

COMMENT

Projection methods require black border removal

G. Nagy¹, S. Seth², M. Viswanathan³

¹ Rensselaer Polytechnic Institute, Troy, NY

² University of Nebraska – Lincoln, Lincoln, NE

³ IBM T.J. Watson Research Center, Yorktown Heights, NY

Abstract

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

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 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 U. Washington dataset [5], which was evidently scanned against a *black* (or non-reflective) background. As indicated in Fig. 6 of [1], black borders probably account for most of the RXYC segmentation errors.

A reasonable motivation for a non-reflective 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 data base (presumably to allow testing different methods).

Like Mao and Kanungo, Shafait, Keysers and Breuel 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 not 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, 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, March 2001.
- [3] G. Nagy, S. Seth, 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. on Pattern Analysis and Machine Intelligence*, vol. 15, no, 7, pp. 737-747, 1993.
- [5] I. Guyon, R. M. Haralick, J. J. Hull, and I. T. Phillips, "Data sets for OCR and document image understanding research," in *Handbook of character recognition and document image analysis*, H. Bunke and P. Wang, Eds. World Scientific, Singapore, 1997, pp. 779–799.