

SpringerBriefs in Computer Science

Showmik Bhowmik



# Document Layout Analysis

# **SpringerBriefs in Computer Science**

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic.

Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8–12 weeks after acceptance. Both solicited and unsolicited manuscripts are considered for publication in this series.

**\*\*Indexing: This series is indexed in Scopus, Ei-Compendex, and zbMATH \*\***

Showmik Bhowmik

# Document Layout Analysis

 Springer

Showmik Bhowmik  
Department of Computer Science  
and Engineering  
Ghani Khan Choudhury Institute of  
Engineering and Technology  
Malda, West Bengal, India

ISSN 2191-5768 ISSN 2191-5776 (electronic)  
SpringerBriefs in Computer Science  
ISBN 978-981-99-4276-3 ISBN 978-981-99-4277-0 (eBook)  
<https://doi.org/10.1007/978-981-99-4277-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.  
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

Documents often possess information regarding the social, economic, educational, and cultural status of a particular place or person for a specific time period. Historical documents are the witnesses of the developments that have been made in various sectors of a country since the early days. Therefore, different archives are established around the world to preserve these documents in their physical form. However as these were created long years ago, there is a risk to preserve them in their physical form. These documents need to be digitized. Considering this fact, a widespread initiative of converting paper-based documents into electronic documents has been taken. There are lots of advantages of electronic documents over paper documents, such as compact and lossless storage, easy maintenance, efficient retrieval, and fast transmission. However, mere electronic conversion of these documents would not be of pronounced help for properly preserving the documents as well as automatic information retrieval, unless we provide a system for efficiently analysing the layout of these documents.

Documents have an explicit structure; it can be segregated into a hierarchy of physical modules, such as pages, columns, paragraphs, text lines, words, tables, and figures or a hierarchy of logical modules, such as titles, authors, affiliations, abstracts, and sections or both. This structural information would be very beneficial and convenient in indexing and retrieving the information contained in the documents.

The objective of document layout analysis is to detect these physical modules present in an input document image to facilitate effective indexing and information retrieval.

Malda, West Bengal, India

Showmik Bhowmik

# Contents

<b>1</b>	<b>Introduction</b> . . . . .	1
1.1	Document Image Processing System . . . . .	2
1.2	Structure of a Document . . . . .	3
1.3	Categories of Document Layout . . . . .	4
1.4	Document Layout Analysis (DLA) . . . . .	6
1.5	Why DLA Is Still an Open Area of Research . . . . .	8
	References . . . . .	10
<b>2</b>	<b>Document Image Binarization</b> . . . . .	11
2.1	Different Types of Degradations and Noise . . . . .	12
2.2	Pre-processing . . . . .	15
2.3	Document Image Binarization . . . . .	16
2.4	Different Binarization Methods . . . . .	17
2.4.1	Threshold-Based Methods . . . . .	17
2.4.2	Optimization Based Methods . . . . .	19
2.4.3	Classification-Based Methods . . . . .	20
2.5	Evaluation Techniques . . . . .	22
2.5.1	Standard Databases for DIB . . . . .	22
2.5.2	Performance Metrics . . . . .	24
	References . . . . .	26
<b>3</b>	<b>Document Region Segmentation</b> . . . . .	31
3.1	Different Document Region Segmentation Methods . . . . .	31
3.1.1	Pixel Analysis-Based Methods . . . . .	32
3.1.2	Connected Component Analysis-Based Methods . . . . .	34
3.1.3	Local Region Analysis-Based Methods . . . . .	35
3.1.4	Hybrid Methods . . . . .	36
3.2	Available Dataset for Page Segmentation . . . . .	37
3.3	Evaluation Metrics . . . . .	38
	References . . . . .	39

<b>4 Document Region Classification</b> . . . . .	43
4.1 Different Types of Document Regions . . . . .	43
4.2 Different Document Region Classification Methods . . . . .	44
4.2.1 Methods for Text/Non-text Classification . . . . .	44
4.2.2 Methods for Text Region Classification . . . . .	49
4.2.3 Methods for Non-text Classification . . . . .	50
4.3 Evaluation Techniques . . . . .	55
4.3.1 Standard Datasets . . . . .	55
4.3.2 Evaluation Metrics . . . . .	57
References . . . . .	58
<b>5 Case Study</b> . . . . .	67
5.1 Analysis of Basic Contents in Documents (ABCD) . . . . .	67
5.1.1 Pre-processing . . . . .	68
5.1.2 Non-text Suppression and Noise Removal . . . . .	69
5.1.3 Text Region Generation . . . . .	70
5.1.4 Non-text Classification . . . . .	73
5.1.5 Experimental Results . . . . .	74
5.2 BINYAS . . . . .	77
5.2.1 Pre-processing . . . . .	77
5.2.2 Isolation of Separators . . . . .	77
5.2.3 Layout Complexity Estimation . . . . .	78
5.2.4 Separation of Large and Small Components . . . . .	78
5.2.5 Text and Non-text Separation . . . . .	78
5.2.6 Thickness Based Text Separation . . . . .	79
5.2.7 Text Region Segmentation . . . . .	80
5.2.8 Non-text Classification . . . . .	80
References . . . . .	81
<b>6 Summary</b> . . . . .	83
References . . . . .	86



# Chapter 1

## Introduction



**Abstract** Documents often carry crucial information, covering almost every aspect of human society. Even in this era of cyber-physical systems, many of us prefer to read paper documents. These facts raise the concern of careful storage and management of these documents. However, preserving these documents in their physical form has many risks and also limit their access. To address these problems, it is required to convert these paper-based documents into their electronic form. However, mere electronic conversion would not be of pronounced help for properly preserving the documents as well as automatic information retrieval. Therefore, a system for efficiently analyzing the layout of these documents becomes a pressing need. In this chapter, a quick introduction to Document layout analysis and its constituent stages are presented. This chapter also discusses various challenges associated with the task of layout analysis.

**Keywords** Document image processing · Document layout analysis · Binarization · Manhattan layout · Non-Manhattan layout · Overlapping layout · DIBCO · Document understanding

From ancient times documents have been used as an important medium of storing and conveying information. Even in this era of cyber-physical systems, many of us prefer to read paper documents. This common preference of human society causes a steady growth in the production of such documents from ancient times. Besides, every document generated for certain purposes irrespective of time always possesses important information. For example, historical documents often carry important historic information regarding a place, person, and event for a certain period. Whereas contemporary documents like books, magazines, reports, and other types cover the current economic, social and cultural status. Therefore, these documents should be carefully stored and managed. However, stockpiling of such ever-increasing paper documents is quite unwieldy. There is always a risk to preserve these in their physical form. These may get lost due to aging, casual handling, natural calamity, and many other reasons. Additionally, such arrangements cause limited access to these documents. For example, despite the importance of historical