{tag}

{/tag}

International Journal of Computer Applications

Volume 33 - Number 1

Year of Publication: 2011

Authors:

Brij Mohan Singh

Ankush Mittal

Vivek Chand

Debashish Ghosh

10.5120/3988-5640

{bibtex}pxc3875640.bib{/bibtex}

Abstract

There are numerous stylish documents which do not have the traditional text layouts where printed text regions are not parallel to each other. Such complex layouts make text line extraction challenging due to multi-orientation of paragraphs. This paper introduces a system for the text line extraction from the complex layout documents. Proposed method is based on the

concept of dilation and histogram profiling. The text regions are extracted using dilation and food fill based approach, then paragraph orientation is determined and individual text lines are extracted. The accuracy of extracted text lines are evaluated using the new proposed concept that is also based on the histogram profiling. The results of proposed approach on the complex layouts are promising.

Reference

- Marinai, S. 2008 Introduction to document analysis and recognition. Studies in Computational Intelligence (SCI), (2008), 90, 1–20.
- Tang, Y.Y., Suen, C.Y. Yan, C.D. and Cheriet, M. 1991. Document analysis and understanding: a brief survey. In Proceeding of the 1st International Conference on Document Analysis and Recognition, Saint-Malo, France, 17-31.
- Plamondon, R., and Srihari, S. N. 2000. On-line and off-line handwritten recognition: A comprehensive survey. IEEE Trans. on PAMI, (2000), Vol.22, 62-84
- Sethi, I. K. and Chatterjee, B. 1977. Machine recognition of constrained hand printed Devnagari. Pattern Recognition, (1977), Vol. 9, 69-75.
- Shaw, B., Parui, S. K., and Shridhar, M. 2008. A segmentation based approach to offline handwritten Devanagari word recognition. PReMI, IEEE, (2008), 528-35.
- Singh, B.M., Mittal, A., and Ghosh, D. 2011. Parallel implementation of Devanagari text line and word segmentation approach on GPU. International Journal of Computer Applications (2011), 24(9):7–14.
- Antonacopoulos, A., Pletschacher, S., Bridson D., and Papadopoulos, C. 2009. ICDAR 2009 page segmentation competition. In Proceeding of International Conference of Document Analysis and Recognition, IEEE, 1370-1374.
- Shafait1, F., Beusekom, J. V., Keysers, D., and Breuel, T. M. 2008. Background variability modeling for statistical layout analysis, In Proceeding of 19th International conference on Pattern Recognition, IEEE, 1-4.
- Fujisawa, H., Nakano, Y., and Kurino, K. 1992. Segmentation methods for character recognition from segmentation to document structure analysis. In Proceeding of the IEEE, Vol.80, pp. 1079-1092. 1992.
- Likforman-Sulem, L., and Faure, C. 1994. Extracting text lines in handwritten documents by perceptual grouping. Advances in handwriting and drawing: a multidisciplinary approach, (1994), 21-38.
- Abuhaiba, I.S.I., Datta, S., and Holt, M.J.J. 1995. 'Line extraction and stroke ordering of text pages. In Proceedings of the Third International Conference on Document Analysis and Recognition, Canada, 390- 393.
- Zahour, A., Taconet, B., Mercy, P. and Ramdane, S. 2001. Arabic hand-written text-line extraction. In Proceedings of the Sixth International Conference on Document Analysis and Recognition, 281–285.
- Weliwitage, C., Harvey A. L., and Jennings, A. B. 2005. Handwritten document offline text line segmentation. In Proceedings of Digital Imaging Computing: Techniques and Applications, 184-187.
- Louloudis, G., Gatos, B., Pratikakis, I. and Halatsis, K. 2006. A block based hough transform mapping for text line detection in handwritten documents. In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 515-520.

- Li, Y., Zheng, Y., Doermann, D., and Jaeger, S. 2006. A new algorithm for detecting text line in handwritten documents. In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 35–40.
- Zahour, A., Taconet, B., Likforman-Sulem, L., and Boussellaa, W. 2008. Overlapping and multi-touching text-line segmentation by block covering analysis. Pattern Analysis and Applications, (2008), Vol. 12, 335-351.
- Goto, H., and Aso, H. 1999. Extracting curved lines using local linearity of the text line. International Journal of Document Analysis and Recognition, (1999), vol. 2, 111–118.
- Hones, F., and Litcher, J. 1994. Layout extraction of mixed mode documents. Machine Vision Applications, (1994), vol. 7, 237–246.
- Liao, S. X., and Pawlak, M. 1996. On image analysis by moments. IEEE Transaction on PAMI, (1996) Vol.18, 254-266.
- Pal, U., Sinha, S., and Chaudhuri, B. B. 2003. English multi-oriented text line extraction. Image Analysis, Springer Verlag, Lecture Notes on Computer Science (LNCS-2749), 1146-1153.
- Roy, P. P., Pal, U., Lladós, J., and Kimura, F. 2008. Multi-oriented English text line extraction using background and foreground information. In Proceeding of Eighth IAPR Workshop on Document Analysis Systems, IEEE, 315-322.
- Roy, P. P., Pal, U., Lladós, J., and Kimura, F. 2008. Convex hull based approach for multi-oriented character recognition from graphical documents. In Proceeding of International Conference on Pattern Recognition, IEEE, 1-4.
- Pal, U., and Tripathy, N. 2004. Multioriented and curved text lines extraction from Indian documents. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, (2004), Vol. 34, No. 4, pp. 1676-1684.
- Pal, U., and Tripathy, N. 2005. Recognition of Indian multi-oriented and curved text. In Proceedings of the Eight International Conference on Document Analysis and Recognition, 141-145.
- Pal, U, and P. P. Roy, P.P. 2004. Multi-oriented and curved text lines extraction from Indian documents. IEEE Transaction on SMC - Part B, (2004), vol.34, 1676-1684.
- Kise, K., Yanagidw, O., and Takamatsu, S. 1996. Page segmentation based on thinning of background. In Proceedings of International Conference on Pattern Recognition, 788-792.
- Wong, K. Y., Casey, R. G., and Wahl, F. M. 1982. Document analysis system. IBM Journal of Research and Development, (1982), vol. 26, no. 6, 647–656.
- Nagy, G., Seth, S. and Viswanathan, M. 1992. A prototype document image analysis system for technical journals. Computer, (1992), vol. 7, no. 25, 10–22.
- O'Gorman, L. 1993. The document spectrum for page layout analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence, (1993) vol. 15, no. 11, 1162–1173.
- Baird, H. S., 1994. Background structure in document images. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), (1994), 8(5):1013–1030.
- Esposito, F., Malerba, D., and Semeraro, G. 1995. A knowledge based approach to the layout analysis. In Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, 466-471.
- Liu, J., Tang, Y. Y., He, Q., and Suen, C. Y. 1996. Adaptive document segmentation and geometric relation labeling: algorithms and experimental results. In Proceedings of the 13th International Conference on Pattern Recognition, vol. 3, 163-767.
- Dori, D., Doermann, D., Shin, C., Haralick, R., Phillips, I., Buchman, M., and Ross, D.

1997. The representation of document structure: A generic object-process analysis. In Handbook of Character Recognition and Document Image Analysis, H. Bunke and P. Wang, Eds. World Scientific, Singapore, 421–456.

- Simon, A., Pret, J.-C., and Johnson. A. 1997. A fast algorithm for bottom-up document layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19:273–277.
- Kise, K., Sato, A., and Iwata, M. 1998. Segmentation of page images using the area Voronoi diagram. Computer Vision and Image Understanding, (1998), vol. 70, no. 3, 370–382.
- Cattoni, R., Coianiz, T., Messelodi, S., and Modena, C.M. 1998. Geometric layout analysis techniques for document image understanding: a review. Downloaded from http://citeseer.nj.nec.com/, IRST, Trento, Italy, Tech. Rep. 9703-09.
- Jain, A.K., and Yu, B. 1998. Document representation and its application to page decomposition. IEEE Transaction on PAMI, (1998), 20(3), 294–308.
- Ishitani, Y., 1999. Logical structure analysis of document images based on emergent computation. In Proceedings of International Conference on Document Analysis and Recognition, Bangalore, India, 189–192.
- Kim, J., Le, D. X., and Thoma, G. R. 2001. Automated labeling in document images. In Proceedings of SPIE Conference on Document Recognition and Retrieval VIII, , San Jose, CA, 111–122.
- Breuel, T. M. 2002. Two geometric algorithms for layout analysis. In Document Analysis Systems, Princeton, NY, 188–199.
- Breuel, T. M. 2003. High performance document layout analysis. In Symposium on Document Image Understanding Technology, Greenbelt, MD.
- Antonacopoulos, A., Gatos, A., and Karatzas, D. 2003. ICDAR 2003 page segmentation competition. In Proceeding of 7th International Conference on Document Analysis and Recognition, Edinburgh, UK, 688–692.
- Mao, S., Rosenfeld, A., and Kanungo, T. 2003. Document structure analysis algorithms: a literature survey. In Proceeding of SPIE Electronic Imaging, vol. 5010, 197–207.
- Antonacopoulos, A., Gatos, B., and Bridson, D. 2005. ICDAR 2005 page segmentation competition. In Proceeding of 8th International Conference on Document Analysis and Recognition, Seoul, Korea, 75–80.
- Antonacopoulos, A., Gatos, B. and Bridson, D. 2007. ICDAR2007 page segmentation competition. In Proceeding of ICDAR2007, Curitiba, Brazil, 1279-1283.
- Cao, H., Prasad, R., Natarajan, P., MacRostie, E. 2007. Robust page segmentation based on smearing and error correction unifying top-down and bottom-up approaches. Ninth International Conference on Document Analysis and Recognition, IEEE, 392-396.
- Marinai, S., Marino, E., and Soda, G. 2005. Layout based document image retrieval by means of XY tree reduction. In Proceeding of 8th International Conference on Document Analysis and Recognition, Seoul, Korea, 432–436.
- Shafait, F., Keysers, D., and Breuel, T. M. 2006. Performance comparison of six algorithms for page segmentation. In Proceeding of 7th IAPR Workshop on Document Analysis Systems, Nelson, New Zealand, 368–379.
- Keysers, D., Shafait, F., and Breuel, T. M. 2007. Document image zone classification - a simple high-performance approach. In Proceeding of 2nd International Conference on Computer Vision Theory and Applications, Barcelona, Spain, 44–51.
- Smith, R. 2009. Hybrid page layout analysis via tab-stop detection. In Proceeding of International Conference of Document Analysis of Recognition, Barcelona, Spain, 241-245.

- Gonzalez, R.C., and Woods, R.E. 2011. Digital image processing. 4th Ed. (DIP/4e), Pearson Education Asia.

Index Terms

Computer Science Pattern Recognition

Key words

OCR Segmentation Profiling Profiling