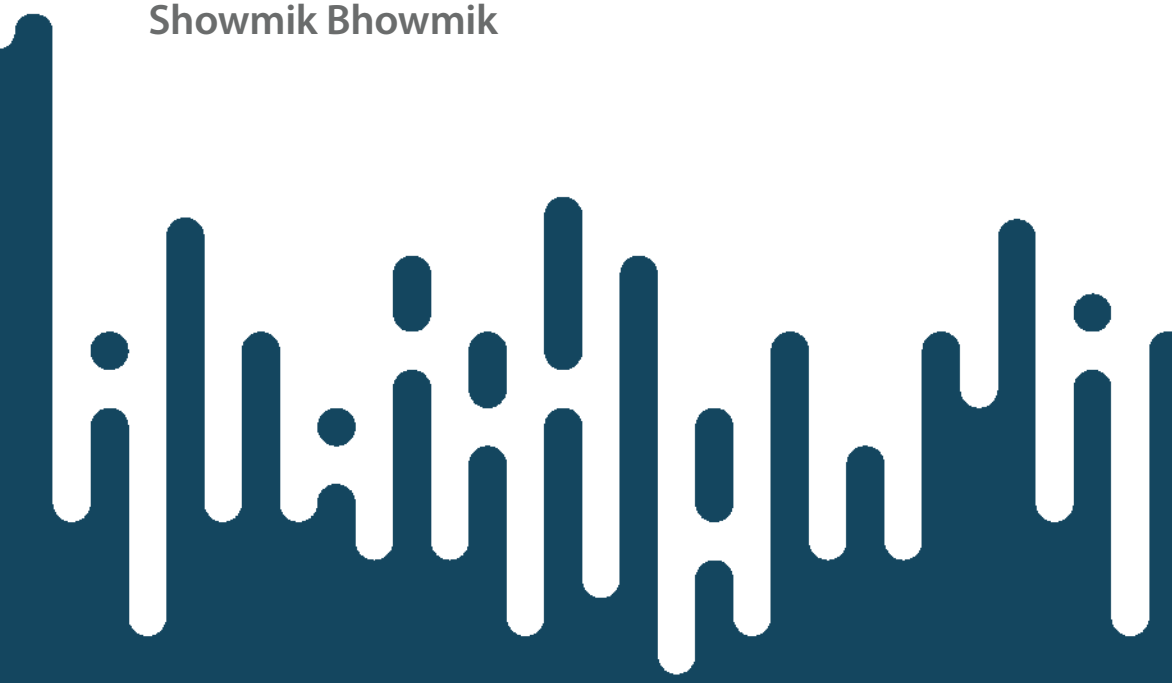
SpringerBriefs in Computer Science

Showmik Bhowmik



Document Layout Analysis

Springer

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

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 Historical documents are the witnesses of the developments that have been made in various sectors of a country since the early days.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

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

S. Bhowmik, *Document Layout Analysis*, SpringerBriefs in Computer Science,
https://doi.org/10.1007/978-981-99-4277-0_1