Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Combined Edge-Based Text Region Extraction from Document Images

C.P.Sumathi*
Department of Computer Science
SDNB Vaishnav College for Women
Chennai, India

N.Priya
Department of Computer Science
SDNB Vaishnav College for Women
Chennai, India

Abstract— Image archival and retrieval systems effectively archive and help in retrieving the documents for content-based retrieval systems, image search engines and script recognition systems. This paper proposes a combined edge-based technique for separating text and non-text regions in a document image. The maximum magnitude of the edge is detected by using the compass masks filtering convolution in eight major directions. Successively, in the localization process the magnitude of the edge can be compared with a threshold value to generate the edge map. Then, morphological operators are applied to find the regions of the non-text connected component. Finally, a statistical feature analysis of the text region is performed and extracted from the background of the image so that a sensible reading is provided for the OCR system. This combined algorithm has been tested on a large set of document images and the result seems to be better when compared to the existing techniques.

Keywords— Compass mask, Threshold, Morphological Operators, Statistical Measures, Text extraction

I. INTRODUCTION

In recent years, document imaging is the conversion of paper documents into electronic images instead of paper filing systems which leads to many applications such as document processing, content-based retrieval systems, image indexing and archiving documents [1]. Text content extraction in document images is a challenging problem due to differences in size, style, orientation, and alignment, as well as low image contrast and complex background. Many works have been developed for text extraction in document images since many years. But still, to find a completely robust and generalized technique for text segmentation, it is difficult to provide appropriate input to the optical character recognition (OCR) system. In the present paper, an efficient combined method has been proposed which is independent of the orientation of the text, background of the image as well as font size of the text.

II. PREVIOUS WORK

Traditionally, text extraction from document images is divided into three categories: top-down, bottom-up and hybrid approaches [2]. Bottom-up techniques [3][4]: Recursively merging and grouping from the characters to words and then to text lines to paragraphs and so on. They are usually based on connected components analysis which does not make assumptions on font style and size and about page layout. This technique is mostly expensive on the use of appropriate thresholds. Most popular bottom-up techniques are mathematical morphology, run length smoothing algorithm, and region growing based methods. Top down Techniques [5]: It starts by detecting the highest level of structures and proceeds by successive splitting until they reach the bottom layer for small scale features. Priori knowledge about the page layout is necessary for this technique. But it is faster than bottom-up techniques and efficient for special purposes (e.g., making all issues of a specific journal into digital format), but they are not suitable for more general purposes. The most well known methods are projection methods, histogram analysis, rule based systems, or space transforms.

Hybrid Techniques: Most of the works do not really fit into any of the two categories mentioned above and hence they are called as hybrids. Examples of these are those based on Gabor filtering and mask convolution, fractal signature and wavelet analysis. The review of literature shows that, many methods for the detection of text regions are based on the hybrid technique. Peeta et.al [6] proposed two texture-based approaches Gabor filters and log-polar wavelets, for separating text from non-text elements in a document image. Both the algorithms compute local energy at some information-rich points, which are marked by Harris' corner detector and compared. It was observed that the Gabor filter based scheme marginally outperforms the wavelet based scheme. Parodi and Piccioli [7] fixed two parameters, the minimum width and precision of the text to be detected. The method works by subdividing the page into overlapping columns whose width and inter-shift depend on width and precision and by performing text lines extraction on each column separately. Successively, a statistical analysis of the text line elements found in each column is performed, and they are connected to form complete text lines. Vijaya and Padma [8] proposed a system based on the characteristic features of top-profile and bottom-profile of individual text lines of the input document image and the features were extracted by finding the behavior of the characteristics of the top and bottom profiles of individual text lines in three different Indian languages. Sachin et.al [9] proposed a simple edge based feature to detect the text from colored

document features. The algorithm is based on the sharp edges of the characters which are missing in images. They find those edges and use them to classify text from images.

These observations shows that, most of the methods fall into any one of the above techniques and also that there is a limitation in each method to give a better detection rate with fewer false alarms without any constraints for text region extraction in document images. Hence, the proposed method tends to provide an efficient and effective combined edge-based approach to extract the text region for a wider range of document images. This method comprises several phases such as

i) detection of magnitude of the edges by using the Compass mask convolution filtering.
ii) creating the edge map from soft thresholding technique.
iii) finding the connected component of the non-text regions using morphological operators
iv) extracting the features of text using statistical measures
v) removing the non-text regions from the document image.

Fig. 1 shows the flow of the complete process of the proposed technique and the rest of the paper is organized as follows. In section 3, the methodology of text region extraction technique is described. Result Analysis and comparison of [6] and [7] with proposed algorithm are presented in section 4 followed by conclusion in Section 5.

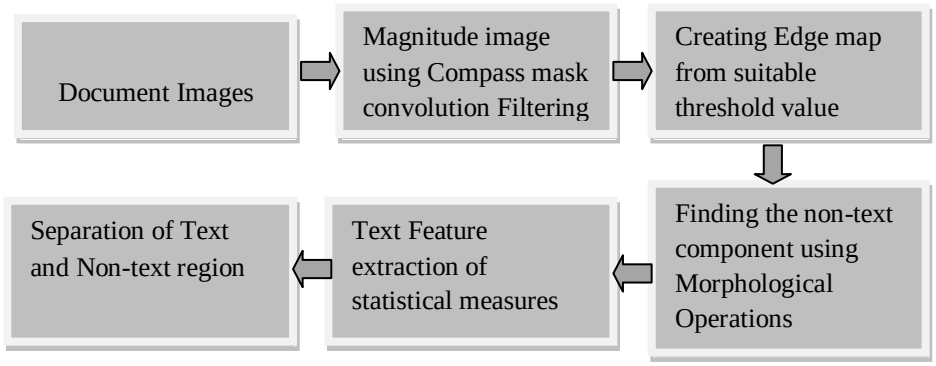


Fig. 1 Flow of Proposed Technique

III. METHODOLOGY

The proposed approach for text content extraction in a document image can be divided into following phases:

A. Phase I

Edges are considered as a very important portion of the perceptual information content in a document image. It's a local concept which represents the significant intensity variations, discontinuities in depth, surface orientation, change in material properties, and light variations. An edge is typically extracted by computing the derivative of the image. This consists of two parts-magnitude of the derivative, which is an indication of the strength/contrast of the edge, and the direction of the derivative vector, which is a measure of edge orientation. The idea is to detect the edges to filter the input gray scale image to get the maximum magnitude of the image. Compass mask is the template matching filters used to perform directional smoothing as they are very sensitive to directions and detect edges in all directions. And also it is a second degree gradient operator and is much more aggressive in enhancing sharp changes. Thus it is best suited to find the sharp edges of the text. By rotating the convolution filter mask in all the eight directions N, NW, W, SW, S, SE, E and NE and if there is a match between the directions and the masks then a maximum gradient magnitude value is produced. The respective masks are given as

$$R_0=\begin{pmatrix}-1&0&1\\-2&0&2\\-1&0&1\end{pmatrix}\quad R_1\begin{pmatrix}0&1&2\\-1&0&1\\-2&-1&0\end{pmatrix}\quad R_2=\begin{pmatrix}1&2&1\\0&0&0\\-1&-2&-1\end{pmatrix}\quad \mathrm{R}_3=\begin{pmatrix}2&1&0\\1&0&-1\\0&-1&-2\end{pmatrix}\quad R_4=\begin{pmatrix}1&0&-1\\2&0&-2\\1&0&-1\end{pmatrix}\quad \mathrm{R}_5=\begin{pmatrix}0&-1&-2\\1&0&-1\\2&1&0\end{pmatrix}$$

$$R_6=\begin{pmatrix}-1&-2&-1\\0&0&0\\1&2&1\end{pmatrix}\quad R_7=\begin{pmatrix}-2&-1&0\\-1&0&1\\0&1&2\end{pmatrix}$$

So the given input image has to be convolved with a mask that is obtained by rotating in eight directions in the range of 0^0 - 315^o in steps of 45^0. Using this mask the appropriate magnitude is given by

$$G(i,j)=\sqrt{G_x^2(i,j)+G_y^2(i,j)} \qquad (1)$$

for a pixel at (i,j). $G_x(i,j)$ and $G_y(i,j)$ are the x- and y- components of the gradient magnitude.

B. Phase II

In this phase, the detected edges are localized and stored in the edge map. The localization process involves determining the exact location of the edge and also takes binary decision whether an image pixel is edge or not. In addition, this stage involves edge thinning and edge linking steps to ensure that the edge is sharp and connected. A prerequisite for the localization stage is normalization of the gradient magnitude. The calculated gradient can be scaled to a specific range say, 0-K by performing this operation. The simplest method is to apply a threshold operation to the edge strength delivered by a compass edge operator using either a fixed or adaptive threshold value, which results in a binary edge image or "edge map". This edge map is stored for further phase operations. $N(x, y)$ is called the normalized edge image and is given as

$$N(x, y) = \frac{G(x, y)}{\max_{i=1..n,\, j=1..n} G(i, j)} \times K \tag{2}$$

The normalized magnitude can be compared with a threshold value T to generate the edge map. The edge map is given as

$$E(x, y) = \begin{cases} 1 & if \quad N(x, y) > 1 \\ 0 & otherwise \end{cases} \tag{3}$$

C. Phase III

A morphological process is intended in this phase to find the connected non-text region from a specific text/background image. Morphology is a very powerful tool for analysing the shapes of the objects present in images. Morphological operations are very effective in the detection of boundary, removal of noise; identify components, convex hull and so on. The most basic morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. These two operations can be combined with opening and closing operations for boundary detection. Opening operation is used to smoothen the inner object contour to break narrow strips and eliminate thin portions of the image. It is also used to remove noise. Closing operation fills the small holes and gaps in a single-pixel object. These operators take a binary image and a mask known as structuring element as input and use it to remove the noise from an image or select objects in a particular direction.

Let $S_{m,n}$ denote a structure element with the size m x n, where m and n are odds in size and larger than zero and $I_{x,y}$ denote a gray-level input image. According to the definition of $S_{m,n}$ some morphological operations such as erosion, closing, opening are mathematically represented as follows

Erosion operation:

$$I(x, y) \Theta S_{m,n} = \min_{|i| \le m/2, |j| \le n/2} I(x - i, y - j) S_{m,n}(i, j), \tag{4}$$

Closing operation:

$$I(x, y) \bullet S_{m,n} = (I(x, y) \oplus S_{m,n}) \Theta S_{m,n}, \tag{5}$$

Opening operation:

$$I(x, y) \circ S_{m,n} = (I(x, y) \Theta S_{m,n}) \oplus S_{m,n}, \tag{6}$$

In the result image after the previous phase, the text and non-text regions are extracted together. To remove the non-text from the resultant image, the size and shape of that region is compared with the text region. Connected components of the non-text whose shapes are obviously different from the text-like connected components and their areas of the text connected component are relatively small comparing to the areas of their non-text. Therefore, they can be easily filled by the above opening, closing and erosion operations. Finally, a fill-hole process is performed to deal with the non-text regions and remove the text regions from a document image.

D. Phase IV

After the previous phase, some portions of the text objects may be merged with the background or portions of the background may appear as a text object. To set this problem, the text consistency test which includes statistical features is performed to distinguish the text regions exactly from the background of the image. Statistical measures mean and standard deviation determines the intensity and contrast values among the text and non-text.

Each connected component consists of a range of some pixel values. These values in each component can be used to calculate the mean and standard deviation which represents the brightness of the intensity and contrast of those pixels. If mean and standard deviation value of a component is high then it means that the component is bright and high contrast in that image and if mean and standard deviation value is low then it means that the component is dark and low contrast in that image. The brightness and contrast can be defined in terms of mean and the standard deviation respectively. It is given as

$$Mean(\mu) = \frac{1}{N} \sum_{i=0}^{N-1} x_i \tag{7}$$

$$\text{Standard Deviation}(\sigma^2) = \frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^2 \quad (8)$$

E. Phase V

More specifically, the above mentioned two features are extracted and computed. TABLE 1 lists the sample statistical features of text and non-text. From the table, one should be able to make a conclusion i.e. high contrast and high brightness gives more textural property than non-text. However, it may be preferable to consider and conclude that high mean and high standard deviation gives a higher probability that the region contains text. After the conclusion, the portion of every text region is well separated from the document image under different backgrounds.

TABLE 1
SAMPLES OF STATISTICAL FEATURES

Images	Mean	Standard Deviation	Text Region	Non-Text Region
Document[1] 1 2 3 4 5	6.61	3.39	49 39 9 427 41	1 1 6 5 3
Document[2] 1 2 3 4 5	6.92	6.14	10 12 117 27 496	3 2 1 6 3
Document[3] 1 2 3 4 5	6.68	3.03	34 193 23 11 57	3 3 1 6 1

The Results of the five phases are shown in the Fig. 2(a-f) and it clearly shows that the proposed method has good abilities to detect all kinds of text regions and gives the good detection rate.

IV. RESULT ANALYSIS AND COMPARISON

This section presents some results of the proposed approach on images of different kinds of documents. The dataset was created as there is no standard dataset available in the literature to include 100 different scanned document images taken from newspapers, journals, books and browsing internet. Figure 2 shows the output of the proposed algorithm in various phases. The true colored version of the original input image is presented in (a). The output of the magnitude image detected by the algorithm is presented in (b) edge map image shown in (c) is calculated by the threshold value , and in the image (d) the non-text regions are separately shown by the morphological operators, finally the text regions extracted from the background based on statistical measures shown in (e and f).

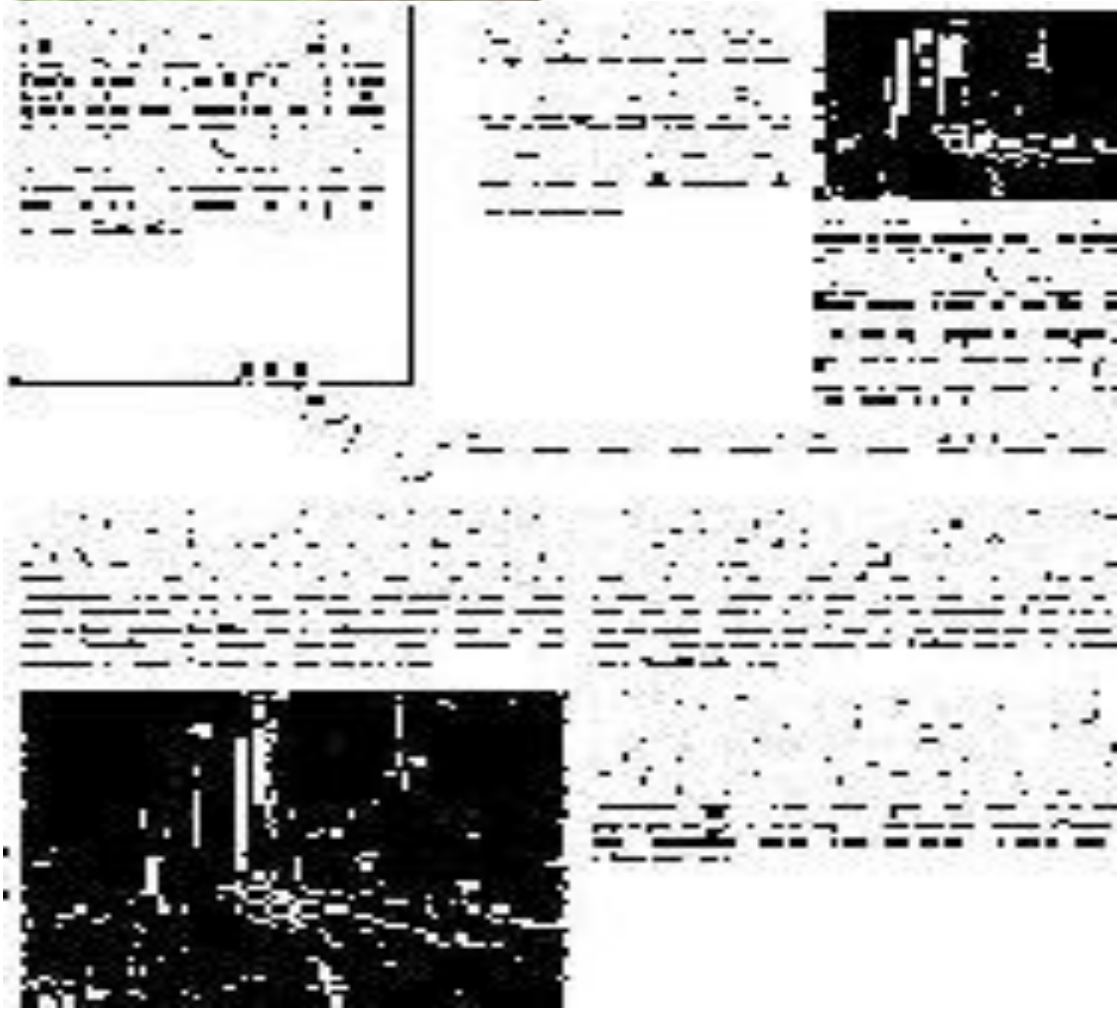


Fig. 2 a)True image b)Magnitude Image c)Edge Map Image

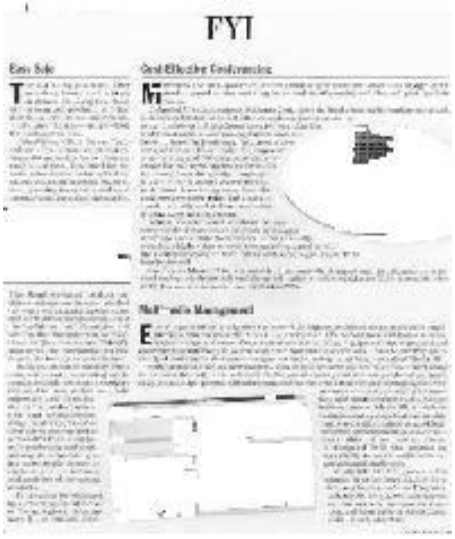
d)Non-Text separation Morphological Operation

e)Finding the text region using Statistical Measures

f)Separation of Text-region using

In another set of experiments, the performance of the text extraction results were compared with two existing algorithms [6] and [7] for the same data set. The first algorithm used the procedure of finding inner, outer and inner-outer corners of the text and non-text regions within an image. The second method extracted the text region only in a horizontal direction and not in all directions. From the Fig. 3, although the text lines were correctly detected by the two algorithms [6] and [7], many false alarms were also detected. This increases the complexity of algorithm to identify the edges at different orientations. The new combined approach has solved the above problems. This algorithm is sensitive to skew and text orientation and the output of the text extraction algorithm is fed to an OCR system to recognize the text region information.

Fig. 3 a)True Image



b) Output of [6]

c) Output of [7]

d) Output of Proposed Technique

To evaluate the performance of this method, detection and error rates were used. *Detection Rate* is the ratio of the of text regions correctly detected by the algorithm to the total number of text regions and the *Error Rate* are those regions in the image which are actually not a text region, but have been detected by the algorithm as text region. Comparisons and performance are shown in the TABLE 2.

TABLE 2
COMPARISON AND PERFORMANCE OF PROPOSED AND OTHER TECHNIQUES

Methods	Detection Rate	Error Rate
Peeta et.al [6]	94 %	5%
Parodi and Piccioli [7]	95%	3%
Proposed Method	96%	2%

V. CONCLUSION

In this paper, an effective combined edge-based approach using compass operators, edge map, morphological operations for feature extraction has been proposed for text region extraction in document images. Experimental results (Fig. 2 and Fig. 3) and comparisons (Table 2) showed that the proposed technique outperforms the existing algorithms in different orientations and different backgrounds and gives a good detection rate. In future, text region extraction can be extended to apply any one of the applications like document processing, content-based retrieval systems, image indexing and archiving documents.

REFERENCES

[1] K. C. Fan, L. S. Wang and Y. K. Wang, *page segmentation and identification for intelligent signal processing*, Signal Processing, 45:329-346, 1995.

[2] L. O. Gorman and R. Kasturi, *Document Image Analysis :A Primer*, Los Alamitos, California, USA: IEEE Computer Society Press, 1995.

[3] D. Wang and S. N. Srihari, *Classification of newspaper image blocks using texture analysis*, Computer Vision Graphics, and Image Processing, vol. 47, pp. 327–352, 1989.

[4] T. Pavlidis and J. Zhou, *Page segmentation and classification*, ComputerVision Graphics, and Image Processing, vol. 56, no. 6, pp. 484–496, 1992.

[5] Q. Yuan and C. L. Tan, *Text extraction from gray scale document images using edge information*, Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on Source: DBLP

[6] Farshad Nourbakhsh_, Peeta Basa Pati, and A.G. Ramakrishnan, *Text Localization and Extraction from Complex Gray Images*, ICVGIP 2006, LNCS 4338, pp. 776–785, 2006. c_Springer-Verlag Berlin Heidelberg 2006

[7] Pietro Parodi and Giulia Piccioli, *A fast and flexiblle statistical method for text extraction in document pagies*, p 619-623, 1063-6919/96 1996 IEEE

[8] P. A. Vijaya and M.C. Padma, *Text Line Identification from a Multilingual Document*, 978-0-7695-3565-4/09 © 2009 IEEE DOI 10.1109/ICDIP.2009.51, PP 302-305

[9] Sachin Grover, Kushal Arora andSuman K. Mitra,*Text Extraction from Document Images using Edge Information*, 978-1-4244-4859-3/09 ©2009 IEEE.

