https://doi.org/10.1142/9789811203244_0001

Chapter 1

Introduction

Andreas Fischer, Marcus Liwicki, and Rolf Ingold

Document image analysis or document recognition refers to the process of extracting valuable information from document images. Although a few optical character reading systems were already available in the 1970's, the fundamental research activities on this challenging task has mainly emerged with the development of the scanner technologies in the 1980's, which allowed affordable document image acquisition. At that time, the main applications were focused on office automation and the interpretation of printed material.

Nowadays, the technology for printed text recognition is considered as mature, even if several unsolved issues remain, typically for processing tables or mathematical formulas. However, the interest for these topics is decreasing because of the gradual reduction of paper documents used for the daily business.

In the meanwhile, a new even more challenging application area has emerged: the analysis and recognition of historical documents. Since the beginning of the new millennium, huge efforts have been made for digitizing historical documents at large scale. The main motivations for such initiatives are twofold: first, the preservation of the cultural heritage, reducing cost and avoiding additional degradations since digital information is supposed to be unalterable; and second, to make these documents available and searchable to a large community of users and notably in human sciences for scholars.

Numerous collections of historical document facsimiles have been made available on the Internet, often linked with additional descriptors and metadata, which provide additional functionalities for selecting, searching, and comparing documents according to these descriptors. However, full text transcription is rarely provided and if it is, it has usually been produced manually.

We could expect that computers should be able to do much more with these documents. It would be desirable to be able to search for historical documents by using an engine like Google, either locally on a given collection or globally working worldwide. Today, such a possibility is considered as a dream, but a dream that will become almost reality in a reasonable future. It is probably a matter of time.

The methods and tools developed for contemporary document analysis are hardly adaptable to be used in the context of historical documents. The goal of this book is to address the new issues that are raised with historical document analysis and to explore the novel research topics the scientific community is now dealing with. It presents the state of the art of several challenges involved. In order to understand the relations between these different tasks, we need to briefly remind the processing chains that are usually applied.

Figure 1.1 shows a typical processing chain composed of several important steps. In this illustration, we voluntarily exclude the image acquisition part, which often requires specific hardware and is therefore considered as a topic for itself. Here, we assume that the images are already available in gray scale or color spaces and focus our discussion on the algorithmic part.

There is not a unique processing chain that is applicable to any kind of documents. The architecture of a document analysis system is highly dependent of the targeted application and the document collection considered for it. Typically for degraded documents, a pre-processing step with sophisticated image enhancement algorithms is required before further processing. At the other end of the chain, post-processing is a matter of cost that must be accommodated to the quality requirements of the results.

However, there are typical steps between that have to be performed in a specific order which are illustrated in Figure 1.1. After the pre-processing stage follows a very important step called layout analysis, which consists in segmenting the entire image into regions of interest. As a preliminary step, foreground and background separation is often required. Then, an important aspect is to separate text from non-text regions (such as ornaments, pictures or large initials), and then to split text blocks into text lines and even

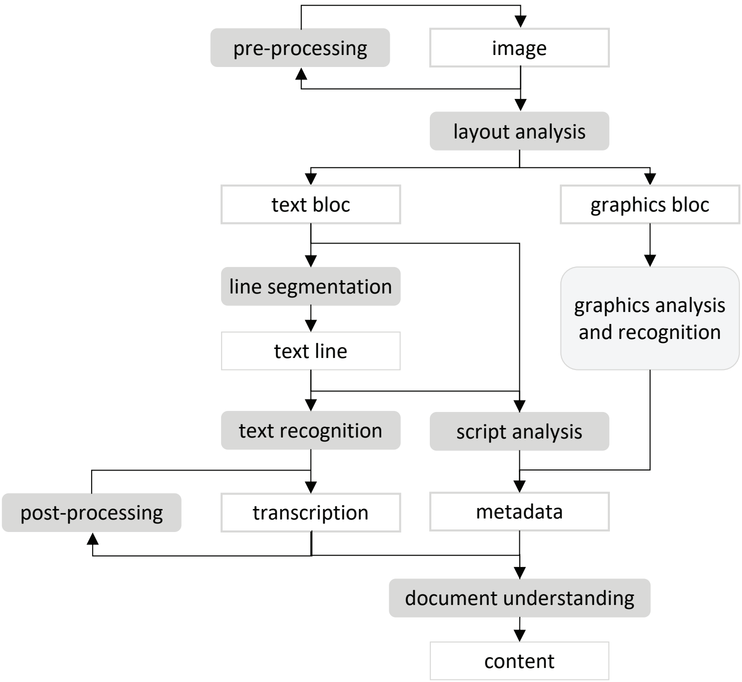

Fig. 1.1: A typical processing chain for historical documents analysis.

further into isolated words. Finally, relations between the regions have to be established for recovering the natural reading order or, in case of a table, to establish the horizontal and vertical alignment properties.

For several reasons, layout analysis of historical documents is much more challenging than for contemporary documents. First, as already mentioned, degradation is a major hindering for accurate analysis and image pre-processing can only hardly improve the image quality and certainly not recover the original information. Additionally, given the preciousness of parchment and paper, the presented information of historical documents is much denser than for contemporary documents. In medieval documents for instance the textual content is very often mixed with ornaments and glosses that have been added later. Classical segmentation algorithms based on spacing properties are not really applicable. Approaches based on texture analysis, inspired by natural image processing seem more appropriate. In the last couple of years, machine learning approaches and especially deep learning has made a considerable progress.

Once the images of text lines or words have been properly extracted, text recognition can be applied. Here we need to distinguish between machine printed and handwritten text. In the first case, OCR technologies are designed to perform text recognition. Most of these software tools are able to cope with almost all kind of alphabets and typefaces. For some specific needs, text recognition might not be enough; as a complementary task it might be important to recognize font variations. In that case, a font recognition module needs to be introduced, either as preliminary task (known as a priori font recognition) or after the text transcription (also named a posteriori font recognition).

Despite the regular progress made over the last two decades, handwritten text recognition has not yet reached a stage to be applicable at a large scale. One of the main reasons for this situation is the fact that characters can only hardly be isolated. Instead of recognizing individual characters (as used in OCR technologies), handwriting recognition needs to be applied on entire words or even sequences of words. This increases immensely the complexity of the task and requires more contextual knowledge, such as language models, dictionaries, or even semantic knowledge relying for instance on domain specific ontologies. To recognize handwritten words, the use of machine learning is essential; hidden Markov models (HMM) or long short-term memory networks (LSTM) seem both to be adapted to model both shape and language properties, provided that enough labeled training data is available, and this is for the moment still the main bottleneck of this technology.

However, for certain document analysis applications, full text transcription is not absolutely required. As an interesting alternative, the so called word spotting approach can be used instead. Here the goal is to recognize and locate specific keywords, reducing considerably the complexity of the task and allowing automatic and accurate indexing by these keywords.

In case of multilingual documents, sometimes by mixing different alphabets, an additional script and language analysis is required. Interestingly, the job can be performed prior to the text recognition task, by extracting relevant features from pixel data. Of course, alternatively, the language can also be recognized by comparing the OCR output with a dictionary. Such an approach is only feasible under the assumption that the text recognition result is accurate enough.

The last important processing step, often considered as optional, is called document understanding. The result of this first step toward interpretation is highly depending of the complexity of the documents and the targeted application. It relies on layout analysis and addresses the logical structure of the document. Concretely, it aims at recognizing titles and headers in order to establish the hierarchy of chapters and sections, of documents. It targets also the recognition of captions, glosses and footnotes. In addition to application-oriented research that aims to develop concrete solutions, the research community also needs to develop appropriate tools to cover its own needs. One major example of that is the development of databases with their ground-truth annotations. This is a crucial element to quantitatively evaluate algorithms and machine learning models and therefore a sound way to stimulate competitions. Providing enough annotated data is far from being trivial; in most cases the ground truth needs to be generated manually by an expert, which is boring, time consuming and therefore expensive. To simplify such work, appropriate tools are developed to automate the process as much as possible, reducing thus the human intervention essentially to the verification and correction of the proposed results.

This book reports about research activities on historical document analysis carried out during the last decade and discusses the results obtained so far. It is composed of two parts. In Part 1 (Chapters 2 to 8), we report about a series of research projects conducted by Swiss universities and largely funded by the Swiss National Science Foundation, which we summarize under the umbrella term “the HisDoc project”. The main objectives of these projects was to develop tools covering the entire processing chain to allow intelligent automatic indexing of handwritten historical documents. Part 2 (Chapters 9 to 12) contains a selection of invited contributions from colleagues from other European countries to give a more global picture of the intense research activities deployed in the domain of cultural heritage preservation.

In Chapter 2, Andreas Fischer describes the effort to develop an initial database that was needed for the HisDoc project to train and evaluate automatic handwriting transcriptions. Three different datasets with great variability have been included. They differ from each other in terms of age (9th, 13th and 18th century), language (Latin, German, and English), layout structure and image quality. The chapter describes not only the ground truth features that are taken into account, but it also presents the different and complex steps that were needed to extract the relevant information. Of particular interest for its originality, we mention here the semi-automatic text alignment process that was used to establish precise links between the text transcription and the text line images.

The contribution of Chapter 3 is provided by Foteini Simistira Liwicki and presents another database developed within the HisDoc project. In contrast with the previous database, the goal of this dataset is intended for layout analysis of medieval documents. The chosen documents belong to a category with very complex layout structures due to a mixture of the original main text and marginal interlinear glosses or corrections that were added later in history. The provided annotations are composed of tight polygons surrounding text lines or graphic elements with labels to discriminate the type of text (main text, glosses or titles). Additionally to the database itself, the chapter also presents protocols and evaluation measures and reports about a competition organized at ICDAR 2017 around this complex task.

Chapter 4 is dedicated to layout analysis. Written by Mathias Seuret, this chapter discusses various techniques used for text line segmentation of medieval documents with complex and irregular layouts taken from the dataset presented in Chapter 3. The focus is put on text line segmentation and the author illustrates the complexity of the tasks and shows that there is not even a strong consensus on the form of the results: the problem can be stated as a splitting task using boundaries or specified as a pixel labeling task. As a consequence, a variety of complementary approaches, either algorithmic or based on machine learning can be applied. Good results are reported by combining a seam carving algorithm with a neural network composed of a stacked autoencoder used for feature extraction and a final pixel classifier.

Text recognition, a topic often seen as the central goal of document analysis is discussed in Chapter 5. Andreas Fischer, the author, shows the inherent difficulty of the topic when applied on historical documents. Even if fully automated transcription will remain a dream for several years, important improvements could be reported. In the HisDoc project, two approaches based respectively on hidden Markov models (HMM) and long short-term memory networks (LSTM) have been evaluated. Both methods rely on the principle that entire word recognition can be performed by combining language models from the transcription with the character models trained on the images. Furthermore, the chapter also addresses briefly the text-image alignment problem.

Chapter 6, written by Volkmar Frinken and Shriphani Palakodety, is dedicated to word spotting, an intriguing alternative to automatic transcription for content based document indexing. It provides an overview of state-of-the-art techniques and describes a complete system well-adapted for word spotting of historical handwriting that was developed during the HisDoc project. The approach is based on sliding windows for feature extraction and on LSTM models for sequence analysis. This system outperforms traditional approaches based on hidden Markov models or dynamic time warping.

The last two chapters related to the HisDoc project present practical tools, which have been developed to better support the research environment useful for the researchers themselves.

The first one, described in Chapter 7 by Marcel Gygli (formerly Marcel Würsch), called DIVAServices is a web-based platform used to access standard algorithms related to document analysis and used for data preparation and management. It includes a large variety of tools ranging from image pre-processing to public OCR and statistical evaluation tools. By sharing such an environment the researchers have no more to deal with the installation and maintenance of all these software packages and can concentrate their efforts on their own contributions.

Finally, Chapter 8, presented by Angelika Garz, describes an innovative interactive tool called GraphManuscribble. Based on a tablet computer with a stylus, it allows to quickly annotate and correct the ground truth for layout analysis of historical documents with complex structures. In a first automatic step, a document graph is computed and overlaid to the original image: the vertices correspond to points of interest extracted from a binarized or gray scale image and the edges correspond to the minimum spanning tree extracted from the Delaunay triangulation. This graph that approximately captures the visual structure of the document, is impaired by a number of incorrect or missing links which would provoke respectively under- and over-segmentation. These errors can then be effectively corrected with a few natural scribbling gestures by a human expert.

The second part of the book contains contributions of several representative research activities conducted in other European countries, which are related to the HisDoc project and aim to achieve similar goals. They describe other fascinating projects on historical analysis and their applicability in many different areas.

Chapter 9 describes a Greek national research project called OldDocPro and dedicated to the automatic indexing of machine-printed and handwritten polytonic historical documents. The main challenge comes from the complexity of the scripts containing great variety of diacritics (more than 270 classes), which the existing OCR technologies cannot manage correctly, even after thorough training. To overcome these difficulties a large specialized database has been created; its ground-truth describes the precise location of more than 100'000 words and more than 170'000 characters. This database was then used to develop several novel methods for text line segmentation, isolated character recognition as well word recognition and keyword spotting.

Chapter 10 is dedicated to a sophisticated keyword spotting system developed at the Polytechnic University of Valencia. The proposed methods are based on a probabilistic approach composed of so called pixel-level "posteriograms" and relevance probabilities estimated on segmented text lines. The methodology has been evaluated by a precision-recall trade-off model and has proven their effectiveness on several large document collections.

The contribution of Chapter 11 comes from the Computer Vision Center CVC of the Autonomous University of Barcelona, another distinguished Spanish research group. It describes a complete system dedicated to the study of historical demography based on marriage and death registers or other similar sources. This interdisciplinary work is based on the analysis of thousands of population records from various sources stored in public, ecclesiastical or private archives and targets the construction of a large knowledge base organized as a historical social network.

Last but not least, Chapter 12 considers the topic of historical document analysis from the big data perspective with the "four V" principles: Volume, Velocity, Variability and Veracity. Lambert Schomaker from the University of Groningen (Netherlands) reports about the observations made with the Monk system, an e-Science service designed for scholars, which is running for more than a decade. The contribution focuses mainly on different strategies for semi-automatic word labeling and draws very interesting conclusions extending thus the vision that is commonly shared by the scientific community dealing with historical document analysis.