Text Line Height Calculation: Methodology for Labeling the Segmented Bengali Text Document

Biswajit Debnath,* Md. Sajid Shahriar,Sabir Ismail and Md. Saiful Islam
Deparment of Computer Science and Engineering,
Shahjalal University of Science and Technology,
Sylhet-3114, Bangladesh
biswajit.sust@gmail.com*

***Abstract*—Calculation of the text line height is the most important part in pre-processing of OCR. After the completion of character recognition part in a OCR system, it starts the post-processing steps and afterwards rebuild the document as a editable document. When the OCR system is rebuilding the document, it must have the track of each text font size, line gap etc. The calculation line height is necessary to extract the font size from the image document. It will also be helpful when the segmented text documents needed to be labeled as Headline, Sub Headline, Paragraph and so on. These are the most contained parts in every documents. The methodology we are introducing in this paper for labeling the Bengali Documents, where the document is well segmented by an OCR System. It starts the work with the textual documents by clusterization of the height of the each connected elements of that document afterward finding out the standard deviation, create a set of Candidate line height for that documents and a set of rules to find out the line height from the candidate sets.**

***Index Terms*—Segmented Text Document, Optical Character Recognition, Text Line Height Detection, Connected Component, Standard Deviation,Layout Analysis, Text Type Labeling.**

I. Introduction

OCR intelligence needs to detect the font size in text documents in order to understand the document image and extract knowledge from the document texts [1], [2]. Because it is one of the main condition for an OCR is to recognize or to specify the text independent font size[3]. If we consider the documents like newspaper, they are constructed with title, section, subsection, paragraphs , columns etc. So the layout of these kind of documents could be reconstructed as it was by extracting the font size information [4]. The texts of this printed document is extracted is also done in the pre-processing step[5]. The extracted text documents don't contain any graphical features [5] so the line height calculation of those extracted text parts is such a simple technique that will solve the problem how to classify the text according to the previously discussed categories.

Working with the scanned documents is doing the processing with an uncompressed image document. Wilcox at el.[2] proposed a technique which use the Line Segmentation process in order to detect the font size.

On the other hand, compressed document gives us the high performance and storage adaptability, so that machines like fax machines [3], Photocopier machines and many others uses the compressed form of the documents [4].If the documents are like run length compressed TIFF documents detection of the font size come handy [4] in terms of complexity like time and space [6], [7], [8]. Mohammed Javed at el. [4] proposed an approach to automate the font size detection straight from the run length compressed TIFF documents by working with the binary documents. These methodology worked without doing any decompression of the document [4]

In our research work we worked with uncompressed Bengali documents. So the methodology we are proposing is for Scanned Newspaper. Newspaper is a very complicated document where the text lines are jam packed. As the lines are kept tight to each other a common problem like overlapping of words of different lines are highly seen. The top priority we gave is to solving some problems which are commonly seen in every Bengali Newspaper like documents in order to have the most accurate font size.

We organized our work in this literature as follows.First we are going to learn about the documents in Section II. A brief discuss about the document structure, need of the methodology we are proposing according to the outline of the document will be discussed there. The methodology will be illustrated in Section III. Then we will evaluate our methodology in IV using the Popular Bengali Newspaper Prothom-Alo. The documents are used for testing purpose are scanned documents. In the end we will give a conclusion about our research work in Section V.

II. Studying About the Document

The Topics in this Section will cover about the Structure. We will try to cover about the document buildup and layout which are technically used in our research work.

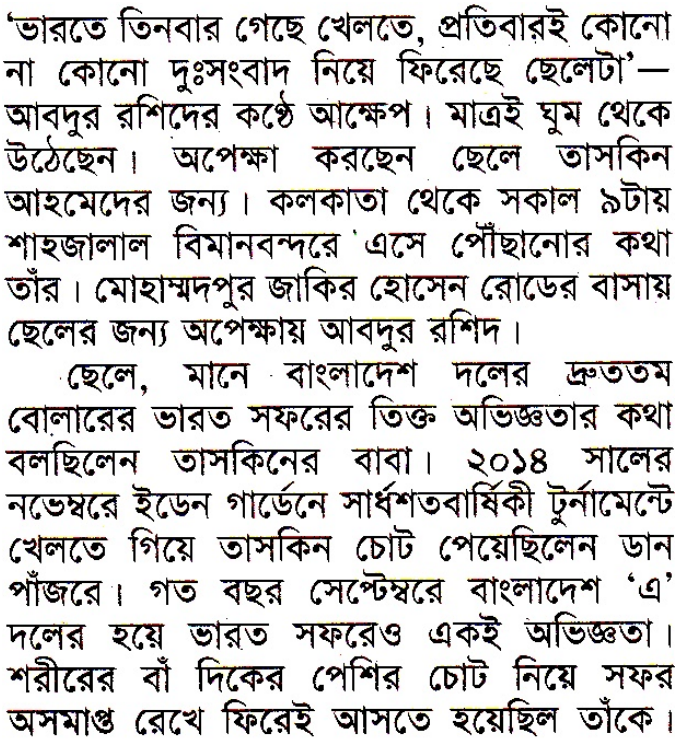
A. Structure

In order to implement the idea of detecting the font size from the line height, we have segmented our document properly . So there will be only textual part of the document. If there was any kind of image or other type of property, it have been segmented already along with the noise removed. Documents may have one to many lines and also it can contain more than one paragraph in each segmented documents. Also it is presumed that there will be only Bengali alphabets, no mix up with other language alphabets. The sample of a document is given in Fig. 1

If the segmented document contain a headline, it will carry the highest line height of that document. and there will be a very few lines than a paragraph. A paragraph will have the most lines and also each lines are congested to each others.

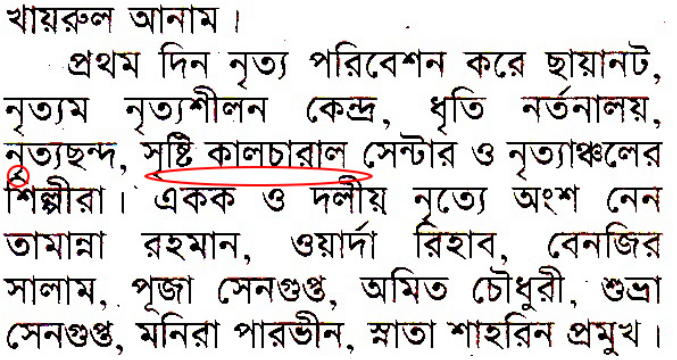
B. Overlapping lines

In Bengali document,if the lines are too much congested,it may create line overlapping case. Bengali words has also


‘ভারতে তিনবার গেছে খেলতে, প্রতিবারই কোনো না কোনো দুঃসংবাদ নিয়ে ফিরেছে ছেলেটা’— আবদুর রশিদের কণ্ঠে আক্ষেপ। মাত্রই ঘুম থেকে উঠেছেন। অপেক্ষা করছেন ছেলে তাসকিন আহমেদের জন্য। কলকাতা থেকে সকাল ৯টায় শাহজালাল বিমানবন্দরে এসে পৌঁছানোর কথা তাঁর। মোহাম্মদপুর জাকির হোসেন রোডের বাসায় ছেলের জন্য অপেক্ষায় আবদুর রশিদ।

ছেলে, মানে বাংলাদেশ দলের দ্রুততম বোলারের ভারত সফরের তিক্ত অভিজ্ঞতার কথা বলছিলেন তাসকিনের বাবা। ২০১৪ সালের নভেম্বরে ইডেন গার্ডেনে সার্ধশতবার্ষিকী টুর্নামেন্টে খেলতে গিয়ে তাসকিন চোট পেয়েছিলেন ডান পাঁজরে। গত বছর সেপ্টেম্বরে বাংলাদেশ ‘এ’ দলের হয়ে ভারত সফরেও একই অভিজ্ঞতা। শরীরের বাঁ দিকের পেশির চোট নিয়ে সফর অসমাপ্ত রেখে ফিরেই আসতে হয়েছিল তাঁকে।

Fig. 1. A Sample Document for font size detection.


খায়রুল আনাম।

প্রথম দিন নৃত্য পরিবেশন করে ছায়ানট, নৃত্যম নৃত্যশীলন কেন্দ্র, ধৃতি নর্তনালয়, নৃত্যছন্দ, সৃষ্টি কালচারাল সেন্টার ও নৃত্যাঞ্চলের শিল্পীরা। একক ও দলীয় নৃত্যে অংশ নেন তামান্না রহমান, ওয়ার্দা রিহাব, বেনজির সালাম, পূজা সেনগুপ্ত, অমিত চৌধুরী, শুভ্রা সেনগুপ্ত, মনিরা পারভীন, স্নাতা শাহরিন প্রমুখ।

Fig. 2. Documents with overlapping Ascender and Descender.

Ascender parts and Descender parts, whether there is no capital letter and small letter. The Ascender parts are above the Matra (A horizontal line-stroke that runs on the top of the letters to create a distinct word.) and the Descender parts are the some of the Diacritic form of the vowels which are use bottom end of a letter. If the lines are too much congested, then we can find out that there is no line gaps. A document is given with marked where the overlapping case is occurred in the Fig. 2.

There was a possibility to detect the line height using the technique proposed by Mohammed Javed at el.[4]. Extracting the histogram of the documents would have been the possibility to get the line height. but this kind of overlapping case will show us two line as one line.

So our proposed method deals with this cases and calculates the line height for the document.

III. Methodology

The methodology we are going to introduced for the calculation of line height, needs seven major steps to complete the detection.

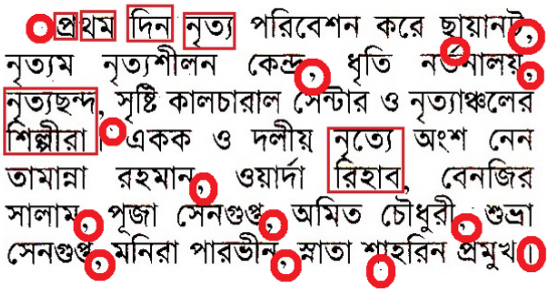

প্রথম দিন নৃত্য পরিবেশন করে ছায়ানট, নৃত্যম নৃত্যশীলন কেন্দ্র, ধৃতি নর্তনালয়, নৃত্যছন্দ, সৃষ্টি কালচারাল সেন্টার ও নৃত্যাঞ্চলের শিল্পীরা। একক ও দলীয় নৃত্যে অংশ নেন তামান্না রহমান, ওয়ার্দা রিহাব, বেনজির সালাম, পূজা সেনগুপ্ত, অমিত চৌধুরী, শুভ্রা সেনগুপ্ত, মনিরা পারভীন, স্নাতা শাহরিন প্রমুখ।

Fig. 3. Non-dominant word examples.

It starts the work by calculating all the connected component which are within the document.The Flood Fill algorithm is efficient way to calculate the line height for the documents. In this process we will get the lists of heights of every connected components. We will call these list of heights as candidate heights.

But in this way we will get a variety of heights from the documents. Because a scanned document can not be fully noise free, though it is went through a noise filter process. Also because of overlapping line problem we will get some heights which are equivalent to two words. Example of this problem is given in Fig. 3

The ratio of these heights don’t have a dominance over the whole document. So these height are labeled as non-dominant height in our methodology. By setting a threshold value, and comparing with it these non-dominant heights are removed from our list of heights. So we can say at this moment our candidate heights is now filtered.

For more filtering we have to calculate two more things, The Average Line Height and the Standard Deviation from the candidate heights. By subtracting the standard deviation from the average heights we will get Height Minimum and by adding will give us Height Maximum. This two value will be the range of needed to filter remaining candidate height. If the candidate height is between the equation below,

$$H_{max} < H_{cadidate} < H_{min} \tag{1}$$

those height will remain through the filtration process. This filtration is shown as data figure in Fig. 4

The final line height will be received by finding out the average height which remained after the filtration process above.

IV. Experimental Result and Analysis

To test our methodology we have collected 300dpi scanned image of total 30 noise and skew free documents where each were perfectly segmented and don’t contain any kind of image. Each of these document is made from hard copy of re-known daily newspaper Prothom alo. Using our methodology we found out the Line Height of the Headline is 155px to 190 px, 50px to 75px for sub-headline and for the paragraph it is around 25px to 32px. The overall accuracy of this about 96.56%.

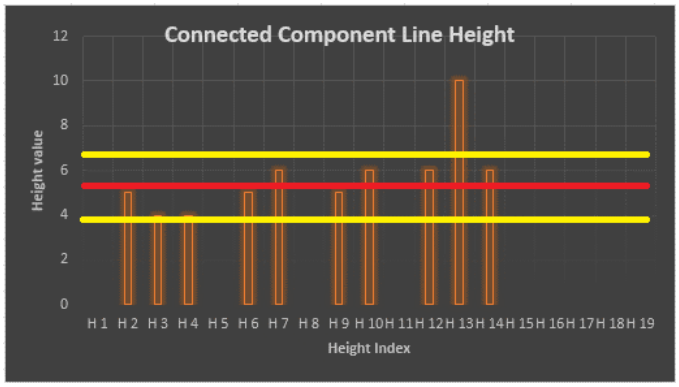


Fig. 4. Data figure of candidate line heights of sample data set.

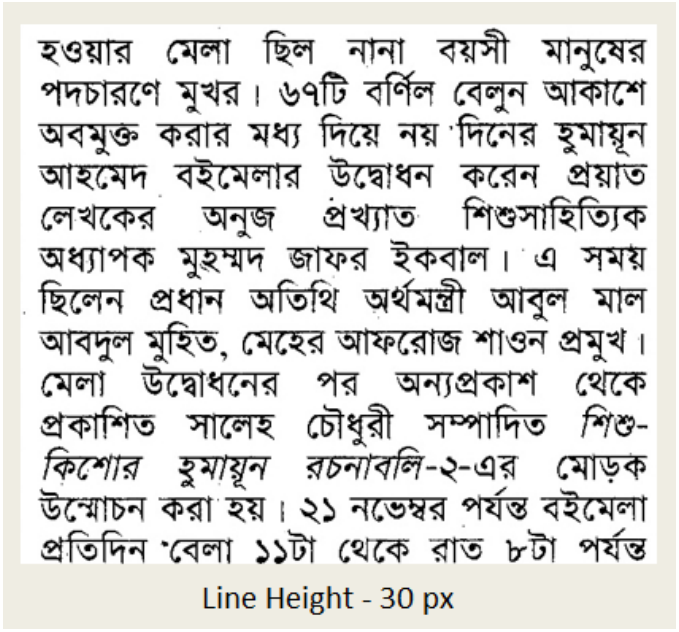


Fig. 5. Result of applying the methodology.

V. CONCLUSION

To evaluate our methodology, we worked only with news paper.Although this evaluation was done with sufficient data set, it wasn't test with different type of documents. So this methodology is currently now works only with documents like newspaper. How ever further research on this work with different type of document, may lead to more perfect methodology.

ACKNOWLEDGMENT

This paper is made possible through the help and support from all sentient beings. Especially, please allow us to dedicate my acknowledgment of gratitude towards our honorable Professor Prof. Muhammed Zafar Iqbal. Throughout their guidance, encouragement and support we were able to continue our works fluently.

REFERENCES

[1] P. Saikrishna and A. G. Ramakrishnan, "Script independent detection of bold words in multi font-size documents," In proceeding of National Conference on Computer Vision, Pattern Recognition Image Processing and Graphics (NCVPRIPG-2013), 2013.

[2] D. S. B. F. R. Chen and L. D.Wilcox, "Detection and location of multicharacter sequences in lines of imaged text," *Journal of Electonic Imaging*, vol. 5, pp. 37 – 49, January 1996.

[3] D. B. K. R. McConnell and R. Schaphorst, *FAX: Digital Facsimile Technology and Applications.* Artech House, 1989.

[4] P. N. Mohammed Javed and B. Chaudhuri, "Automatic detection of font size straight from run length compressed text documents," vol. 5 (1), pp. 818–825, International Journal of Computer Science and Information Technologies, 2014.

[5] A. Z. Laurence Likforman-Sulem and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. vol. 9, no. 2-4, p. 123 138, January 2007.

[6] Y. Lu and C. L. Tan, "Word searching in ccitt group 4 compressed document images," p. 467471, Proceedings of Seventh International Conference on Document Analysis and Recognition (ICDAR03), 2003.

[7] J. Lu and D. Jiang, "Survey on the technology of image processing based on dct compressed domain," p. 786789, International Conference on Multimedia Technology (ICMT), 2011.

[8] P. N. M. Javed and B. Chaudhuri, "Extraction of projection profile, run-histogram and entropy features straight from run-length compressed documents," p. 813817, Proceedings of Second IAPR Asian Conference on Pattern Recognition (ACPR13), Okinawa, Japan, November 2013.