

Preface

Automatic document processing plays a crucial role in the present society, due to the progressive spread of computer-readable documents in everyday life, from informal uses to more official exploitations. This holds not only for new documents, typically born digital, but also for legacy ones that undergo a digitization process in order to be exploited in computer-based environments. In turn, the increased availability of digital documents has caused a corresponding increase in users' needs and expectations. It is a very hot topic in these years, for both academy and industry, as witnessed by several flourishing research areas related to it and by the ever-increasing number and variety of applications available on the market. Indeed, the broad range of document kinds and formats existing today makes this subject a many-faceted and intrinsically multi-disciplinary one that joins the most diverse branches of knowledge, covering the whole spectrum of humanities, science and technology. It turns out to be a fairly complex domain even focusing on the Computer Science perspective alone, since almost all of its branches come into play in document processing, management, storage and retrieval, in order to support the several concerns involved in, and to solve the many problems raised from, application to real-world tasks. The resulting landscape calls for a reference text where all involved aspects are collected, described and related to each other.

This book concerns *Automatic Digital Document Processing and Management*, where the adjective 'digital' is interpreted as being associated to 'processing and management' rather than to 'document', thus including also digitized documents in the focus of interest, in addition to born-digital ones. It is conceived as a survey on the different issues involved in the principal stages of a digital document's life, aimed at providing a sufficiently complete and technically valid idea of the whole range of steps occurring in digital document handling and processing, instead of focusing particularly on any specific one of them. For many of such steps, fundamentals and established technology (or current proposals for questions still under investigation) are presented. Being the matter too wide and scattered, a complete coverage of the significant literature is infeasible. More important is making the reader acquainted of the main problems involved, of the Computer Science branches suitable for tackling them, and of some research milestones and interesting approaches

available. Thus, after introducing each area of concern, a more detailed description is given of selected algorithms and techniques proposed in this field along the past decades. The choice was not made with the aim of indicating the best solutions available in the state-of-the-art (indeed, no experimental validation result is reported), but rather for the purpose of comparing different perspectives on how the various problems can be faced, and possibly complementary enough to give good chance of fruitful integration.

The organization of the book reflects the natural flow of phases in digital document processing: acquisition, representation, security, pre-processing, layout analysis, understanding, analysis of single components, information extraction, filing, indexing and retrieval. Specifically, three main parts are distinguished:

Part I deals with digital documents, their role and exploitation. Chapter 1 provides an introduction to documents, their history and their features, and to the specific digital perspective on them. Chapter 2 then overviews the current widespread formats for digital document representation, divided by category according to the degree of structure they express. Chapter 3 discusses technological solutions to ensure that digital documents can fulfill suitable security requirements allowing their exploitation in formal environments in which legal issues come into play.

Part II introduces important notions and tools concerning the geometrical and pictorial perspective on documents. Chapter 4 proposes a selection of the wide literature on image processing, with specific reference to techniques useful for handling images that represent a whole digitized document or just specific components thereof. Chapter 5 is devoted to the core of processing and representation issues related to the various steps a document goes through from its submission up to the identification of its class and relevant components.

Part III analyzes the ways in which useful information can be extracted from the documents in order to improve their subsequent exploitation, and is particularly focused on textual information (although a quick glance to the emerging field of image retrieval is also given). Chapter 6 surveys the landscape of Natural Language Processing resources and techniques developed to carry out linguistic analysis steps that are preliminary to further processing aimed at content handling. Chapter 7 closes the book dealing with the ultimate objective of document processing: being able to extract, retrieve and represent, possibly at a semantic level, the subject with which a document is concerned and the information it conveys.

Appendices A and B briefly recall fundamental Machine Learning notions, and describe as a case-study a prototypical framework for building an intelligent system aimed at merging in a tight cooperation and interaction most of the presented solutions, to provide a global approach to digital documents and libraries management.

The book aims at being self-contained as much as possible. Only basic computer science and high-school mathematical background is needed to be able to read and understand its content. General presentation of the various topics and more specific aspects thereof are neatly separated, in order to facilitate exploitation by readers interested in either of the two. The technical level is, when needed, sufficiently detailed to give a precise account of the matter presented, but not so deep and pervasive as to discourage non-professionals from usefully exploiting it. In particular,

most very technical parts are limited to sub-subsections, so that they can be skipped without losing the general view and unity of the contents. To better support readers, particular care was put in the aids to consultation: the index reports both acronyms and their version in full, and in case of phrases includes entries for all component terms; the glossary collects notions that are referred to in different places of the book, so that a single reference is provided, avoiding redundancy; the acronym list is very detailed, including even items that are used only once in the text but can be needed in everyday practice on document processing; the final Reference section collects all the bibliography cited in the text.

The main novelty of this book lies in its bridging the gaps left by the current literature, where all works focus on specific sub-fields of digital document processing but do not frame them in a panoramic perspective of the whole subject nor provide links to related areas of interest. It is conceived as a monograph for practitioners that need a single and wide-spectrum *vade-mecum* to the many different aspects involved in digital document processing, along with the problems they pose, noteworthy solutions and practices proposed in the last decades, possible applications and open questions. It aims at acquainting the reader with the general field and at being complemented by other publications reported in the References for further in-depth and specific treatment of the various aspects it introduces. The possible uses, and connected benefits, are manifold. In an academic environment, it can be exploited as a textbook for undergraduate/graduate courses interested in a broad coverage of the topic.¹ Researchers may consider it as a bridge between their specific area of interest and the other disciplines, steps and issues involved in Digital Document Processing. Document-based organizations and final users can find it useful as well, as a repertoire of possible technological solutions to their needs.

Although care has been put on thorough proof-reading of the drafts, the size of this work makes it likely that some typos or other kinds of imprecisions are present in the final version. I apologize in advance for this, and will be grateful to anyone who will notify me about them.

Bari, Italy

Stefano Ferilli

¹The included material is too much for a semester, but the teacher can select which parts to stress more and which ones to just introduce.



<http://www.springer.com/978-0-85729-197-4>

Automatic Digital Document Processing and
Management

Problems, Algorithms and Techniques

Ferilli, S.

2011, XXVI, 297 p., Hardcover

ISBN: 978-0-85729-197-4