International Journal For Technological Research In Engineering
Volume 5, Issue 4, December-2017 *ISSN (Online): 2347 - 4718*

TEXT EXCLUSION FROM DOCUMENT IMAGES AND NUMBER PLATE USING MATLAB

Er. Mubarak Ahmad Wani[1], Er. Priyanka Mehta[2]
[1]M.Tech Scholar, [2]HOD CSE, UIET lalru, Department Of Computer Science And Engineering

Abstract: ***Text regions extraction from document images containing both texts and graphics is an important step of any optical character recognition system. This paper describes an improvement over existing methods for localization of text regions from document images. The improvement is achieved by accommodating distinctive features like regularity in frequency, orientation, width, area, spatial cohesion etc. to identify text blocks in a document image containing both text and graphics. Proposed technique is tested on MARG dataset of multiple layouts and large varieties of color document images collected from web. Experimental result confirms the improvement of extraction accuracy by suppressing the false alarms notably.***

Keywords: ***Discrete wavelet transform (DWT), document image segmentation, Haar wavelet transform (HWT), text localization.***

I. INTRODUCTION

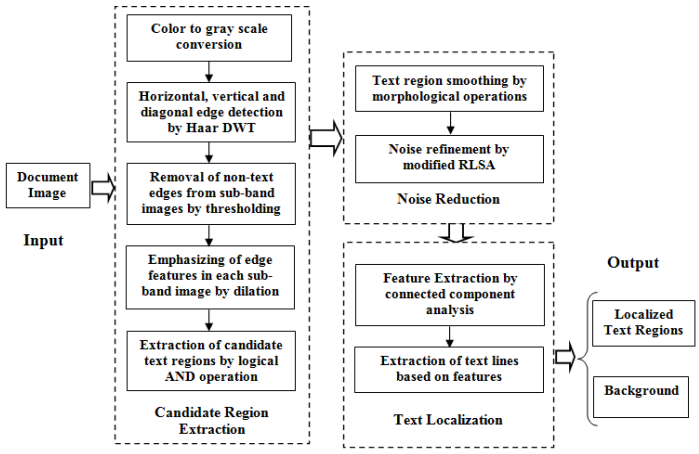
TEXT is usually the main source of information in documents and accurate text detection can greatly facilitate optical character recognition. Automatic recognition, reading, and storing information are the demands of modern technology. Therefore, text localization and extraction is a key area of research in document image analysis. However, locating and extracting textual data is not an easy task. Since texts are often embedded in different font styles, sizes, orientations and colors against a complex background. Moreover, low contrast between the text and the Complicated background often makes text detection extremely challenging. To address these problems, a large number of new methods for text localization, extraction and optical character recognition have been proposed recently. Some of the wellknown approaches are: (i) morphology based segmentation, ii) pixels based classification,(iii)connected component based classification , iv) edge based segmentation], v) texture based segmentation ,(vi) frequency based classification The survey papers enlist more techniques for layout analysis of document images. In text extraction process, the most important step is to find approximate locations of text lines in a gray-scale image. In this paper we address the problem of locating the textual data in an image. Our proposed system employs both connected component and discrete wavelets to localize text from complex document layouts. The paper is organized as follows. Section 2 deals with the related work. Section 3 gives a step by step description of proposed method. Experimental results are illustrated in Section 4. Finally, conclusions and future works are summarized in Section 5.

II. PREVIOUS WORK

Many researchers have been investigating various wavelet based techniques to retrieve textual information present in document and scene images. Li and Gray [8] used distribution characteristics of wavelet coefficients for document image segmentation. Liang and Chen [9] employed. Haar wavelet transforms to detect edges of candidate text regions. Kumar et al. [10] proposed globally matched wavelet filters and Markov random field (MRF) based processing for text extraction from document and scene text images. In 2004, Liang and Chen's proposed a Simple approach and it performs well for separating captions and titles from video and scene images. However, it is often unable to differentiate between text and non-text components in document images and hence produces large false alarms especially when the layout is complex. In this paper, we introduce an improvement of Liang and Chen's segmentation algorithm to suppress false alarm and generalize it for separating text and non-text components from document images as well.

III. PROPOSED METHOD

We propose an improved and efficient method to extract text regions from document images containing both text and graphics. The whole text extraction process is divided into three
Distinct parts:
a. Candidate region extraction
b. Noise reduction
c. Text localization



IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate the performance of proposed approach, we have selected various downloaded images of books, journals, and magazines from the Internet containing complex

backgrounds, graphics, different font sizes, and overlapping styles as the experimental data set. To demonstrate The efficiency of our method we have also tested proposed method in another database which contains 45document images from MARG [17] dataset. The latterdataset is created by randomly picking 5 images from each of nine classes of the page layouts of MARG. Our proposed method performed equally well for regular and irregular layouts along with complex background.

V. CONCLUSIONS

Text extraction from document images is a challenging task because of the complex background and multiresolution criteria. Moreover, degradations introduce during scanning or copying a paper document. This paper presents an efficient and simple method to locate texts in documents. To improve the accuracy we modified Liang and Chen's approach by accumulating RLSA with connected component analysis. Our experimental results show that, along with improving accuracy, our method reduces false alarms from resultant images. Moreover compared with other methods our technique relied on adaptability of predefined text region features.

REFERENCES

[1] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben, "A Robust algorithm for Text detection in images," Proc. of the 3rd International Symposium on Image and Signal Processing and Analysis, 2003.

[2] Bukhari, S. S., Shafait, F., and Breuel, T. M., "Document image segmentation using discriminative learning over connected components," Proc. 9th IAPR Workshop on Document Analysis Systems, pp. 183-190, 2010.

[3] M. Ly u, J. Song, M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 2, pp. 243 –255, 2005.

[4] A.K. Jain, Y. Zhong, "Page segmentation using texture analysis, "Pattern Recognition," vol. 29, no. 5, pp.743–770, 1996.

[5] J. Li and R. M. Gray, "Context based multi-scale classification of document images using wavelet coefficient distribution, " IEEE Trans. Image Process., vol. 9, no. 9, pp. 1604 –1616, Sep. 2000.

[6] Chung-Wei Liang and Po-Yueh Chen, "DWT based text localization," International Journal

[7] S.Rajakumar, Dr.V.Subbiah Bharathi, "Century Identification and Recognition of Ancient Tamil Character Recognition"- International Journal of Computer Applications,Volume 26– No.4, July 2011.

[8] Dayashankar Singh, Sanjay Kr. Singh, Dr. (Mrs.) Maitreyee Dutta, "Hand Written Character Recognition Using Twelve Directional Feature Input and Neural Network"- International Journal of Computer Applications Volume 1 – No. 3, 2010

[9] Anita Pal,Dayashankar Singh, "Handwritten English Character Recognition Using Neural Network"- International Journal of Computer Science & CommunicationVol. 1, No. 2, July-December 2010.

[10] C.Sureshkumar, Dr.T.Ravichandran, "Handwritten Tamil Character Recognition and Conversion using Neural Network"- International Journal on Computer Science and EngineeringVol. 02, No. 07, 2010.

[11] Dharamveer Sharma, Deepika Gupta, "Isolated Handwritten Digit Recognition using Adaptive Unsupervised Incremental Learning Technique"- International Journal of Computer Applications Volume 7– No.4, September 2010

