International Journal of Engineering, Science and Mathematics
Vol.6Issue 4, August 2017,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com Double-Blind
Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Text regionseparation in document images: A Review

Baljinder Kaur*
Balwinder Singh**

Abstract

Extracting text blocks from scanned document images is a crucial step in optical character recognition systems, as images, graphics,equations and tables are also part of documents. Other applications of text block detection are image compression, indexing and archiving. In this paper, an attempt has been made to list and analyze various text-graphics separation schemes for document images. The different techniques for various scripts has been studied and merits and demerits of various methods has been listed.

Keywords:

Text graphics separation;
Document images;
Text region;
Text identification;
Printed text.

Author correspondence:

Baljinder Kaur,
M.Tech. Research Scholar, Computer Engg. Section
Yadavindra College of Engineering, Talwandi Sabo, India

1. Introduction

The aim of geometrical layout analysis is to identify the text and non-text parts such as drawing and figures in a document image [1]. There are various non-text objects present in documents such as images, drawings, mathematical equations, tables etc., which must be filtered before applying OCR onto text area of the documents. Text separation is a process to segment the text block from non-text area in document images. Figure 1 shows the example different kind of scanned document images containing both text and non-text parts.

In this paper, various methods available in the literature to separate text blocks in various scripts for different document images has been reviewed. Though many studies has been carried out on image analysis, but only few research papers are available in context of Indian scripts. Different kind of images such as historical, newspaper, applications, scanned images, novels and books containing different scripts such as Devanagari, Roman, Gurumukhi and related scripts are considered in this study.

The rest of the paper is organized as follows. Section 2 discusses various text segmentation techniques for documents images in detail from existing literature. Section 3 presents the conclusion and future scope.

*M.Tech. Research Scholar, Computer Engg. Section, Yadavindra College of Engineering, Talwandi Sabo, India
**Assistant Prof. (Computer Science), Yadavindra College of Engineering, Talwandi Sabo, India

Figure 1. Document images containing text and graphics

2. Literature Survey

Various pre-sprocesing techniques has been also applied prior to text classification such as enhancemnt or noise filtering. Post-processing operations such as flitering touching text, has been applied to improve the results of classification. Segmentation phase for document images starts with text block segmentation, which plays a main role in correct word segmentation and recognition process.Based on exisiting methods [2], the text classification technique can be grouped into connected component analysis, edge, projection profile, filter and block based approaches. The various text region segmentation techniques are described in following section.

Yadav and Ragot [2] proposed a simple technique for text extraction based on identifying corner points. Fast corners points have been identified after dividing the image into equal size blocks. The density of points in each block has been computed and based on computed threshold, blocks with more dense points has been designated as text blocks. The connectivity of neighboring blocks has been checked to group them into full text blocks. The method has been tested on different scripts, handwritten and printed, skewed, noisy, documents with various resolutions with precision and recall higher than 90%.

Fletcher and Kasturi [3] proposed a highly efficient algorithm for text separation firstly by generating connected components, then filtering area/ratio of it. Then colinear components has been grouped with Hough transform and finally text has been grouped into words. The method separates both text and image into individual images. The results demonstrate the superiority of technique as compared to other methods.

Khedekar et.al. [4] describes a projection-profile based technique to separate text blocks from document images using header-line feature of Devanagari script. Features such as frequency, orientation etc. has been used to classify the blocks into regular (text) and irregular blocks. The method is based on fact that text area has irregular level script in text areas, while having a uniform pattern in case of graphics portion.

Arvind et.al. [5] proposed a method to classify text from other non-text region based on features horizontal projection profile (HPP), Eigen profile (EP) vector and Fisher profile vector. Nearest neighbor (NN), Support vector machine and neural network has been used as classifier for text separation. Authors reported better results for HHP features in transform domain and EP-NN feature-classifier combination achieved accuracy of 96.5%.

Jain &Bhattacharjee [6] proposed a Gabor filter based approach to segment text and non-text area in document images using texture features. The idea is that text and remaining parts of a documents basically have dissimilar texture. The results presented demonstrate that method works successfully for different styles of font, skewed and handwritten documents.

Garg et.al. [7] presented a method for separating text and graphics in Indian language newspapers by using texture properties of the script. Expectation maximization approach has been used to optimize the various parameters and adaptive segmentation is applied. Total 600

images from six Indian scripts has been used for testing and an accuracy in range of 87% has been reported.

Lu [8] proposed a method for detection of text regions by removing non-text parts from engineering drawings. The method works by first erasing horizontal and vertical linear parts such as straight lines. Then left-out graphics parts has been removed by utilizing strokes density of identified connected components. The output generated from previous operation has been subjected to morphological operations erosion and dilation to remove the overlapping characters and graphics. Finally, text strings has been extracted from image after circumscribing the rectangle around text parts. The method has been able to successfully locate Chinese and western characters, but is immune to noise in images.

Kumar et.al. [9] presented a method to identify text area in an image using a clustering technique to estimate globally matching wavelets and Fisher classifier. The results have been optimized by using Markov random field pixel labeling. The method has been tested with encouraging results on document and general images with complex background.

Grover et. al. [10] proposed a methodology for text extraction from colour document images using text edges. Edge image has been obtained using Sobel edge operator after conversion of colour into from grayscale image. Edge image has been divided into non-overlapping small sized rectangle blocks and each block has been classified as text or non-text block based on computation of edge-features for each block. The methods has been found to work well for newspaper and magazine documents images scanned at various resolutions.

4. Conclusion

Text extraction has many applications such as page segmentation, layout analysis, image compression and document archiving and image indexing. The paper is very helpful for researchers carrying out research in field of optical character recognition. The paper discussed various text area segmentation techniques for document images. The various techniques has been studied and merits and demerits of different methods has been listed.

Although many text separation techniques exist for various Indian scripts, but only few papers are available for Gurmukhi script. In future, a novel technique or improved technique will be proposed for text area separation from Gurmukhi script.

References

[1] Cattoni R., T. Coianiz, S. Messelodi, and C. M. Modena. "Geometric layout analysis techniques for document image understanding: a review." IRST, Trento, Italy (1998).

[2] Yadav Vikas, Nicolas Ragot “Text Extraction in Document Images: Highlight on Using Corner Points” In Document Analysis Systems (DAS),12th IAPR Workshop on, pp. 281-286 Institute of Electrical and Electronics Engineers Standard (IEEE), 2016.

[3] Fletcher Lloyd Alan and Rangachar Kasturi “A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images.” Institute of Electrical and Electronics Engineers(IEEE) transactions on pattern analysis and machine intelligence,vol. 10, no. 6, pp. 910-918, Nov. 1988.

[4] Khedekar Swapnil, Vemulapati Ramanaprasad Srirangaraj Setlur and Venugopal Govindaraju “Text-Image Separation in Devanagari Documents.” 7th international conference on document analysis and recognition (ICDAR) Institute of Electrical and Electronics Engineers Standard, vol. 3, pp. 1265, 2003.

[5] Arvind K. R., Peeta Basa Pati and A. G. Ramakrishnan "Automatic Text Block Separation in Document Images." In Intelligent Sensing and Information Processing, ICISIP 4th International Conference on pp. 53-58, Institute of Electrical and Electronics Engineers Standard (IEEE), 2006.

[6] Jaiun Anil K., and Sushil Bhattacharjee. "Text segmentation using Gabor filters for automatic document processing." Machine Vision and Applications, vol. 5, no. 3, pp.169-184, 1992.

[7] Garg Ritu, Anukriti Bansal, Santanu Chaudhury, and Sumantra Dutta Roy. "Text graphic separation in Indian newspapers." In Proceedings of the 4th International Workshop on Multilingual OCR, p. 13. ACM, 2013.

[8] Lu Zhaoyang. "Detection of text regions from digital engineering drawings." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 4, pp. 431-439, 1998.

[9] Kumar Sunil, Rajat Gupta, Nitin Khanna, Santanu Chaudhury, and Shiv Dutt Joshi. "Text extraction and document image segmentation using matched wavelets and MRF model." IEEE Transactions on Image Processing, vol. 16, no. 8, pp. 2117-2128, 2007.

[10] Sachin Grover, Kushal Arora, and Suman K. Mitra. "Text extraction from document images using edge information." In proceedings of India Conference (INDICON), 2009 Annual IEEE, pp. 1-4. IEEE, 18-20 Dec. 2009.