2009

Projection Methods Require Black Border Removal

George Nagy
Rensselaer Polytechnic Institute, nagy@ecse.rpi.edu

Sharad C. Seth
University of Nebraska - Lincoln, seth@cse.unl.edu

Mahesh Viswanathan
IBM T.J. Watson Research Center, Somers, NY, maheshv@us.ibm.com

Follow this and additional works at: http://digitalcommons.unl.edu/csearticles

Part of the Computer Sciences Commons

http://digitalcommons.unl.edu/csearticles/24

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in CSE Journal Articles by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
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


For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.