Home > Journal > Business & Economics | Computer Science & Communications > IIM

Indexing View Papers Aims & Scope Editorial Board Guideline Article Processing Charges

IIM> Vol.2 No.2, February 2010

OPEN ACCESS

Text Extraction in Complex Color Document Images for Enhanced Readability

PDF (Size:2745KB) PP. 120-133 DOI: 10.4236/iim.2010.22015

Author(s)

P. Nagabhushan, S. Nirmala

ABSTRACT

Often we encounter documents with text printed on complex color background. Readability of textual contents in such documents is very poor due to complexity of the background and mix up of color(s) of foreground text with colors of background. Automatic segmentation of foreground text in such document images is very much essential for smooth reading of the document contents either by human or by machine. In this paper we propose a novel approach to extract the foreground text in color document images having complex background. The proposed approach is a hybrid approach which combines connected component and texture feature analysis of potential text regions. The proposed approach utilizes Canny edge detector to detect all possible text edge pixels. Connected component analysis is performed on these edge pixels to identify candidate text regions. Because of background complexity it is also possible that a non-text region may be identified as a text region. This problem is overcome by analyzing the texture features of potential text region corresponding to each connected component. An unsupervised local thresholding is devised to perform foreground segmentation in detected text regions. Finally the text regions which are noisy are identified and reprocessed to further enhance the quality of retrieved foreground. The proposed approach can handle document images with varying background of multiple colors and texture; and foreground text in any color, font, size and orientation. Experimental results show that the proposed algorithm detects on an average 97.12% of text regions in the source document. Readability of the extracted foreground text is illustrated through Optical character recognition (OCR) in case the text is in English. The proposed approach is compared with some existing methods of foreground separation in document images. Experimental results show that our approach performs better.

KEYWORDS

Color Document Image, Complex Background, Connected Component Analysis, Segmentation of Text, Texture Analysis, Unsupervised Thresholding, OCR

Cite this paper

P. Nagabhushan and S. Nirmala, "Text Extraction in Complex Color Document Images for Enhanced Readability," *Intelligent Information Management*, Vol. 2 No. 2, 2010, pp. 120-133. doi: 10.4236/iim.2010.22015.

References

[1] A. K. Jain and S. K. Bhattacharjee, " Address block location on envelopes using Gabor filters, " Pattern Recognition, Vol. 25, No 12, pp. 1459 – 1477, 1992.

[2] V. Wu, R. Manmatha, and E. M. Riseman, " Textfinder: An automatic system to detect and recognize text in images, " IEEE PAMI, Vol. 21, No. 11, pp. 1224 – 1229, 1999.

[3] D. Chen, H. Bourland, and J. P. Thiran, " Text identification in complex background using SVM, " Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 621 – 626, 2001.

[4] U. Garain, T. Paquet, and L. Heutte, " On foreground- background separation in low quality document images, " International Journal of Document Analysis and Recognition, Vol. 8, No. 1, pp. 47 – 63, 2006.

• Open Special Issues
• Published Special Issues
• Special Issues Guideline

IIM Subscription

Most popular papers in IIM

About IIM News

Frequently Asked Questions

Recommend to Peers

Recommend to Library

Contact Us

Downloads:	144,622
Visits:	361,759

[5] H. Hase, M. Yoneda, S. Tokai, J. Kato, and C. Y. Suen, “Color segmentation for text extraction,” IJDAR, Vol. 6, No. 4, pp. 271–284, 2003.

[6] T. Kasar, J. Kumar, and A. G. Ramakrishnan, “Font and background color independent text binarization,” Proceedings of 2nd International Workshop on Camera Based Document Analysis and Recognition, pp. 3–9, 2007.

[7] E. Kavallieratou and E. Stamatatos, “Improving the quality of degraded document images,” Proceedings of 2nd International Conference on Document Image Analysis for Libraries, pp. 340–349, 2006.

[8] G. Leedham, Y. Chen, K. Takru, J. H. N. Tan, and L. Mian, “Comparison of some thresholding algorithms for text/background segmentation in difficult document images,” Proceedings of 7th International Conference on Document Analysis and Recognition, pp. 859–864, 2003.

[9] W. Niblack, “An introduction to image processing,” Prentice Hall, Englewood Cliffs, 1986.

[10] S. Nirmala, P. Nagabhushan, “Isolation of foreground- text in document images having known complex background,” Proceedings of 2nd International Conference on Cognition and Recognition, pp. 99–106, 2008.

[11] N. Otsu, “A threshold selection method from gray level histograms,” IEEE Transactions on Systems, Man & Cybernetics, Vol. 9, No. 1, pp. 62–66, 1979.

[12] M. Pietik?inen and O. Okun, “Text extraction from grey scale page images by simple edge detectors,” Proceedings of the 12th Scandinavian Conference on Image Analysis (SCIA), pp. 628–635, 2001.

[13] J. Sauvola and M. Pietik?inen, “Adaptive document image binarization,” Pattern Recognition, Vol. 33, No. 2, pp. 225–236, 2000.

[14] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” Journal of Electronic Imaging, Vol. 13, No. 1, pp. 146– 165, 2004.

[15] K. Sobottka, H. Kronenberg, T. Perroud, and H. Bunke, “Text extraction from colored book and journal covers,” IJDAR, Vol. 2, No. 4, pp. 163–176, 1999.

[16] C. L. Tan and Q. Yaun, “Text extraction from gray scale document image using edge information,” Sixth International Conference on Document Analysis and Recognition, pp. 302–306, 2001.

[17] O. D. Trier and A. K. Jain, “Goal directed evaluation of binarization methods,” IEEE PAMI, Vol. 17,