

Preface

Pattern recognition systems usually consist of two main parts. On the one hand, the data acquisition and learning stage and, on the other hand, the classification of this data to a certain category. In order to recognize which category a certain query element belongs to, a set of pattern models must be provided beforehand. An off-line learning stage is needed to train the classifier and to offer a robust classification of the patterns. Within the pattern recognition field, we are interested in the document image analysis topic and, in particular, in the recognition of graphics appearing within documents rich in graphical information. In the particular case of graphical symbol recognition, descriptors are extracted from the symbol to recognize and are subsequently matched with the set symbol models. In this context, one of the main concerns is to see if the proposed systems remain scalable with respect to the data volume to be able to handle growing number of symbol models. In order to avoid working with a database of reference symbols, symbol spotting and on-the-fly symbol recognition methods have been introduced in the past years.

Generally speaking, the symbol spotting problem can be defined as the identification of a set of regions of interest from a document image, which are likely to contain an instance of a certain queried symbol without explicitly applying the whole pattern recognition scheme. Our application framework consists in indexing a collection of graphic-rich document images. This collection is queried by example with a single instance of the symbol to look for and, by means of symbol spotting methods, to retrieve the regions of interest where the symbol is likely to appear within the documents. This kind of applications are known as focused retrieval methods.

In order that the focused retrieval application can handle large collections of documents, there is a need to provide an efficient access to the large volume of information that might be stored. Indexing strategies are used in order to efficiently retrieve by similarity the locations where a certain part of the symbol appears. In that scenario, graphical patterns should be used as indices for accessing and navigating the collection of documents. These indexing mechanisms allow the user to search for similar elements using graphical information rather than textual queries.

In this book, we present a spotting architecture and different methods aimed at building a complete focused retrieval application dealing with a graphic-rich document collections.

Different symbol descriptors encoding geometric and structural information are proposed. These descriptors aim at describing parts of the symbols in a very compact and efficient way. Vectorial signatures, attributed strings and off-the-shelf shape descriptors are used to cluster parts of the symbols by similarity.

Several strategies to search for graphical information by similarity are used in this book. In order to retrieve locations from the document collection where parts of the symbols appear, we use lookup tables and grid files indexed by graphical patterns. A final validation phase is introduced to validate the hypothetic locations where a symbol is likely to be found. This validation stage is formulated in terms of spatial and relational information.

In addition, a protocol to evaluate the performance of symbol spotting systems in terms of recognition abilities, location accuracy and scalability is also studied. Evaluation measures allowing to determine the weaknesses and strengths of the methods under analysis are presented. All the methods under analysis have been tested on an experimental scenario consisting of a collection of architectural drawings with its corresponding ground-truth.

Structure

This book is divided into four parts. Part I is of introductory nature consisting of two chapters. Chapter 1 presents the symbol spotting and focused retrieval problems and outlines the proposed architecture. Chapter 2 reviews the related work to symbol spotting which has been proposed in the last years.

Part II is centered on the application of well-known methods of Computer Vision for recognizing objects in scenes to the specific problem of spotting graphical symbols in documents. Chapter 3 presents, as a running example, an application of logo spotting for a document categorization application. The method processes incoming document images such as invoices or receipts. The categorization of these document images is done in terms of the presence of a certain graphical entity detected without segmentation.

Part III is centered on the use of geometrical and structural constraints as symbol description techniques. Chapter 4 presents a method to determine which symbols are probable to be found in technical drawings by the use of vectorial signatures as symbol descriptors. Chapter 5 presents a spotting method which uses a prototype-based search as the basis for the focused retrieval task. Finally, Chapter 6 presents an indexing method to retrieve locations of interest where a query symbol is likely to be found. In order to foster the querying speed, a hashing technique is used in order to retrieve primitives by similarity very efficiently.

Part IV including just Chapter 7 is centered on the performance evaluation of spotting systems. Since symbol spotting systems and focused retrieval applications shall have the ability to recognize and locate graphical symbols in a single step, the measures to evaluate the performance of a symbol spotting system are defined in terms of recognition abilities, location accuracy and scalability.

Finally, Chapter 8 gives some concluding remarks about this study, and specifies some possible future research lines on symbol spotting techniques. Throughout this book, different symbolic databases have been used to perform the experiments. All these databases are explained in Appendix A.

Audience

This book is intended for researchers and practitioners from the field of graphics recognition who are interested in the problem of symbol spotting and focused retrieval applications in the context of digital libraries. Some basic knowledge of pattern recognition, document image analysis and graphics recognition is assumed.

Acknowledgments

We would like to thank the friends and colleagues from the Computer Vision Center. It is a pleasure to share research and teaching with all of them. Special thanks go to Prof. Juanjo Villanueva for making all this possible. We also thank the people from the Document Analysis Group for their insightful comments on our work. In particular, the authors would like to thank Dr. Ernest Valveny, Dr. Dimosthenis Karatzas, Dr. Alicia Fornés, Dr. Agnès Borràs, and Joan Mas, to name just a few.

The authors would also like to thank Prof. Karl Tombre and Dr. Philippe Dosch from the LORIA laboratory in Nancy, France, and Prof. Jean-Marc Ogier and Dr. Karell Bertet from the Université de La Rochelle, France, for their collaboration in many aspects of this work, and for hosting us during many research visits.

We would also like to acknowledge the Spanish Ministry of Research for funding this work through the grants TIN2006-15694-C02-02, TIN2009-14633-C03-03 and CONSOLIDER – INGENIO 2010 (CSD2007-00018).

Barcelona, Spain

Marçal Rusiñol
Josep Lladós

Symbol Spotting in Digital Libraries
Focused Retrieval over Graphic-rich Document
Collections

Rusiñol, M.; Lladós, J.

2010, XIV, 180 p., Hardcover

ISBN: 978-1-84996-207-0