecting unattached text area and segmenting connected components of Bengali/Devnagari characters from an image. Their online and automated system of Bengali character recognition was based on calculating features of connected components such as elongations ratio, number of holes present in the segmented character, aspect ratio and object to background pixels ratio. However the approach is restricted to capture pictures of high lighted text only. This algorithm of text extraction can be extended to scanned copy of the Bengali pictorial text also.

3. ALGORITHM

The algorithm for the first stage is presented as follows [5, 6, 7]:

1. Convert the RGB image into a grayscale image.
2. Use the Otsu's method to compute a global threshold of the grayscale image and use that value in order to convert the grayscale image to a logical image.
3. For each pixel in the logical image, repeat steps 4 to 6.
4. Create an image histogram for the neighborhood of that particular pixel.
5. Count the number of pixels belonging to each of the two intensities, '0' and '1' and store these values in p_0 and p_1 respectively.
6. Calculate the entropy value E around the neighborhood for that particular pixel using the formula,
$$E = -(p_0 \times log_2 p_0 + p_1 \times log_2 p_1) \quad (1)$$
7. Convert the entropy filtered image to a logical one.
8. For each connected component obtained in the logical image, repeat steps 9 to 11.
9. Trace the exterior boundary of the component and store the coordinates of boundary pixels in an array.

10. Using the boundary pixels in step 9, extract the corresponding component from the logical image obtained in step 2.
11. Store the component obtained as a separate logical image.
12. Algorithm ends.

The algorithm for the second stage is presented as follows:

1. For each image representing an extracted component obtained in the first stage, repeat steps 2 to 7.
2. Calculate the relative density for each row in the image using the formula

$$Relative\ density = \frac{number\ of\ occupied\ pixels\ in\ a\ row}{total\ number\ of\ pixels\ in\ each\ row} \times 100 \quad (2)$$

3. Plot a graph of relative density, where the abscissa represents the row number and the ordinate represents the relative density.
4. If the maximum peak of the graph is greater than or equal to 70, go to step 5 else the image does not represent Bengali text and go to step 8.
5. Scan the left neighborhood of the maximum peak. If a peak is encountered having value greater than 20, the image does not represent Bengali text and go to step 8 else go to step 6.
6. Scan the right neighborhood of the maximum peak. If a peak is encountered having value more than 60, the image does not represent Bengali text and go to step 8 else go to step 7.
7. Scan the image for the presence of few features typical of Bengali text:
   (i) The last row having a relative density above 70 will be connected to the subsequent row at a number of points. The width of each connection will be lesser than 5% of the number of pixels in each row.
   (ii) Below the last row having relative density above 70, the connected components will have uniform spacing between them except in very less number of cases when the components may be overlapped.
   (iii) The width of each non-overlapping component scanned in step 7(ii) will fall in two different ranges; one range (modifiers) will be much lesser than the other (characters).
   (iv) Check the pattern of the component to determine the presence of strokes and curves as shown in Fig 2(f) that make up Bengali characters.

   If the above conditions are satisfied, the image represents Bengali text else, else the image does not represent Bengali text.
8. Algorithm ends.

4. PROCEDURE

In a multimedia image, objects of different shapes and sizes may occur with textual portion in it. In order to detect the Bengali textual portion in an image, the image is passed through a two stage process.

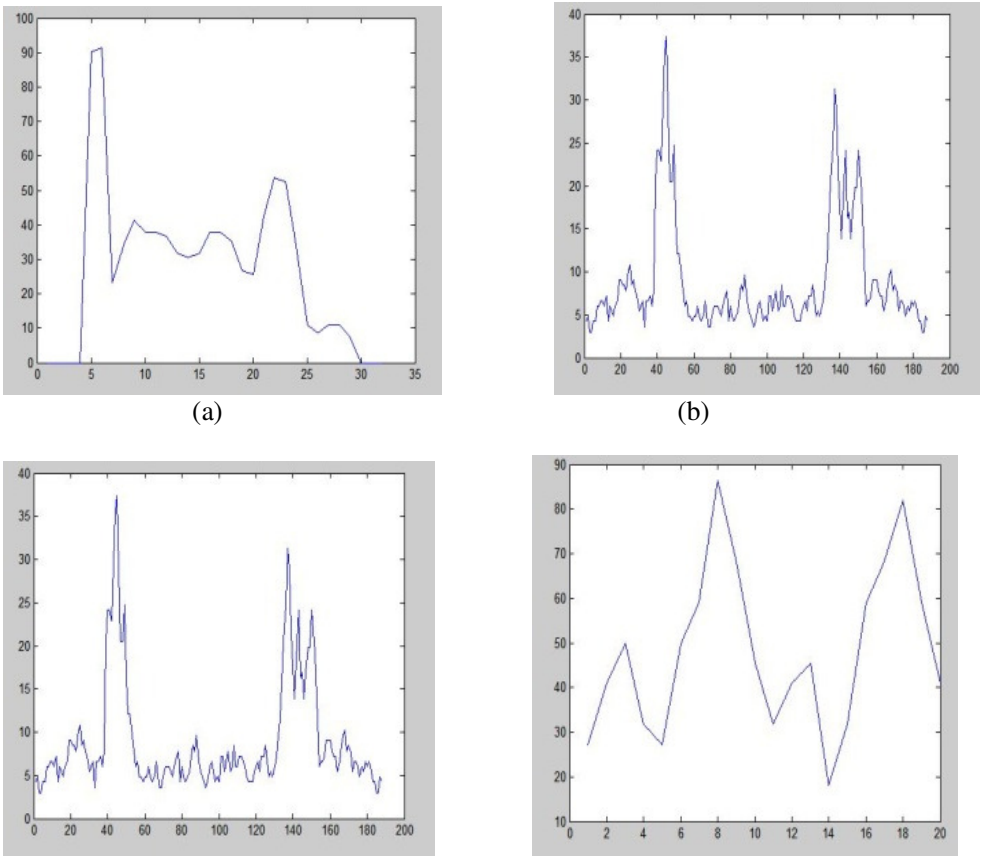
4.1 GENERATION OF COMPONENTS

The first stage converts the RGB image to a grayscale image. This is followed by obtaining a threshold value by Otsu's method to convert the grayscale image to binary form. Now, the image will consist of pixels of values '0' and '1', where a '0' corresponds to the background and a '1' corresponds to the foreground. At this stage, if we directly try to obtain the connected

components, we might get erroneous results as there will be a large number of connected regions which are actually fragments of a bigger component. This is because in the grayscale image, the same component will be represented by pixels of varying intensities and when we try to convert the grayscale image to a binary image using Otsu's method, some pixel values of that component in the grayscale image might have values lesser than the global threshold calculated by the Otsu's method and thus they will be converted to a '0' value in the binary image. When this takes place for quite a number of pixels within a component, it might lead to fragmentation. So, we have to apply various pre-processing steps in order to correctly extract the component. In this approach, we calculate the entropy value for each pixel in the binary image. Entropy is a statistical measure of randomness and it helps to classify the texture of the image. In order to calculate the entropy of a pixel, we first obtain the histogram for the neighbourhood of that pixel and then using the formula given in (1), we calculate the entropy. The entropy filtered image will then clearly show each connected component in the image very distinctly since it has clustered together all the components falling in the same intensity range. Then we convert the entropy filtered image to a logical one. After that, the boundaries of the connected components can be easily traced and corresponding to the boundary traced, the original component can now be extracted from the logical image which was obtained in step 2 of our first algorithm.

4.2 TEXT DETECTION

The second stage uses various features of Bengali text and heuristics that were obtained after extensive studies and thereby tries to identify which components obtained in the first stage represent Bengali text and discards the other components. Bengali text has a thick horizontal line known as the "matra". The characters are connected to the "matra" from below forming the lower part. Sometimes a part of the character may extend above the 'matra" forming the upper part. Now, if we calculate the row-wise relative density for Bengali text with "matra" using the formula given in (2), and plot a graph of relative density versus row number, we will always obtain a graph of the form shown in Fig 1(a). This is because the rows representing the "matra" will always occupy the maximum number of pixels and the relative density of occupied pixels will be greater than or equal to 70, giving rise to the high peak, as the "matra" will extend across the entire length of the text. If the peak is less than 70, we can reject the image as non-text. Such a plot is represented in Fig 1(b).If the maximum value of the peak is more than 70, yet there are other peaks to the left side of the maximum peak that have value greater than 20, we conclude that the image does not represent Bengali text. This is based on the feature that above the "matra" there can only be few curves corresponding to upper parts of characters, which will have relative density of at most 20. Such a plot is represented in Fig 1(c). If the value of maximum peak is 70 and yet there are peaks to the right side of the maximum peak that have value more than 60, then we conclude that the image does not represent Bengali text as there is more than one peak having a value very close to the global maximum. This is based on the feature that the portion of image consisting of the characters will not have a relative density that is as high as 60. Such a plot is represented in Fig 1(d). If the image satisfies the above properties, then we say the image has a possibility of representing Bengali text so we check in the image for more features that match with Bengali text. We find out the number of components connected to the "matra" in the image and also the width of each connection. This width will be less than 5% of the number of pixels in a row of that image. The characters are separated by a column of pixels of value '0' except sometimes when the characters may overlap. Also, the width of all the characters should have a same range of values. Sometimes there are features (vowel and consonant modifiers) appearing before and/or after the characters that will be occupying very less width compared to the width of the character. Finally, we check the patterns of the components to see if they match with the strokes and curves which make up the Bengali characters. If the image satisfies these features, then we can classify that image as Bengali text. The above algorithm can in fact detect text from all the languages which are of the form of Bengali that is the characters are attached to a "matra".



(a) (b)

(c) (d)

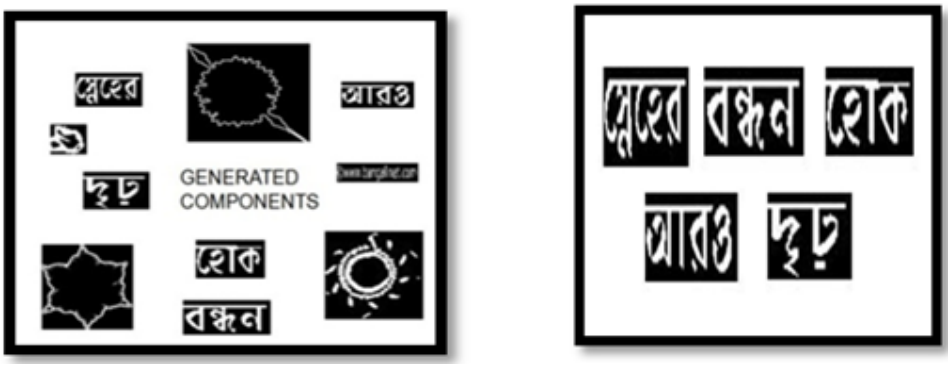
Fig 1. Plots representing relative density of occupied pixels per row (in percentage) for the images obtained

(a) Image representing Bengali text
(b) Image does not represent Bengali text as the maximum peak is less than 70.
(c) Image does not represent Bengali text as there are peaks to the left side of maximum peak having value more than 20.
(d) Image does not represent Bengali text as there are peaks to the right side of maximum peak having value more than 60.

Fig 2 illustrates the proposed method with the help of an example. An input image is passed through the different stages of the algorithm and the images obtained at each of the steps are shown.

(a) (b) (c)

(d) (e) (f)



(g) (h)

Fig 2. An image passing the different stages of the algorithm

(a) Original RGB image
(b) RGB image converted to grayscale image.
(c) Grayscale image converted to logical image
(d) Entropy filtered image
(e) Entropy filtered image converted to logical image
(f) Some strokes and curves to be matched
(g) Connected components extracted from the logical image
(h) Bengali text distinguished from non-textual regions

5. RESULTS AND DISCUSSION

In order to test the accuracy of the proposed method, a large number of input images of different types are collected and the above algorithm was run on these images. The images were of the following types:

(i) Images with only text.
(ii) Images where the object and text were completely separated.
(iii) Images where the object and text was overlapped.
(iv) Images where the text was present in between the objects.

A sample image for each case is shown in Fig 3.

(a) (b) (c) (d)

Fig 3. Sample image for each case
(a) Images with only text
(b) Images where the object and the text were completely separated
(c) Images where the object and text was overlapped
(d) Images where the text was present in between the object

The results are shown in Table 1. From the experimental results, it is observed that in all the four cases, the algorithm has given fairly high results, successfully identifying the textual components in the image. In case of simple images, where only text is present or text is present separately from the object, we almost get full accuracy in every test image. However, in case of complex images, when the text is overlapped with an object, the accuracy is bit lower than in the first two cases. This is because, in some test images, some part of the Bengali character in the text is very intricately combined with the object and thus can not be distinguished with full accuracy. In case of images where text is present in between objects, the algorithm is also able to separate the connected components properly.

Table 1.Heading and text fonts

Image Type	No. of sample images tested	No. of images where the textual portions were correctly identified	Accuracy
Images with only text	50	50	100 %
Images where the object and text were completely separated	50	49	98%
Images where the object and text were overlapped	50	47	94%
Images where the text was present in between the objects	50	48	96%

6. CONCLUSIONS AND FUTURE SCOPE

The algorithm presented in this paper, when implemented is able to identify Bengali text from a certain level of complex multimedia images with a desired accuracy level. In some cases it may so happen that there is a component that is actually an object but because of its features, may misguide the algorithm to identify the component as text. Also, in some complex graphical images like cartoon images, the Bengali text may be present without the "matra". So it is difficult to design a generalized algorithm that will be able to identify textual portions from all kinds of images. However, the algorithm presented in this paper has been designed to cover a broad scope. Future work on this paper will include working on more number of images with higher levels of complexity so that more complex patterns of Bengali text represntation may be identified.

ACKNOWLEDGEMENTS

The authors would like to thank Soumendu Das, an M.Tech student of West Bengal University of Technology for his valuable contribution towards the survey of literature and overall discussion.

REFERENCES

[1] Mohammed Jasim Uddin, Mohammed Towhidul Islam and Md. Abdus Sattar, *Recognition of Printed Bangla Characters Using Graph Theory,* National Conference on Computer and Information System-NCCIS, Dec 9-10, 1997, Dhaka, Bangladesh.

[2] A.O.M Asaduzzaman, Md. Khademul Islam Molla and M. Ganjer Ali, *Printed Bangla Text Recognition using Artificial Neural Network with Heuristic Method,* Proc. ICCIT'2001, 28-29 December, East West University, Dhaka, Bangladesh.

[3] U. Bhattacharya, S. K. Parui and S. Mondal, *Devanagari and Bangla Text Extraction from Natural Scene Images,* Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on 26-29 July 2009.

[4] Ranjit Ghoshal, Anandarup Roy, Tapan Kumar Bhowmik and Swapan K. Parui, *Headline based Text Extraction from Outdoor Images,* Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, 2011, Volume 6744/2011, 446-451, DOI: 10.1007/978-3-642-21786-9_72.

[5] Fakulteta za stronjnistvo, Character Recognition, *Handwritten character Recognition: Training a Simple NN for classification using MATLAB*, Technical Report.

[6] Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing" 2nd Edition: Prentice Hall.

[7] Image Processing Toolbox, User's Guide.

Authors

Ankita Sikdar has done her schooling form Mahadevi Birla Girls' Higher Secondary School. She is at present a fourth year student of West Bengal University of Technology pursuing B.Tech in Computer Science and Engineering. She is going to pursue Phd in The Ohio State University, research area : artificial intelligence

Payal Roy has done her schooling from Carmel School and Hem Sheela Model School. She is at present a fourth year student of West Bengal University of Technology,Kolkata pursuing B.Tech in Computer Science and Engineering. She has bagged a few job offers and will be soon joining the industry.

Somdeep Mukherjee has completed his schooling from St. Xavier's Collegiate School, Kolkata. He is at present a fourth year student pursuing B.Tech in Computer Science and Engineering from West Bengal University of Technology, Kolkata. He has got job offers and is about to join the industry.

Moumita Das is a fourth year student of West Bengal University of Technology pursuing B.Tech in Computer Science and Engineering. She is about to join the industry.

Dr Sreeparna Banerjee obtained her B. Sc., M.Sc., and Ph.D degrees all in Physics. She has taught in universities in India and abroad. Her current research interests include Physics of space plasmas: Molecular Dynamics and Monte Carlo simulations,charge transfer, nonlinear dynamics; Neural networks, Pattern Recognition and Soft Computing applications in Astrophysics, Meteorology and Medical Imaging; Data Mining.