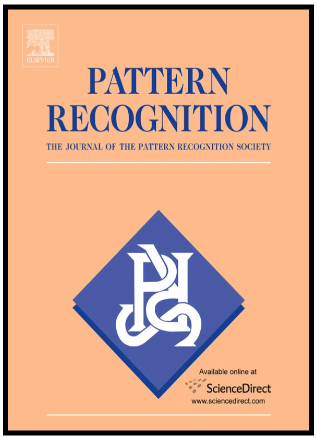

www.elsevier.com/locate/pr

A comprehensive survey of mostly textual document segmentation algorithms since 2008

Sébastien Eskenazi, Petra Gomez-Krämer, Jean-Marc Ogier

PII: S0031-3203(16)30339-9
DOI: http://dx.doi.org/10.1016/j.patcog.2016.10.023
Reference: PR5931

To appear in: *Pattern Recognition*

Received date: 3 June 2016
Revised date: 17 October 2016
Accepted date: 19 October 2016

A comprehensive survey of mostly textual document segmentation algorithms since 2008

Sébastien Eskenazi*,a, Petra Gomez-Krämer*,a, Jean-Marc Ogier*,a

[a] L3i laboratory - La Rochelle University, Avenue Michel Crépeau, 17042 La Rochelle, France

Abstract

In document image analysis, segmentation is the task that identifies the regions of a document. The increasing number of applications of document analysis requires a good knowledge of the available technologies. This survey highlights the variety of the approaches that have been proposed for document image segmentation since 2008. It provides a clear typology of documents and of document image segmentation algorithms. We also discuss the technical limitations of these algorithms, the way they are evaluated and the general trends of the community.

Key words: Document, Segmentation, Survey, Evaluation, Trends, Typology

1. Introduction

Industrial document digitization, document archiving with destruction of the original copy and security technologies based on document processing create an increasing need for reliable document processing algorithms. A thorough list of the available algorithms would be of great use to choose them correctly. A typical paper document content extraction process is shown in Figure 1. Document segmentation aims at dividing the document image into meaningful parts. These parts can be glyphs, words, text lines, paragraphs, regions (usually with one type of content such as text or graphic). These parts are usually used

*Email: {sebastien.eskenazi, petra.gomez, jean-marc.ogier}@univ-lr.fr