Comments

Comment: Projection Methods Require Black Border Removal

George Nagy, *Fellow*, *IEEE*,
Sharad Seth, *Fellow*, *IEEE*, and
Mahesh Viswanathan, *Sr. Member*, *IEEE*

Abstract—A persistent flaw in the evaluation of page segmentation algorithms is examined.

Index Terms—X-Y tree, page segmentation, layout analysis.

A detailed comparative evaluation of six page segmentation methods, reported recently by Shafait et al. [1], follows the experimental protocol of an earlier study by Mao and Kanungo [2] that purported to show that recursive X-Y cut (RXYC) segmentation is much more error-prone than competitive methods, even on isothetic layouts. It does not, however, require much experimentation or reflection to discover that all pixel-projection methods (for either segmentation or skew determination) require removing any black background introduced by optical scanning.

We reported good results for RXYC segmentation in 1992 [3], from which Shafait et al. extracted the algorithm, and added to the evidence in [4]. The documents used in our experiments were either scanned or copied against a *white* background or obtained from an IEEE CD-ROM (this was pre-Web!) with all-white backgrounds. The page images tested in [1] and [2] were drawn from the University of Washington data set [5], which was evidently scanned against a *black* (or nonreflective) background. As indicated in Fig. 6 of [1], black borders probably account for most of the RXYC segmentation errors.

A reasonable motivation for a nonreflective background is that detecting the edges of the paper greatly simplifies eliminating black pixels that do not belong to the page, as well as estimating and removing any skew introduced in scanning. Neither of these steps was performed in the preparation of the database (presumably to allow testing different methods).

Like Mao and Kanungo, Shafait et al. suggest that the poor performance of the X-Y tree method is due to its vulnerability to noise. (Mao and Kanungo also called the black borders "noise," suggested removing them, but did not.) Black borders are not *noise*. In image or signal processing, *noise* denotes random, unpredictable phenomena, not entirely predictable artifacts. Even low-end fax scanners avoid adding black borders.

Performing the simple but essential step of border removal would not have violated the spirit of Mao's and Kanungo's general approach and the valuable pixel-based ground truth and "vectorial" performance measure proposed by Shafait et al. It would have much improved these otherwise careful and thorough evaluations of page segmentation algorithms.

References

[1] F. Shafait, D. Keysers, and T.M. Breuel, "Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 6, pp. 941-954, June 2008.

[2] S. Mao and T. Kanungo, "Empirical Performance Evaluation and Its Application to Page Segmentation Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 242-256, Mar. 2001.

[3] G. Nagy, S. Seth, and M. Viswanathan, "A Prototype Document Image Analysis System for Technical Journals," *Computer,* vol. 25, no. 7, pp. 10-22, July 1992.

[4] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan, "Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no, 7, pp. 737-747, July 1993.

[5] I. Guyon, R.M. Haralick, J.J. Hull, and I.T. Phillips, "Data Sets for OCR and Document Image Understanding Research," *Handbook of Character Recognition and Ddocument Image Analysis,* H. Bunke and P. Wang, eds., pp. 779-799, World Scientific, 1997.

- *G. Nagy is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, JEC 6020, 110 8th Street, Troy, NY 12180-0115. E-mail: nagy@ecse.rpi.edu.*
- *S. Seth is with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Avery Hall, Room 359, Lincoln, NE 68588-0115. E-mail: seth@cse.unl.edu.*
- *M. Viswanathan is with the IBM T.J. Watson Research Center, 294 Route 100, Mail Stop: SOM4/4L04, Somers, NY 10589. E-mail: maheshv@us.ibm.com.*

Manuscript received 30 May 2008; accepted 14 July 2008; published online 22 July 2008.
Recommended for acceptance by L. O'Gorman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-05-0314.
Digital Object Identifier no. 10.1109/TPAMI.2008.192.

0162-8828/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society