Foreground Text Extraction in Color Document Images for Enhanced Readability

S. Nirmala and P. Nagabhushan

Dept of Studies in Computer Science, University of Mysore, Mysore-570 006, India
nir_shiv_2002@yahoo.co.in, pnagabhushan@compsci.uni-mysore.ac.in

Abstract. Quite often it is observed that text information in documents is printed on colorful complex background. Smooth reading of text content in such documents is difficult due to background patterns and mix up of foreground text color with background color. Further the character recognition rate when such documents are OCRed, is low. In this paper we are presenting a novel approach for extraction of text information in complex color document images. The proposed approach is a three stage process. In the first stage the edge map is obtained utilizing the Canny edge operator. The edge map is split into blocks of uniform size and image blocks are classified as text or non-text. In each text block the possible text regions are identified and enclosed in tight bounding boxes using x-y cut on edge pixels. Further the text regions that are immediate adjacent to each other in vertical direction in which the character(s) are split horizontally are merged so as to enclose the character(s) fully in one text region. In the second stage certain amount of false text regions are eliminated based on a property of printed text. In the last stage the foreground text in each text region is extracted by unsupervised thresholding using the data of refined text regions. We conducted exhaustive experiments on documents having variety of background complexities with printed foreground text in any color, font and tilt. The experimental evaluations show that on an average 98.03% of text is identified. The processed document images showed better performance when OCRed compared with the corresponding unprocessed source document images.

Keywords: Color document image, Complex background, Foreground Text extraction, Text region detection, Unsupervised thresholding, OCR.

1 Introduction

Often we find many documents that are designed deliberately with colorful and complex background for instance news paper articles, advertisements, magazine pages. Background patterns, high level variation of background color(s), combination of foreground text color and background color cause non smooth readability of the document contents. Further, automatic OCRing of such documents result in low recognition accuracy. In past many efforts were reported on separation of foreground from background of document images [1]-[5]. Thresholding is a simple and effective method of isolation of foreground from background of a

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 387–392, 2009.

document [1]. In [3] the performance of five popular local thresholding methods on four types of 'difficult' document images is evaluated and it is reported that no single algorithm works well for all types of images. Most of the thresholding methods are based on the apriori knowledge on foreground and background intensity. Practically it is not possible to know the polarities of foreground and background intensities which call for a specialized binarization technique. In [5], a specialized binarization method is proposed to extract the characters in color document images. Text-regions in a document image can be detected either by connected component analysis [2] or by texture analysis method [6].The connected component based methods detect text at faster rate but are not very robust for text localization. Also they result in false text regions for images with complex background. On the other hand texture based methods [6]are robust in detecting the text regions but they are very expensive.Most of the works discussed earlier on separation of foreground in color/gray document images show some serious shortcomings and impossible to apply on documents with complex color background with foreground text in any color and tilt. To overcome the above drawbacks, in this work we propose a three stage novel approach to extract the foreground text in complex color document images. The rest of the paper is organized as follows. Section 2 introduces the proposed approach. Experimental results are provided in section 3.Conclusions drawn from the current study are summarized in section 4.

2 Proposed Method

In this paper we propose a novel approach for extraction of printed foreground text in complex color documents which are of low resolution and scanner based. Fig.1 shows the block diagram of the proposed method. The proposed method is based on a property of printed characters that they form the edges against the background due to high intensity gradient.

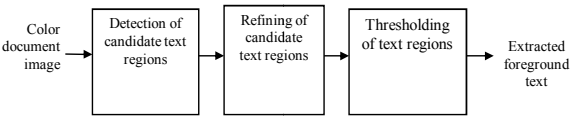

Fig. 1. Block diagram of proposed method

2.1 Detection of Candidate Text Regions

As Canny edge detector has low probability of missing an edge, we propose Canny edge operator for detection of text edge pixels. We have set the threshold values 0.3 and 0.4 for the hysteresis thresholding step of Canny edge detection. To avoid loss of text edges, edge detection is carried out in each color channel of RGB color model using Canny edge operator [4]. The final edge map is

formed by assimilating the results of edge detection in all the three color components [4]. Suppose E_R, E_G and E_B are the edge images of red, green and blue components the final edge map 'E' is given by, E=$E_R \vee E_G \vee E_B$, where '$\vee$'represents logical 'OR' operator. As the resolution of the image is very low the broken edges are connected using 'imclose' operation with a structuring element in vertical direction. We conducted experiments to set the size of the structuring element. Structuring element size = 4 pixels result into maximum reduction of false text regions without loss of true text regions in stage-2 of the proposed method. With structuring element size $<$ 4 pixels result in high reduction of false text regions but certain percent of true text regions are lost. Hence structuring element size is empirically set to 4 pixels. The modified edge map is split into blocks of uniform size. The blocks in the edge map that do not contain even a single edge pixel are classified as non-text blocks and these non-text blocks are ignored as they compose only the background pixels. The blocks that contain at least one edge pixel is considered as candidate text block. In each candidate text block the text regions are enclosed in tight bounding boxes by performing x-y cut on edge pixels. We identify the text regions which are originally placed in adjacent text blocks in vertical direction and merge those adjacent text regions in which the horizontal split of character strings appear. Fig 2. shows the output of sub processes of stage-1 in sequence for an initial block of size 25 $\times$ 25.

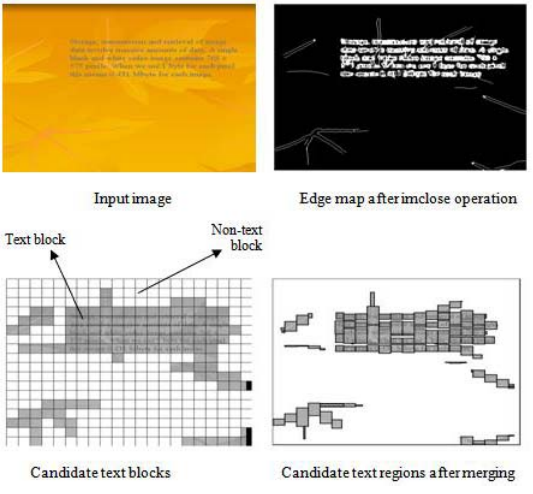

Fig. 2. Output of sub processes in stage-1

2.2 Refining of Candidate Text Regions

Due to high intensity of some background objects the edges of background objects might be detected by Canny edge operator. The candidate text regions that

contain only the edges of background objects (false text region) are identified and eliminated in stage-2. To develop a criterion to identify the false text region the following feature of text is observed [6]. The fact we observed is that the number of text edge pixels is more in true text region compared to false text region. Suppose 'W' and 'H' are the width and height of the text region. If the text edge pixel count in a text region is greater than maximum(2 * W, 2 * H) we classified it as true text region else as false text region. The false text regions are removed in this stage and true text regions are considered for further processing which is described in the subsection that follows.

2.3 Thresholding of Text Regions

The true text regions which are obtained from stage-2 are thresholded locally to extract the foreground pixels and deposited in proper position on a uniform white background. We considered the gray scale equivalent of the corresponding text region for thresholding the text region. A specialized unsupervised thresholding is designed based on foreground pixel intensity and background pixel intensity and deposited foreground characters in black on uniform white background. The approximate background intensity is computed by averaging the intensity values of non edge pixels in the updated edge map. For each text region the approximate foreground intensity is computed by averaging the intensity values of edge pixels of updated edge map. Threshold value 'Th' for each text region is computed as follows:

If (abs(average foreground pixel intensity - average background pixel intensity) > 40)
Th=average foreground intensity
Else
Th= 0.5 * (average foreground pixel intensity + average background pixel intensity).

Fig. 3. shows the output images obtained from stage-2 and stage-3.

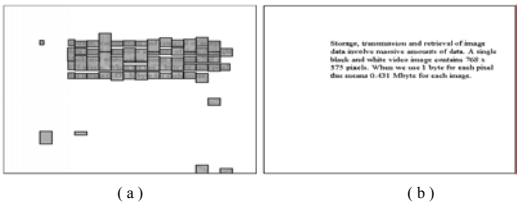

Fig. 3. (a) Output image from stage-2 (b) output image from stage-3

3 Experimental Results

Since no standard corpus of images is available for this work we created our own corpus of printed color document images by scanning the documents from

various sources viz. magazines, story books, postal envelopes and newspapers. In addition we created another corpus of synthesized images. All the images in both the corpus are of low resolution. Irrespective of the foreground font color the output image is created by depositing black characters on white background. The performance of amount of text detection is evaluated in terms of Recall ((correct detects / (correct detects + missed detects)) and Precision (correct detects / (correct detects + false alarms)). Table 1. shows the average value of Precision and Recall in percentage for document images in the corpus.

Table 1. Performance evaluation of text detection

No. of samples	Total number of characters	Recall(%)	Precision(%)
160	31633	98.03	97.80

Table 2. OCR results

Initial Block size	OCR Recognition rate (before processing)	OCR Recognition rate (after processing)
25×25	56.25	66.84
50×50	56.25	62.97
75×75	56.25	58.29

(a) (b)

Fig. 4. Result of sample a document image with foreground text in different font, color and tilt : (a) Input , (b) Output

Reading of the extracted text is evaluated on Readiris 10.04 pro OCR. Readiris 10.04 pro OCR handles color /gray document images. In this work readability of the extracted foreground text is evaluated in terms of character recognition rate. Table 2. shows the average character recognition rate by OCR before and after applying the proposed approach. Although the average performance after processing appears to be around 66% it should be noted that in some specific difficult cases the recognition rate drastically improved to nearly 100% (after processing) from recognition rate of 0% (before processing).It is also observed that the better performance is with initial block of size 25×25 which is evident from table 2. Fig. 4. shows result of a sample document image with foreground text in different color, font and tilt.

4 Conclusion

In this paper a novel approach is presented for extraction of foreground text from complex background in color document images which are of low resolution and scanner based. The candidate text blocks are identified based on a characteristic property of the printed characters. Identified text regions in each block are enclosed in tight bounding boxes. By designing a criterion based on text pixel count in a text region the false text regions are filtered out. An unsupervised thresholding is devised to extract the foreground text in the refined text regions. The proposed approach detects 98.03% of text content in the source document. We achieved 66.84% of OCR character recognition accuracy. The proposed approach fails to extract the characters that are too thick. Devising an effective criterion for elimination of false text regions so as to improve the recall rate and designing a thresholding technique to extract the foreground characters to improve the character recognition accuracy by OCR are considered as future works of the current study.

References

1. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic imaging 13, 146–165 (2004)
2. Pietikäinen, M., Okun, O.: Text extraction from grey scale page images by simple edge detectors. In: Proceedings of the 12th Scandinavian Conference on Image Analysis, SCIA, Norway, pp. 628–635 (2001)
3. Leedham, G., Chen, Y., Takru, K., Tan, J.H.N., Mian, L.: Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In: Proceedings of seventh International Conf. on Document Analysis and Recognition (ICDAR), pp. 859–864 (2003)
4. Shivananda, N., Nagabhushan, P.: Separation of Foreground Text from Complex Background in Color Document Images. In: Proceedings of Seventh international conference on advances in pattern recognition, ISI Kolkata, pp. 306–309 (2009)
5. Kasar, T., Kumar, J., Ramakrishnan, A.G.: Font and Background Color Independent Text Binarization. In: Proceedings of 2nd Intl. workshop on Camera Based Document Analysis and Recognition (workshop of CBDAR), pp. 3–9 (2007)
6. Liu, Y., Goto, S., Ikenaga, T.: A contour based robust algorithm for text detection in color images. IEICE Transactions on Information and Systems 89, 1221–1230 (2006)