Response to 'Projection methods require black border removal'

Faisal Shafait, Daniel Keysers, and Thomas M. Breuel

Abstract

In contrast to prior experimental work, our results support the conclusion that RXYC can perform well after marginal noise removal. However, marginal noise removal on page images like those found in UW3 remains a hard problem, and it therefore remains an open question whether RXYC can actually achieve competitive performance on such databases.

Index Terms

Document page segmentation, OCR, performance evaluation, performance metric

Nagy et al. [1] misrepresent our work [2], [3] when they write:

"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."

We do not conclude that RXYC has "poor performance". In fact, our paper strongly argues against such a simplistic, one-dimensional view of performance evaluation. The stated purpose of our paper was to introduce a novel evaluation method based on a vectorial score, and demonstrate its utility and validity by comparing it to the results obtained using Mao and Kanungo's method [4]. Our analysis shows, among other things, that Mao and Kanungo's conclusion that RXYC is the worst of the algorithms needs to be modified, and our very first recommendation in [3] is (Section IV.C):

"For clean documents with little or no skew, the x-y cut algorithm might be a good choice as it is fast and easy to implement."

F. Shafait and D. Keysers are with the Image Understanding and Pattern Recognition (IUPR) research group at the German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany. E-mail: {faisal.shafait, daniel.keysers}@dfki.de

T.M. Breuel is with the Computer Science Department at the Technical University of Kaiserslautern, Germany. E-mail: tmb@informatik.uni-kl.de

In different words, if black borders and marginal noise have been successfully removed, and if documents have been successfully deskewed, we tentatively recommend the use of RXYC. We can derive such a recommendation from our data precisely because the vectorial score lets us draw conclusions about the behavior of algorithms without testing all possible combinations of preprocessing methods and layout analysis methods. This is one illustration of the advantages of the vectorial score over a simple score and is the primary point of our paper.

Nagy et al. imply in their letter that the necessary document cleanup step is simple; for example, they write:

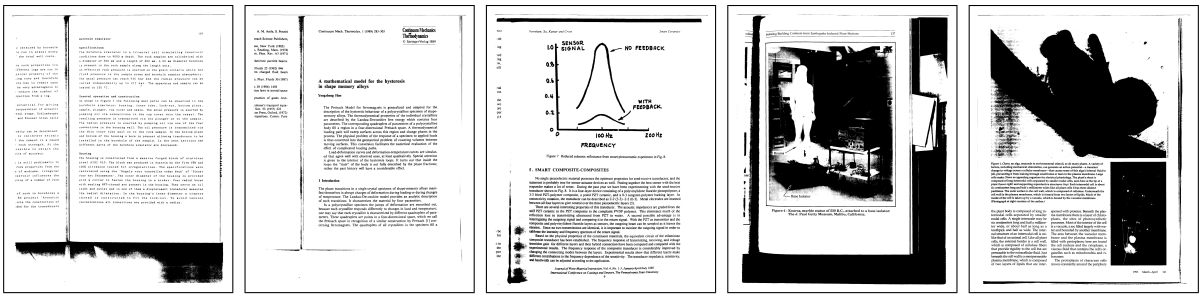
"The page images tested in [1] and [2] were drawn from the U. Washington dataset [5], which was evidently scanned against a black (or non-reflective) background. [...] A reasonable motivation for a non-reflective background is that detecting the edges of the paper greatly simplifies eliminating black pixels that do not belong to the page"

In fact, Nagy et al. are wrong regarding the origin and nature of the marginal noise in UW3. The source of the marginal noise in UW3 is documented [5]: UW3 contains pages from a wide variety of scanning conditions, including different page sizes and scans after manual photocopying. Contrary to what Nagy et al. state, UW3 does not contain "black borders" designed to be easy to remove, it contains unpredictable and variable marginal noise. This is also evident from looking at samples of UW3 page images (see Figure 1).

In addition, a growing literature on marginal noise removal [6]–[12], suggests that marginal noise removal remains a difficult problem. We do not know of any algorithm (simple or otherwise) capable of reliably removing marginal noise components on UW3 page images to the degree required by RXYC. Therefore, although we considered it, testing combinations of RXYC and different marginal noise removal methods is not a "simple" experiment that we could have carried out as part of our evaluation, and as it was not directly relevant to the actual conclusions of our paper, we decided to leave this for future work.

We believe that the source of the "persistent flaw" that Nagy et al. perceive in subsequent work may lie in their own publications: neither [13], nor [14], disclose a limitation of RXYC to documents scanned against a white background, nor provide a border removal algorithm. If they had done so, subsequent work would have taken that into account.

Our paper makes the statement that our experimental results support: RXYC methods can perform well if marginal noise can be removed. This result represents a strong improvement



(a) A00IBIN (b) A001BIN (c) A031BIN (d) D035BIN (e) E001BIN

Fig. 1. Sample images from the University of Washington Database 3 (UW3). The samples illustrate the variability and unpredictability of marginal noise in UW3. Across the entire database, noise outside the page margins of the scanned page consists of connected components at many different sizes and shapes, including actual text, and ranges from nearly absent to dominating the image. On other pages, illustrations outside the text area may resemble marginal noise and black borders.

over Mao and Kanungo's results, which simply stated that RXYC works poorly on UW3. Nagy et al.'s statement that border removal is "simple", however, is evidently false for UW3 and similar real-world databases; we note that even in their letter, they fail to cite or state such an algorithm.

Determining whether RXYC can actually achieve competitive performance on document image collections like UW3 therefore remains a complex and open question that needs to be explored in future work.

REFERENCES

[1] G. Nagy, S. Seth, and M. Viswanathan, "Projection methods require black border removal," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22 July 2008, preprint.

[2] F. Shafait, D. Keysers, and T. M. Breuel, "Performance comparison of six algorithms for page segmentation," in *7th IAPR Workshop on Document Analysis Systems*, ser. Lecture Notes in Computer Science, vol. 3872, Nelson, New Zealand, Feb. 2006, pp. 368–379.

[3] ——, "Performance evaluation and benchmarking of six-page segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, 2008.

[4] S. Mao and T. Kanungo, "Empirical performance evaluation methodology and its application to page segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 242–256, 2001.

[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.

[6] K. C. Fan, Y. K. Wang, and T. R. Lay, "Marginal noise removal of document images," *Pattern Recognition*, vol. 35, no. 11, pp. 2593–2611, 2002.

[7] L. Cinque, S. Levialdi, L. Lombardi, and S. Tanimoto, "Segmentation of page images having artifacts of photocopying and scanning," *Pattern Recognition*, vol. 35, no. 5, pp. 1167–1177, 2002.

[8] B. T. Avila and R. D. Lins, "Efficient removal of noisy borders from monochromatic documents," in *Int. Conf. on Image Analysis and Recognition*, Porto, Portugal, Sep. 2004, pp. 249–256.

[9] W. Peerawit and A. Kawtrakul, "Marginal noise removal from document images using edge density," in *4th Information and Computer Engineering Postgraduate Workshop*, Phuket, Thailand, Jan. 2004.

[10] F. Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Page frame detection for marginal noise removal from scanned documents," in *SCIA 2007, Image Analysis, Proceedings*, ser. Lecture Notes in Computer Science, vol. 4522, Aalborg, Denmark, June 2007, pp. 651–660.

[11] N. Stamatopoulos, B. Gatos, and A. Kesidis, "Automatic borders detection of camera document images," in *2nd Int. Workshop on Camera-Based Document Analysis and Recognition*, Curitiba, Brazil, Sep. 2007, pp. 71–78.

[12] F. Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Document cleanup using page frame detection," *Int. Jour. on Document Analysis and Recognition*, 2008, accepted for publication.

[13] G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals," *Computer*, vol. 7, no. 25, pp. 10–22, 1992.

[14] 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.