

Tools for Document Image Retrieval in Digital Libraries: the AIDI System

Simone Marinai, Giovanni Soda

Dipartimento di Sistemi e Informatica
University of Florence, Italy

In the last few years, Digital Libraries became one important application area for Document Image Analysis and Recognition research [1]. In this field, a relevant line of research is Document Image Retrieval (DIR) that aims at finding relevant documents relying on image features only. DIR techniques are used to index not only the textual content of a document, but also its layout, graphical objects, mathematical equations, and handwritten text. By integrating these capabilities with other traditional indexing approaches we expect it will be possible to define new search strategies that could be especially useful in scientific and technical collections. In this abstract we summarize our recent research on Document Image Retrieval techniques in the field of Digital Libraries that we integrated in the AIDI prototype system.

1 Document Image Retrieval techniques

Document Image Retrieval aims at finding relevant documents from a corpus of digitized pages relying on image features only and is closely related to Content-Based Image Retrieval (CBIR) [2] [3]. Important sub-tasks include the retrieval of documents on the basis of layout similarity and on the basis of the textual content.

1.1 Word indexing

One very important sub-topic of text-based DIR is word-level indexing, that addresses the efficient identification of the occurrences of a given word in the indexed documents (e.g. [4] [5] [6] [7]). When the use of OCR is not advisable, either due to the low quality of images or the use of uncommon fonts, then image-based word retrieval is a viable alternative. In methods based on character-like coding some objects (that might correspond to characters) are extracted from each word. The word is then represented by concatenating the codes assigned to the objects on the basis of shape similarity [5]. The word indexing implemented in the AIDI system relies on character-like coding and is described with more details in [6].

1.2 Layout Indexing

The layout of a page conveys some semantics that is important for both scholars and general readers, but is often neglected by DL's indexing approaches. For instance, users could be interested on identifying the pages having a *marginalia*

in the right side, or would like to retrieve a page containing a figure on some specific position in the left column in the page. Another example is the retrieval of the title page of scientific papers that, starting from a general structure, can have different actual layouts. Some systems have been proposed in the past to index various types of documents, such as forms [8] [9] and journal papers [10].

In the AIDI system we represented the page layout with a hierarchical description based on the XY tree [11]. XY trees have been demonstrated to be useful when dealing with documents containing ruling lines and can deal with multi-column pages as well as with pages where the pictures cover more text columns [10]. Leaves of the tree correspond to homogeneous regions in the page. To perform the page retrieval, the MXY trees are encoded into a fixed-size representation that is subsequently used to rank the pages. The layout indexing implemented in the AIDI prototype has been described in various papers (e.g. see [10] [12]).

1.3 Early printed books

Early printed books, such as the Latin Gutenberg Bible, look very similar to medieval manuscripts, since they contain illuminated letters (hand painted) and several ligatures and abbreviations that were standard in manuscript writing and have been slowly abandoned in the technological progress of printing. Indexing early printed documents is therefore a task that is closely related to handwriting indexing. To demonstrate the feasibility of these approaches, we recently addressed the indexing and retrieval of the Gutenberg Bible with a Query by Example (QbE) retrieval mechanism implemented in a prototype tool [13].

1.4 Mathematical Symbol Indexing

The recognition of mathematical symbols is particularly difficult for three main reasons: the very large number of symbol classes, the reduced script size for superscripts and subscripts, and the lack of linguistic tools (such as dictionaries) that could help in the recognition. Document image retrieval techniques have been seldom used to process mathematical expressions. However, several researchers envisage the utility of math search systems that would be able not only to search for text, but also for “fine-grain mathematical data” (e.g. equations and functions) [14]. In [15] we recently described our current work on the development of one mathematical symbol indexing and retrieval module.

2 The AIDI system

The Automatic Indexing of Document Images (AIDI) system, that has been developed by our research group, integrates a font-independent word indexing and a layout-based document retrieval into a unique framework [10] [16]. Figure 1 shows a snapshot of the AIDI user interface. On the top-left there are ten thumbnails that contain either the browsed pages or the retrieval results. The image on the right is one selected page that can be further enlarged in the zoom area. The bottom-left part contains the buttons used to perform the queries. In

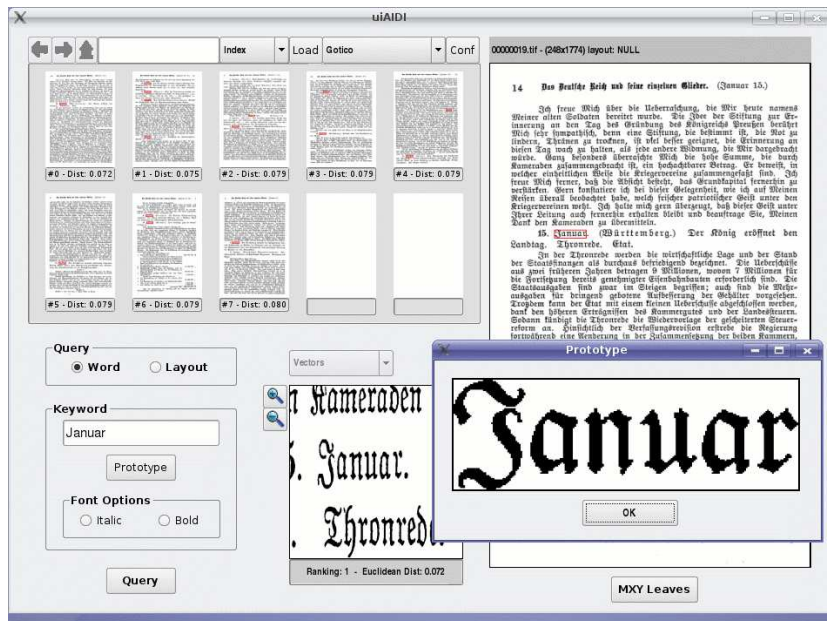


Fig. 1. The user interface of the AIDI system.

the case of textual queries the user enters the query word in the appropriate field. The “Prototype” window shows the generated prototype (in this case a word printed with the Gothic font). Layout-based queries are made with a QbE approach. Therefore, the user selects a page of interest from the list of thumbnails and performs the query by pressing the appropriate button.

During the indexing, the pages are first processed by a layout analysis tool that extracts homogeneous regions. Textual regions are subsequently analyzed so as to identify the words, that are encoded with appropriate character labels. At the same time the layout is encoded to obtain a page-level representation of the documents. The pages can be retrieved by taking into account both textual and layout queries. In the first case, a query prototype is obtained by rendering the word entered by the user with the \LaTeX package (see the “Prototype” window in the Figure). The prototype is encoded similarly to the indexed words that are lastly ranked according to their similarity with the query. Likewise, a query page can be represented in the same way of indexed pages that can be ranked according to their layout similarity.

3 Conclusions

Image-based indexing techniques can be adopted for large collections only if scalable approaches are available to index feature vectors. We are currently testing some indexing techniques to the problem of word image indexing with interesting results for both effectiveness and efficiency of the retrieval [17].

References

1. Baird, H.S.: Digital libraries and document image analysis. In: Int'l Conference on Document Analysis and Recognition. (2003) 2–14
2. Doermann, D.: The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding* **70**(3) (June 1998) 287–298
3. Mitra, M., Chaudhuri, B.: Information retrieval from documents: A survey. *Information Retrieval* **2**(2/3) (2000) 141–163
4. Rath, T.M., Manmatha, R.: Word spotting for historical documents. *IJDAR* **9**(2-4) (2007) 139–152
5. Lu, S., Li, L., Tan, C.L.: Document image retrieval through word shape coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(11) (Nov. 2008) 1913–1918
6. Marinai, S., Marino, E., Soda, G.: Font adaptive word indexing of modern printed documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(8) (2006) 1187–1199
7. Meshesha, M., Jawahar, C.V.: Matching word images for content-based retrieval from printed document images. *IJDAR* **11**(1) (2008) 29–38
8. Liu, J., Jain, A.: Image-based form document retrieval. *Pattern Recognition* **33** (2000) 503–513
9. Duygulu, P., Atalay, V.: A hierarchical representation of form documents for identification and retrieval. *IJDAR* **5**(1) (November 2002) 17–27
10. Marinai, S., Marino, E., Cesarini, F., Soda, G.: A general system for the retrieval of document images from digital libraries. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), 23-24 January 2004, Palo Alto, CA, USA, IEEE Computer Society (2004) 150–173
11. Nagy, G., Seth, S.: Hierarchical representation of optically scanned documents. In: Int'l Conference on Pattern Recognition. (1984) 347–349
12. Marinai, S., Marino, E., Soda, G.: Tree clustering for layout-based document image retrieval. In: Proc. 2nd International Workshop on Document Image Analysis for Libraries (DIAL 2006), 27-28 April 2006, Lyon France, IEEE Computer Society. (2006) 243–251
13. Marinai, S.: Text retrieval from early printed books. In: Third Workshop on Analytics for Noisy Unstructured Text Data, ACM Press (2009) 33–40
14. Youssef, A.: Roles of math search in mathematics. In: Mathematical Knowledge Management MKM 2006, Springer Verlag- LNCS 4108 (2006) 2–16
15. Marinai, S., Miotti, B., Soda, G.: Mathematical symbol indexing using topologically ordered clusters of shape context. In: Proc. 10th Int'l Conference on Document Analysis and Recognition, Washington, DC, USA, IEEE Computer Society (2009) 1041–1045
16. Marinai, S., Marino, E., Soda, G.: Exploring digital libraries with document image retrieval. In Kovács, L., Fuhr, N., Meghini, C., eds.: ECDL. Volume 4675 of Lecture Notes in Computer Science., Springer (2007) 368–379
17. Marinai, S., Marino, E., Soda, G.: Embedded map projection for dimensionality reduction-based similarity search. In: Proc. Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, SSPR/SPR. (2008) 582–591