Script Independent Detection of Bold Words in Multi Font-size Documents

Pedamalli Saikrishna
Dept. of Electrical Engg.
Indian Institute of Science,
Bangalore, 560012, India
Email: sk.iitg@gmail.com

A.G.Ramakrishnan
Dept. of Electrical Engg.
Indian Institute of Science,
Bangalore, 560012, India
Email: ramkiag@ee.iisc.ernet.in

***Abstract*—A script independent, font-size independent scheme is proposed for detecting bold words in printed pages. In OCR applications such as minor modifications of an existing printed form, it is desirable to reproduce the font size and characteristics such as bold, and italics in the OCR recognized document. In this morphological opening based detection of bold (MOBDoB) method, the binarized image is segmented into sub-images with uniform font sizes, using the word height information. Rough estimation of the stroke widths of characters in each sub-image is obtained from the density. Each sub-image is then opened with a square structuring element of size determined by the respective stroke width. The union of all the opened sub-images is used to determine the locations of the bold words. Extracting all such words from the binarized image gives the final image. A minimum of 98 % of bold words were detected from a total of 65 Tamil, Kannada and English pages and the false alarm rate is less than 0.4 %.**

I. Introduction

When we want to make minor changes to a printed document such as an application form, we can OCR the form, make those modifications and again print the same. In such circumstances, it is important that the OCR recognizes the different font sizes for the title, sub-heading, etc. and reproduce the same in the output text document. Similarly, it s desirable that bold or italic text in the original document is reproduced as it is in the output OCR text. This requirement calls for detection of bold text as against normal text, and this detection must be independent of the font size. This aspect of an OCR is distinct from the need to recognize the text independent of the font type, size, italicization or bold face.

In this paper, we propose a script and font size independent method to retrieve the information about the positions of the bold words in the given page of text using elegant morphological image processing techniques.

The rest of the paper is organized as follows. Section II discusses the various methods existing in the literature. Section III discusses the proposed method followed by the experimental results in section IV. Section V presents the conclusion.

II. Existing Methods for Detection of Bold Text

Very few researchers have reported work on detecting the bold text in a page. A brief overview of these methods is given here. Two of the three methods discussed in this section make use of morphological image processing techniques, showing their usefulness to detect bold text.

Bloomberg [1] used various morphological image processing techniques to detect bold text. The binarized image is thinned progressively, removing pixels alternatively from both the ends of the stems of a character. Each thinning operation is followed by opening the thinned image using a vertical structuring element. This leads to the reduction in the number of ”ON” pixels present in the image after each iteration. The ratio between the number of ”ON” pixels in the present to the previous iteration is calculated after each iteration. If this ratio is below 40%, the process is stopped and the resultant image is used to determine the positions of the bold text in the given page.

Doermann et. al [2] make use of erosion and opening to perform the task. The binarized image is eroded and opened alternatively till the total number of ON pixels in the result is less than 20% of that in the binarized image. The resultant image is then used to determine the positions of the bold text in the given page.

Unlike the previous methods, Chaudhuri and Garain [3] did not use morphological image processing techniques. The run lengths of the black pixels (ON in this case) is obtained in different directions for every character. The minimum of these is considered to be the stroke width. As the stroke width depends on the height of the character, the stroke width and the height of the character are compared to determine the position of the bold characters. This method also detects if the character is italic or not, type of font and whether the character is capital or not. However, the emphasis is only on Roman script and no results are given for any other script.

Since our aim is to develop a method applicable to all scripts, we make use of morphological image processing techniques. However, the existing morphological methods have a few drawbacks. The methods do not consider the font size information while determining the bold text. This results in misclassification of higher font size non-bold text as bold text and missed detection of bold text of lower font size. This provided us with the scope to improve on these methods, resulting in the method reported here.

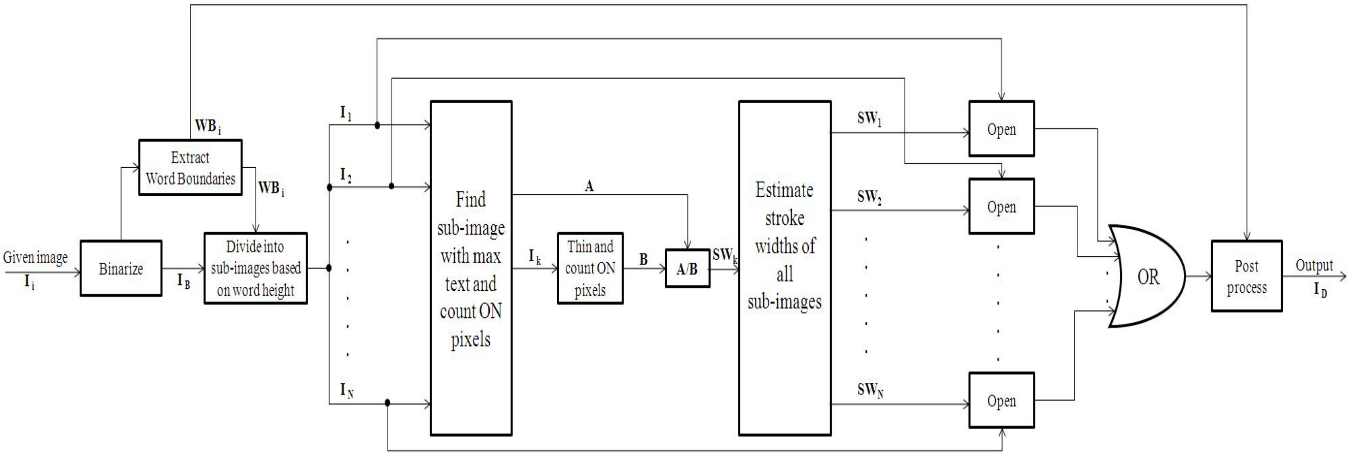


Fig. 1. Block diagram of the proposed MOBDoB method .

III. MOBDoB Algorithm

Figure 1 shows the block schematic of the proposed morphological opening based detection of bold (MOBDoB) algorithm. The algorithm starts with binarized images from which all the graphics have already been removed and thus contain only text.

A. *Detect Distinct Font Sizes*

From the binarized image, the word boundaries are obtained. Based on their heights, the word images in the document page are clustered into distinct sub-images. The words in each resulting cluster have nearly uniform font sizes. However, any cluster may contain multiple disjoint regions, depending upon the nature of the document. If we obtain N clusters, it means that the image contains text of at least N distinct font sizes. The image is divided into sub-images such that the following two conditions are met:

$$\bigcup_{s=1}^{N} I_s = Text\, in\, Binarized\, image \tag{1}$$

$$I_i \cap I_j = \phi \begin{cases} i, j \in 1, 2, ..., N \\ and \\ i \neq j \end{cases} \tag{2}$$

B. *Stroke width correlates from thinning*

The total number of words present in each sub-image is counted. The sub-image (I_k) with the highest number of words is selected to obtain a rough estimate of the stroke density, which is a correlate of the stroke width (SW_k) of the characters of the principal font size. Stroke density is estimated as the ratio of the total number of "ON" pixels in this sub-image (A) to that in the completely thinned (thinned to single pixel width characters) sub-image (B). Using this value, the stroke widths of other sub-images are estimated. This can be done because, for the same font type, the stroke density is nearly correlated to the font size, which can be approximated by the word height. The equation that provides the respective stroke widths for other sub-images is given below.

$$SW_l = \tfrac{H_l}{H_k} \times (SW_k), \quad l \in 1, 2, ..., N \tag{3}$$

where, H_l and H_k are the minimum heights of the words excluding the ascenders and descenders in sub-images I_l and I_k, respectively. The estimated stroke widths are rounded off to the nearest integers.

C. *Detect Bold By Morphological Opening*

The stroke width of the bold characters is higher than that of the normal characters. Opening a sub-image using an appropriate structuring element removes the normal text while retaining some parts of the bold text. The dimension of the structuring element is determined by the stroke width value of each sub-image. Every sub-image (I_s) is opened by a square structuring element of size $((SW_s) + 1) \times ((SW_s) + 1)$. A resultant image is formed by "OR"ing all the opened sub-images.

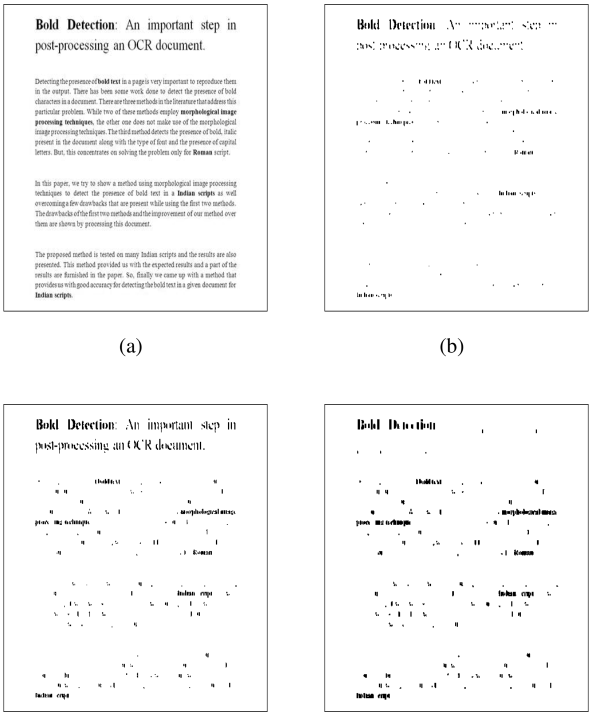
$$I_R = \bigcup_{s=1}^{N} I_s^o \tag{4}$$

D. *Extract Bold Words*

In the ideal case, the resultant image provides us with the positions of bold characters alone in the given page. In practice, there are a few places in a few characters, where the stroke width is considerably larger and is comparable with the bold character's stroke width. In many fonts of most scripts, the stroke width of the bold characters is not uniform. This may result in loss of a part of the bold character. To remove this undesired presence of the "ON" pixels at the positions of the normal characters and effectively retain the parts of bold characters, we post-process the resultant image.

In post-processing, the ratio of number of "ON" pixels in the resultant image to that in the binarized image is calculated for each word using the word bounding box information (coordinates). The ratio is then thresholded to determine whether the word under consideration is bold or not. The words detected as bold are retained and others are removed, thus resulting in a bold-word-only image.

IV. Results and Discussion



Bold Detection: An important step in post-processing an OCR document.

Detecting the presence of **bold text** in a page is very important to reproduce them in the output. There has been some work done to detect the presence of bold characters in a document. There are three methods in the literature that address this particular problem. While two of these methods employ **morphological image processing techniques**, the other one does not make use of the morphological image processing techniques. The third method detects the presence of bold, italic present in the document along with the type of font and the presence of capital letters. But, this concentrates on solving the problem only for **Roman** script.

In this paper, we try to show a method using morphological image processing techniques to detect the presence of bold text in a **Indian scripts** as well overcoming a few drawbacks that are present while using the first two methods. The drawbacks of the first two methods and the improvement of our method over them are shown by processing this document.

The proposed method is tested on many Indian scripts and the results are also presented. This method provided us with the expected results and a part of the results are furnished in the paper. So, finally we came up with a method that provides us with good accuracy for detecting the bold text in a given document for **Indian scripts**.

(a) (b)

(c) (d)

Fig. 2. Performance comparison with other techniques in the literature. (a) Original image. (b) Result after opening using [1]; a few bold words of lower font size are missed. (c) Result after opening using [2]; normal words of large font size text are misrecognized as bold. (d) Intermediate result of proposed method, before postprocessing - union of opened sub-images of different font sizes; bold words of all the font sizes are properly detected.

The MOBDoB algorithm has been tested on 65 images of printed pages from Tamil, Kannada and English languages. The pages are selected in such a way that 50% of the pages in each language contain a single font size with a part of the text being bold, 30% pages have multiple font sizes with bold characters and remaining pages contain no bold text at all in the entire page.

The proposed method is compared with the two other morphological methods in the literature and the results for all the three algorithms are shown for a sample image in Fig 2. It can be seen from the figure that other methods detect the normal font of large font size text to be bold, while our method handles it appropriately.

Figure 3 shows the results of some key intermediate steps involved in the proposed method. In this figure, the binarized image is divided into two sub-images. Using sub-image 1 and its thinned image, the estimated stroke width is 3.8 and the word height is 39. The word height of sub-image 2 is 70. By using (3), the SW of sub-image 2 is estimated to be 6.8. Rounding these stroke widths, we obtain $SW_1 = 4$ and $SW_2 = 7$. Therefore, the sub-images 1 and 2 are opened using square structuring elements of sizes 5×5 and 8×8, respectively. The union of the opened sub-images is checked against the word boundaries obtained from the binarized image. In any word boundary, if more than one-third of the ”ON” pixels of binarized image is retained in the resultant image, then the word is detected as bold; otherwise, the word is declared to be normal. Figure 4 shows the results of MOBDoB technique on a segment of the image of a Kannada printed page.

The detection results are analyzed in terms of the number of words detected as bold or normal. The proposed method provided desirable results for all the scripts tested. The performance for different scripts is tabulated in Table I. The sixth and eighth columns in Table I correspond to false alarms and missed detections of bold words, respectively. From the figures and the table, it can be seen that the proposed method performs well irrespective of the script present in the image under test. In addition to the above scripts, we tested a few documents of Hindi, Malayalam, Telugu and Bengali scripts and found that similar results are obtained. However, since only a few pages were tested, those results are not furnished in Table I. Since Bloomberg [1] and Doermann et al [2] do not provide quantitative results of their methods in terms of the percentage of words correctly detected as bold or otherwise, we are not in a position to quantitatively compare our method with theirs.

V. Conclusion

This paper proposed an elegant method to determine the positions at which bold text is present in a given page. The method, based only on morphological processing steps, is able to effectively detect bold words with different font sizes in the same page. The results clearly demonstrate that the technique is script independent at least with respect to the seven scripts tested. Also, the accuracies obtained over the pages tested is very good across all the scripts.

VI. Acknowledgment

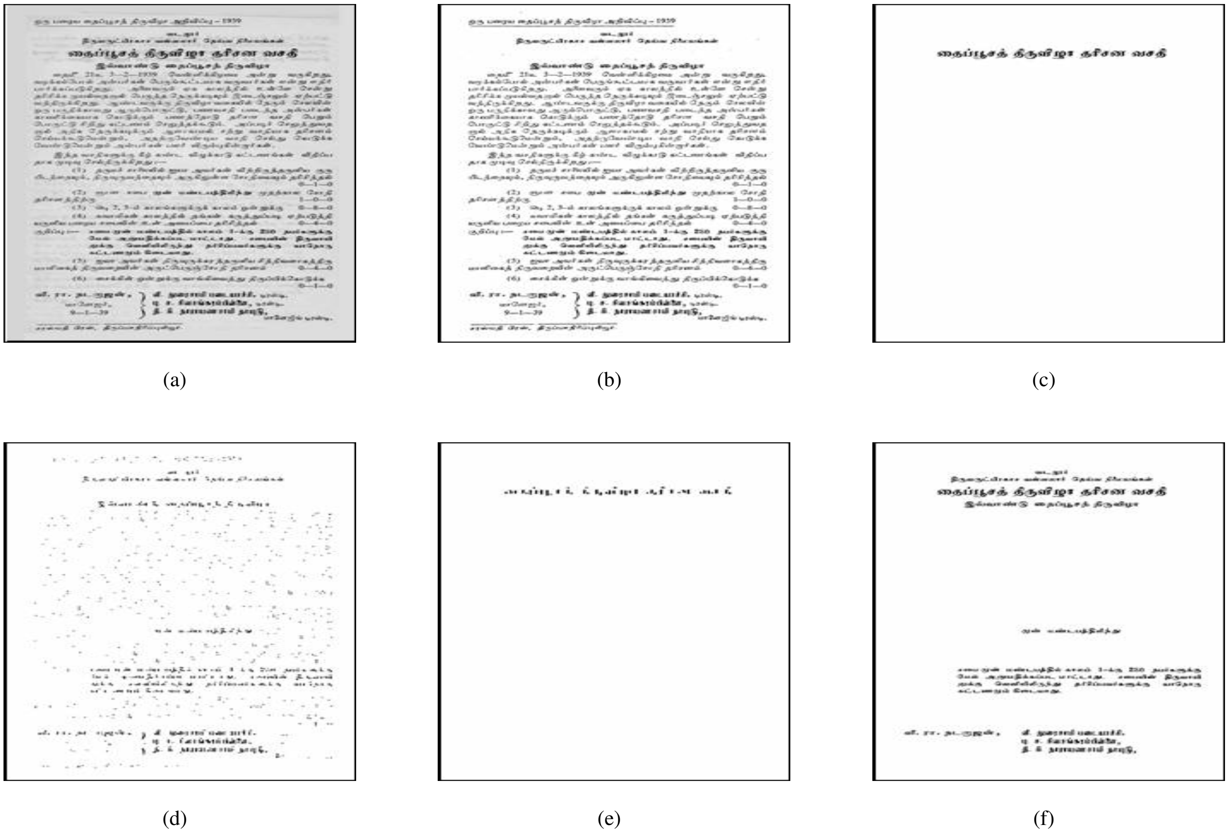
The authors acknowledge Technology development for Indian languages (TDIL), Department of information technology, Government of India for funding this research under the consortium project 11(19)/2010-HCC(TDIL) titled, ”Development of robust document image understanding system for documents in Indian scripts - Phase II”.

References

[1] Dan S. Bloomberg ”Multiresolution morphological analysis of document images,” *SPIE Visual Communications and Image Processing*, 1818, 648-662, 1992.

[2] Doermann, D.; Rosenfeld, A.; Rivlin, E.; , ”The function of documents,” *International Conference on Document Analysis and Recognition*, vol.2, pp.1077-1081, 1997.

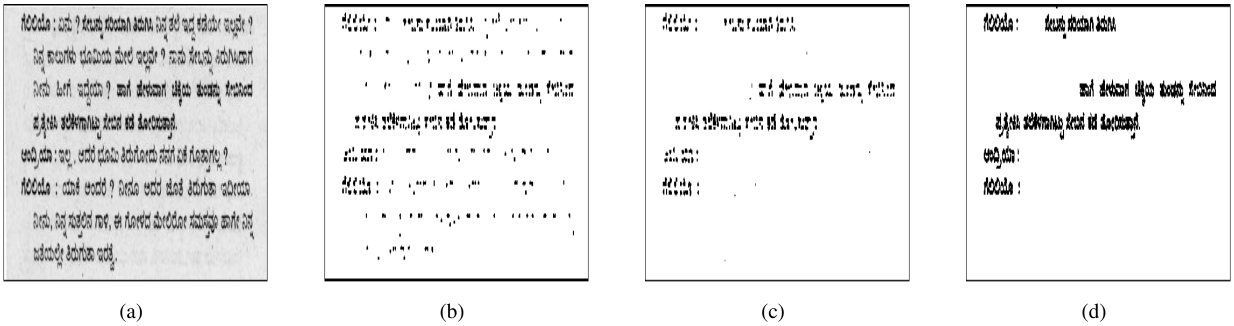
[3] Chaudhuri, B.B.; Garain, U.; , ”Automatic detection of italic, bold and all-capital words in document images,” International Conference on Pattern Recognition, vol.1, pp.610-612, 1998



(a) (b) (c)

(d) (e) (f)

Fig. 3. Performance of the proposed method on a page of Tamil text (a) Original image, (b) Binarized Sub-image 1 of lower font size, (c) Binarized Sub-image 2 of higher font size, (d) Opened sub-image 1, (e) Opened sub-image 2 and (f) Image with the bold words extracted.



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

Fig. 4. Performance of the proposed method on a segment of a page of Kannada text (a) Original Image, (b) Opened Image, (c) Detected positions of the bold words and (d) Image with the bold words extracted.

TABLE I
BOLD WORD DETECTION PERFORMANCE OF MOBDOB METHOD ON DIFFERENT SCRIPTS.

Script	No. of pages	Words present		Classified as				Overall Performance	
		Bold	Normal	Bold		Normal			
				Correct	False Alarms	Correct	Missed Detections	Correct	Wrong
Tamil	30	300	5200	295 (98.33 %)	19 (0.37 %)	5181 (99.63 %)	5 (1.67 %)	5476 (99.56 %)	24 (0.44 %)
Kannada	30	713	4927	698 (97.9 %)	15 (0.3 %)	4912 (99.7 %)	15 (2.1 %)	5610 (99.47 %)	30 (0.53 %)
Roman	5	142	929	141 (99.3 %)	2 (0.22 %)	927 (99.78 %)	1 (0.7 %)	1068 (99.72 %)	3 (0.28 %)