A Survey of Document Image Retrieval in Digital Libraries

Simone Marinai

A Survey of Document Image Retrieval in Digital Libraries

Simone Marinai

Dipartimento di Sistemi e Informatica
University of Florence, Italy

marinai@dsi.unifi.it

Abstract :

In this paper, we analyze the current trends in the applications of Document Image Retrieval techniques in the field of Digital Libraries. We present the different techniques in a single framework in which the emphasis is put on the representation level at which the similarity between the query and the indexed documents is computed.

Keywords : Digital Library, Document Image Retrieval, Handwriting, Layout Analysis, OCR.

1 Introduction

In the last few years, Digital Libraries (DL) became one important application area for Document Image Analysis and Recognition (DIAR) research. This new trend in the DIAR research is demonstrated by the large number of papers related to DLs that have been published in conferences and journals. An important line of research is Document Image Retrieval (DIR) that aims at finding relevant documents relying on image features only. Until today, the largest portion of documents belonging to libraries is made by printed books and journals. The electronic counterparts of these physical objects are scanned documents that are traditionally the main subject of DIAR research.

Information Retrieval (IR) is one of the principal components of a modern digital library and it is an important factor for building efficient DLs [BAI 03]. The first applications of computers in libraries were related to the MARC (MAchine Readable Cataloging) standard introduced to define a shared format for library records (e.g. [ARM 00]). The MARC files were first shared by exchanging data on magnetic supports, whereas in the 1990s the Internet became the preferred way to exchange bibliographic information as well as to distribute this information to final users. The next step, from a conceptual point of view, was the creation of databases of abstracts typed by hand in machine readable format (mostly in ASCII). The size of these databases grew quickly thus providing a perfect application domain for IR. In the mid 1980s some of the largest libraries installed local computers with the aim of allowing full-text search (through catalog indexes or collections of abstracts) to local users. The next steps of this story are well known: in the mid 1990s the field of digital libraries exploded and several institutions began large digitization programs with the aim of preserving rare holdings for future generations, and ease the access to these collections.

Today, several factors influence the current state of digital libraries and the interest related to the DIR applications. First, the generalized reduction of the budgets lead to a decrease of the number of digitization projects. Second, beginning from the late 1990s many key publishers in several disciplines began distributing on-line "digital-born" documents thus eluding the need for retro-conversion of document's contents. Third, in December 2004 Google announced its *Google Print Library Project* with the ambitious task of digitizing and make available over the Web large portions of the print book collections of five major research libraries in US and UK. These factors could give us a perspective of uselessness of the DIR research in digital libraries.

In this paper, we briefly review the current research in DIR with special interest in applications to digital libraries. We also provide pointers to the most promising research directions for the near future.

The organization of the literature adopted in this paper is described in Section 2. In Section 3 we shortly recall the approach used to find interesting information in libraries. The following two sections analyze the main strategies that can be adopted to use document image analysis techniques to perform DIR. We close the paper discussing the performance evaluation and some research directions.

2 Retrieval paradigms

From a high level point of view, the document retrieval from digital libraries relyes on three main steps: document storage (or indexing), query formulation, and similarity computation with subsequent ranking of the indexed documents with respect to the query. All the retrieval approaches proposed so far can be described on the basis of these three components and the main difference between the methods is the "level" at which the similarity computation occurs. To explain this point of view, in Figure 1 we summarize the main steps performed during indexing as well as some interactions with the retrieval system from the user point of view.

Even the traditional access to libraries can be interpreted under this framework (Section 3): in this case the similarity is evaluated by the user when browsing physical catalog cards or physical books.

The first main stream of approaches adopted to perform DIR was based on a similarity computation at the symbolic level (Section 4). These methods assume that a recognition engine can extract all the information from the digitized documents and possible errors will not affect too much the retrieval performance. The methods belonging to this category have two main advantages; first, they are easy to integrate into a standard IR framework (usually based on ASCII text);

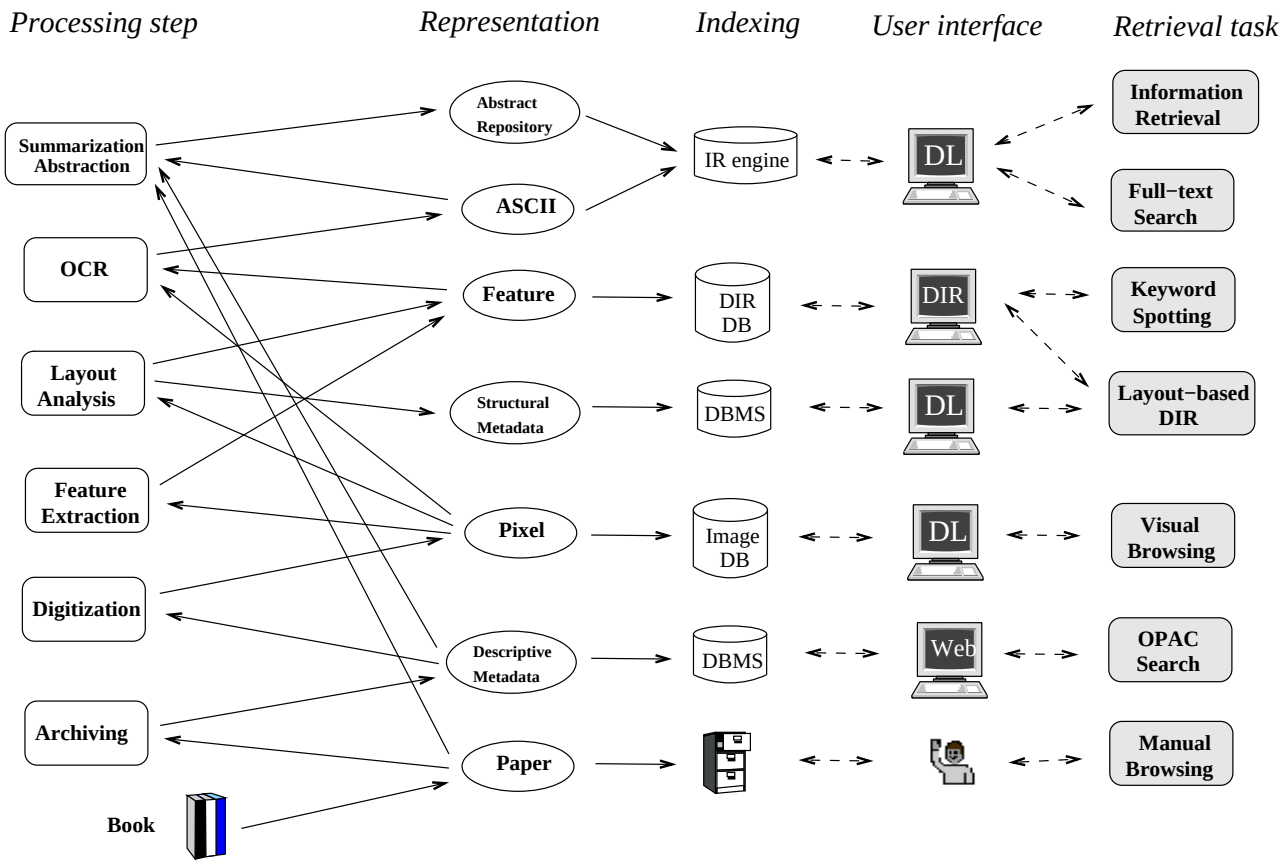

Figure 1: User interaction in retrieval from libraries. Left: the operations performed (either manually or automatically) during the indexing. Right: some retrieval strategies with the corresponding user interfaces (**DL**: Digital Library, **Web**: a web interface to the library catalog, **DIR**: Document Image Retrieval interface).

second, the similarity computation and results ranking has a lower computational cost. However, the recognition-based approach has some limitations when dealing with documents having a high level of noise or containing multi-lingual text printed with non-standard fonts with a variable layout. All the latter problems are peculiar to old documents (ancient or early modern) that populate most libraries.

To solve these problems several recognition-free approaches have been proposed in the last few years with encouraging results (Section 5). Two factors pushed this stream of research: the increased performance of modern computers that allows us to adopt more expensive matching algorithm, and a new interest for the processing of historical documents. Unfortunately, most approaches belonging to the latter category are still at the prototypical stage and their integration into standard DLs frameworks appears to be quite problematic.

3 Search in libraries

Search in libraries has been performed for a long time by using catalog cards and manually encoded information such as printed bibliographies or collections of abstracts ("Manual Browsing" in Fig.1). The direct evolution of this approach was the electronic storage of descriptive meta-data collected with a manual process. From the user point of view the interface with this information is made through "OPAC search" (Fig. 1).

One of the main services of the first digital libraries was the ability to browse collections (still identified through OPAC indexes) by looking at individual pages, stored in appropriate image databases, with the help of customized user interfaces ("Visual Browsing", Fig.1). The most significant services of these DLs were the use of IR techniques associated with full-text search dealing with manually encoded information. However, the cost of the text encoding bounds the size of collections that can be accessed in this way. Therefore, the full-text search can be performed only for few documents.

To allow the full-text search from large collections, OCR packages have been adopted. These approaches follow the recognition-based framework that is analyzed in the next section.

4 Recognition-based retrieval

The first DIR applications in digital libraries were based on the recognition-based paradigm, where document image analysis techniques (and mostly OCR packages) are used to recognize the informative content in the documents to be archived. By taking into account the general framework described in Section 2 in this class of methods the similarity is computed either at the symbolic (ASCII text and structural meta-data) or at the semantic level (abstracts). Some of the earliest methods adopted for the recognition-based approach, and in particular for the text retrieval re-

lying on OCR, have been described in two comprehensive surveys [DOE 98, MIT 00].

4.1 Retrieval from OCRed documents

It is widely known that state of the art OCR packages can recognize documents printed using the most common Latin scripts with a few errors in each page when the quality of the images is quite high and the layout of the page obeys to standard styles. Several strategies have been proposed to deal with OCR errors [MAR 97, TAG 96]. In most approaches the uncorrected OCR output is used for text indexing and words are compared with the query by means of *string edit distance* algorithms. This strategy has been adapted by introducing "ad hoc" edit costs for most common OCR errors (e.g. [LOP 96]).
If there is no need to exactly index each printed word, but the purpose of the word recognition is limited to perform IR, then several studies demonstrated that a low number of OCR errors are not too problematic [TAG 96, MAR 97]. However some problems remain for short texts where the redundancy cannot be exploited. Recent investigations furtherly analyzed the effect of OCR errors on the IR performance (e.g. [TAG 06, TAG 01].
One related technique is the use of electronic abstracts to increase the size of domain-specific vocabularies exploited to increase the recognition of OCR engines in digital libraries [LI 06]. Another topic is the caption detection and recognition from videos that is frequently addressed by using OCR packages on the video frames [MAN 06] with applications also in slide retrieval systems [DAD 05].

4.2 Citation analysis

When reliable text extraction techniques are available, then it is possible to automatically perform citation analysis. This task has a long tradition based on manual annotation of bibliographic data and cross-references. Some well-known examples are the Science Citation Index (a commercial system) and the DBLP server that is feed by volunteers [PET 05].
An alternative approach relyes on a Web crawling focused at collecting freely available papers. The downloaded PDF (or PS) files can be processed with DIAR techniques to extract meta-data and references. To this purpose the bibliography recognition can be helpful [BES 04, OKA 04].

4.3 Handwriting recognition

Handwriting recognition in constrained environments (e.g. postal address recognition) is one of the most successful applications of DIAR. However, the processing of historical handwritten documents presents an important challenge for these techniques due to the large size of lexicons involved [GOV 04]. Working applications can be built in constrained situations, for instance considering single writer collections (e.g. [RAT 03]) or using the recognition algorithms for a partial provisional annotation of handwritten documents to be furtherly refined by manual checking [COU 04].

4.4 Layout recognition

Meta-data, "data about data", provide high level information about a set of data. In the field of DLs, the meta-data are usually divided into three main categories: administrative (e.g. the ISBN code), descriptive (e.g. the number of pages of a book), and structural (e.g. the title of a chapter). Structural meta-data can be extracted from a digital book only after an accurate analysis of the book content and layout analysis techniques can be used to obtain this information.
Layout analysis techniques have been widely investigated (e.g [JAI 98a, MAR 05a]) and are now used to process historical documents [LEB 04]. A widely used approach relies on page classification ([DUY 02, APP 01, SHI 01]), using page representations and similarity measures analogous to those adopted in layout-based document image retrieval (e.g. [BEU 06]). One important problem in page classification is the a-priori definition of an exaustive set of classes, that will not change later. To address these issues the techniques discussed in Section 5.4 have some advantages.

4.5 Born-digital documents

As already mentioned, in the last ten years several publisher made available in Internet (usually under restricted access) electronic versions of the published material. Usually, these documents are distributed as PDF files. The presence of already coded text avoids the use of OCR packages and we can assume to deal with error-free text. However, there is still the need for techniques related to the document image analysis research since the PDF documents usually do not explicitly contain some relevant information such as the structural meta-data (e.g. sections heading, figure captions) or the reading order. Lastly, some parts of the documents encoded as images (for instance corresponding to figures in the papers) still need some document image analysis approach. Some applications in this field have been recently proposed [CHE 06] [ESP 06].

5 Retrieval without recognition

In the recognition-free retrieval approaches, the similarity computation between the indexed documents and the query is made at the raw data or at the feature level, avoiding the explicit recognition during the indexing (see the interfaces "DIR" in Fig. 1). This approach has been exploited both for word indexing and for layout-based retrieval. This is a promising direction for poor quality documents, however the scalability and easy integration into existing digital libraries are more problematic with respect to the recognition-based methods.

5.1 Word indexing and keyword spotting

Keyword spotting, whose goal is to locate user defined words from an information flow (e.g. audio streams or sequences of digitized pages, such as faxes) [CUR 95, WIL 00], is one of the first examples of the recognition-free paradigm. In the first approaches the similarity computation took place considering the image or low level features and demonstrated the feasibility of the general idea with low expectations concerning the scalability toward large data-sets. Some recent applications concern the processing of historical documents [TER 05]. Other systems addressed the processing of larger and heterogeneous datasets [TAN 02, MAR 06] or the integration of word image matching at feature level into an existing DL framework (Greenstone) [BAL 06]. The lit-

erature on this domain is very large, and we invite interested readers to refer to [DOE 98, MIT 00, MAR 06].

5.2 Graphical items

The recognition/retrieval of graphical items allows the user to identify interesting documents from a new perspective. Graphical items can be recognized with steps similar to those used in OCR (pre-processing, feature extraction, classification). However, some peculiarities suggest the use of retrieval techniques. The large number of classes to be considered as well as the variable number of classes are problematic for the classifier design and training. Moreover, graphical symbols can have different sizes and are prone to segmentation problems being frequently connected with other parts of the documents.

Logo retrieval has been earlier addressed in [JAI 98b], whereas the retrieval of architectural symbols is described in [TER 03]. In the domain of digital libraries a related problem is the retrieval of graphical drop-caps from historical documents as it is under investigation by a project involving several institutions [PAR 06].

5.3 Handwriting

The design of systems for the retrieval of handwritten documents working at the image or feature level is still at the beginning. The main problems are the large variability in writing style and the large size of the vocabulary. Interesting approaches have dealt with single writer manuscripts. For instance in [RAT 03] the manuscripts of George Washington collection are used as test-bed for the comparison of feature sets for handwritten word retrieval; the similarity is computed by means of one Dynamic Time Warping (DTW) algorithm. Indexing of handwritten documents is approached also in [ZHA 04] where word images are represented with binary features corresponding to gradient, structural, and concavity features. Using the binary features it is possible to speed-up the matching process with respect to the DTW approach without decreasing the retrieval performance. Other applications are the processing of on-line handwritten documents [JAI 03] and the signature-based document retrieval [SRI 06].

5.4 Layout retrieval

Document image retrieval based on layout similarity offers to users a new retrieval strategy that was possible before only by manually browsing documents (either by interacting with physical books-journals or dealing with on-line images on DLs). From the user point of view the retrieval by layout similarity is similar Content Based Image Retrieval (CBIR).

In most cases, a fixed-size feature vector is obtained by computing some features in the regions defined by a grid superimposed to the page [HU 00, TZA 02]. To overcome the problems due to the choice of a fixed grid size, hierarchical representations of the page layout have been considered [MAR 05b, DUY 02].

In the system proposed in [HUA 05] the documents are ranked according to the similarity with respect to a query document selected by the user. Instead of performing a complete document analysis the system extracts text lines and describes the layout by means of relationships between pairs of these lines. A similar approach has been proposed to retrieve similar documents with different resolutions, different formats and multiple languages [LIU 05]. These techniques are appropriate also when performing document retrieval from camera-based devices [NAK 06]. At the crossroad between classification and retrieval are some methods devoted to the slide retrieval in the domain of E-learning [BEH 05]. Another interesting retrieval mechanism is based on the integration of text and layout retrieval that has been proposed, for instance, in [MAR 04], whereas text and graphics in technical manuals are processed in [WOR 01].

5.5 System integration

One important issue for the widespread use of the techniques discussed in this paper is the integration of recognition-free methods into a standard DL framework (text-based or browsing based) that currently only adopts OCR-based techniques. Two main approaches could be exploited. One approach consists in the evolution of the prototypical methods up to the system level by adding functionalities. The alternative approach is based on the close integration of the retrieval techniques into existing DL frameworks. The two approaches have advantages and disadvantages.

In the first case some problems already "solved" should be somehow re-invented and adopted. In particular the interface with existing archiving programs (dealing with OPAC data and specific meta-data) should be provided. In addition, new user interfaces (less familiar for current users) should be provided. However, the main drawback of this approach is the scalability issue, since several methods proposed for DIR only marginally considered the processing cost of effective, but expensive, similarity computations.

The integration into existing DL frameworks has the clear advantage that the transition to the new services is more smooth. However, it is difficult to use sophisticated matching algorithms, even if examples of integration already exist [ASC 05].

6 Performance evaluation

With analogies to information retrieval, the performance evaluation of document image retrieval systems has been mostly addressed by comparing Precision-Recall plots. In some cases, since the answer set is open, or not manually labeled, the top-n precision (the precision measured over an answer set of n items) has been used as well. When the methods deal with classification approaches, then confusion tables and error rates are considered. However, similarly to content based image retrieval, the evaluation of DIR systems can take into account alternatives to better measure the performance in cases where th P-R plots are difficult to obtain or give misleading results (e.g. [MUE 01]).

Another issue concerns the standardization of data-sets that is still far away. Almost each paper deals with different data-sets and this is due also to the variety of problems addressed by different applications. On the other hand, it is clear that digital library can provide to the DIAR community a unique test-bed for the availability of data and meta-data [NAG 06].

7 Conclusions

In this paper, we shortly analyzed the current state of the art of document image retrieval for digital libraries. Several new approaches in this field have been proposed in the last few years. The main limit of our presentation is clearly the low number of pointers to the literature as well as the low description depth of the various papers analyzed.
There are some issues for the future research that emerge from this discussion. The most important challenge is the integration of ad-hoc methods into a standard DL framework. One solution that we envisage is the integration of already developed retrieval engines (e.g. Lucene) with state of the art retrieval mechanism based on the recognition-free paradigm. The latter methods usually take into account symbolic or sub-symbolic data representations that are difficult to incorporate into an existing IR system. However, some efforts should be made in this direction, even at the risk of reduced retrieval effectiveness.

References

[APP 01] APPIANI E., CESARINI F., COLLA A., DILIGENTI M., M.GORI, S.MARINAI, G.SODA, Automatic document classification and indexing in high-volume applications, *International Journal on Document Analysis and Recognition*, vol. 4, n° 1, 2001, pp. 69–83.

[ARM 00] ARMS W. Y., *Digital Libraries*, MIT Press, 2000.

[ASC 05] ASCHENBRENNER S., JSTOR: adapting lucene for new search engine and interface, *DLib Magazine*, vol. 11, 2005.

[BAI 03] BAIRD H. S., Digital Libraries and Document Image Analysis, *Proc. 7th ICDAR*, 2003, pp. 1–13.

[BAL 06] BALASUBRAMANIAN A., MESHESHA M., JAWAHAR C., Retrieval from document image collections, *Proc. DAS*, 2006, pp. 1–12.

[BEH 05] BEHERA A., LALANNE D., INGOLD R., Enhancement of layout-based identification of low-resolution documents using geometric color distribution, *Proc. 8th ICDAR*, 2005, pp. 468–472.

[BES 04] BESAGNI D., BELAID A., Citation recognition for scientific publications in digital libraries, *Proc. DIAL*, 2004, pp. 244–252.

[BEU 06] VON BEUSEKOM J., KEYSERS D., SHAFAIT F., BREUEL T. M., Distance measures for layout-based document image retrieval, *Proc. DIAL*, 2006, pp. 232–242.

[CHE 06] CHEN N., SHATKAY H., BLOSTEIN D., Use of figures in literature mining for biomedical digital libraries, *Proc. DIAL*, 2006, pp. 180–197.

[COU 04] COUASNON B., CAMILLERAPP J., LEPLUMEY I., Making handwritten archives documents accessible to public with a generic system of document image analysis, *Proc. DIAL*, 2004, pp. 270–277.

[CUR 95] CURTIS J. D., CHEN E., Keyword spotting via word shape recognition, *Proceedings of the SPIE - Document Recognition II*, 1995, pp. 270-277.

[DAD 05] DADDAOUA N., ODOBEZ J., VINCIARELLI A., OCR based slide retrieval, *Proc. 8th ICDAR*, 2005, pp. 945–949.

[DOE 98] DOERMANN D., The Indexing and Retrieval of Document Images: A Survey, *Computer Vision and Image Understanding*, vol. 70, n° 3, 1998, pp. 287–298.

[DUY 02] DUYGULU P., ATALAY V., A Hierarchical Representation of Form Documents for Identification and Retrieval, *International Journal on Document Analysis and Recognition*, vol. 5, n° 1, 2002, pp. 17–27.

[ESP 06] ESPOSITO F., FERILLI S., BASILE T., MAURO N. D., Automatic content-based indexing of digital documents through intelligent processing techniques, *Proc. DIAL*, 2006, pp. 204–219.

[GOV 04] GOVINDARAJU V., XUE H., Fast handwriting recognition for indexing historical documents, *Proc. DIAL*, 2004, pp. 314–320.

[HU 00] HU J., KASHI R., WILFONG G., Comparision and Classification of Documents Based on Layout Similarity, *Information Retrieval*, vol. 2, n° 2/3, 2000, pp. 227–243.

[HUA 05] HUANG M., DEMENTHON D., DOERMANN D., GOLEBIOWSKI L., Document ranking by layout relevance, *Proc. 8th ICDAR*, 2005, pp. 362–366.

[JAI 98a] JAIN A. K., YU B., Document representation and its application to page decomposition, vol. 20, n° 3, 1998, pp. 294–308.

[JAI 98b] JAIN A., VAILAYA A., Shape-based retrieval: a case study with trademark image databases, *Pattern Recognition*, vol. 31, n° 9, 1998, pp. 1369–1390.

[JAI 03] JAIN A. K., NAMBOODIRI A. M., Indexing and Retrieval of On-line handwritten documents, *Proc. 7th ICDAR*, 2003, pp. 655–659.

[LEB 04] LEBOURGEOIS F., TRINH E., ALLIER B., EGLIN V., EMPTOZ H., Document images analysis solutions for digital libraries, *Proc. DIAL*, 2004, pp. 1–24.

[LI 06] LI L., TAN C., Improving OCR text categorization accuracy with electronic abstracts, *Proc. DIAL*, 2006, pp. 82–87.

[LIU 05] LIU H., FENG S., ZHA H., LIU X., Document image retrieval based on density distribution feature and key block feature, *Proc. 8th ICDAR*, 2005, pp. 1040–1044.

[LOP 96] LOPRESTI D. P., Robust Retrieval of noisy text, *Proc. of ADL'96*, 1996, pp. 76–85.

[MAN 06] MANOHAR V., SOUNDARARAJAN P., BOONSTRA M., RAJU H., GOLDGOF D., KASTURI R., GAROFOLO J., Performance Evaluation of Text Detection and Tracking in Video, *Proc. DAS*, 2006, pp. 576–587.

[MAR 97] MARUKAWA K., HU T., FUJISAWA H., SHIMA Y., Document retrieval tolerating character recognition errors - Evaluation and application, *Pattern Recognition*, vol. 30, n° 8, 1997, pp. 1361-1371.

[MAR 04] MARINAI S., MARINO E., CESARINI F., SODA G., A general system for the retrieval of document images from digital libraries, *Proc. DIAL*, 2004, pp. 150–173.

[MAR 05a] MARINAI S., GORI M., SODA G., Artificial Neural Networks for Document Analysis and Recognition, *IEEE Transactions on PAMI*, vol. 27, n° 1, 2005, pp. 23–35.

[MAR 05b] MARINAI S., MARINO E., SODA G., Layout based document image retrieval by means of XY tree reduction, *Proc. 8th ICDAR*, 2005, pp. 432-436.

[MAR 06] MARINAI S., MARINO E., SODA G., Font Adaptive Word Indexing of Modern Printed Documents, *IEEE Transactions on PAMI*, vol. 28, n° 8, 2006.

[MIT 00] MITRA M., CHAUDHURI B., Information retrieval from documents: A Survey, *Information Retrieval*, vol. 2, n° 2/3, 2000, pp. 141–163.

[MUE 01] MUELLER H., MUELLER W., SQUIRE S. M., MARCHAND-MAILLET S., PUN T., Performance evaluation in content-based image retrieval: overview and proposals, *Pattern Recognition Letters*, vol. 22, n° 9, 2001, pp. 593–601.

[NAG 06] NAGY G., LOPRESTI D., Interactive document processing and digital libraries, *Proc. DIAL*, 2006, pp. 2–11.

[NAK 06] NAKAI T., KISE K., IWAMURA M., Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval, *Proc. DAS*, 2006, pp. 541–552.

[OKA 04] OKADA T., TAKASU A., ADACHI J., Bibliographic Component Extraction Using Support Vector Machines and Hidden Markov Models, *ECDL 04*, Springer Verlag - LNCS 3232, 2004, pp. 501–512.

[PAR 06] PARETI R., UTTAMA S., SALMON J., OGIER J.-M., TABBONE S., WENDLING L., ADAM S., VINCENT N., On defining signatures for the retrieval and the classification of graphical drop caps, *Proc. DIAL*, 2006, pp. 220-231.

[PET 05] PETRICEK V., COX I. J., HAN H., COUNCILL I. G., GILES C. L., A Comparison of On-Line Computer Science Citation Databases, *ECDL 05*, Springer Verlag - LNCS 3652, 2005, pp. 438–449.

[RAT 03] RATH T., MANMATHA R., Features for word spotting in historical manuscripts, *Proc. 7th ICDAR*, 2003, pp. 218–222.

[SHI 01] SHIN C., DOERMANN D., ROSENFELD A., Classification of document pages using structure-based features, *International Journal on Document Analysis and Recognition*, vol. 4, n° 3, 2001, pp. 232–247.

[SRI 06] SRIHARI S. N., SHETTY S., CHEN S., SRINIVASAN H., HUANG C., ADAM G., FRIEDER O., Document image retrieval using signatures as queries, *Proc. DIAL*, 2006, pp. 198–203.

[TAG 96] TAGHVA K., BORSACK J., CONDIT A., Evaluation of model-based retrieval effectiveness with OCR text, *ACM TOIS*, vol. 14, n° 1, 1996, pp. 64–93.

[TAG 01] TAGHVA K., STOFSKY E., OCRSpell: an interactive spelling correcton system for OCR errors in text, *International Journal on Document Analysis and Recognition*, , n° 3, 2001, pp. 125–137.

[TAG 06] TAGHVA K., BECKLEY R., COOMBS J., The effects of OCR error on the extraction of private information, *Proc. DAS*, 2006, pp. 348–357.

[TAN 02] TAN C. L., HUANG W., YU Z., XU Y., Imaged document text retrieval without OCR, *IEEE Transactions on PAMI*, vol. 24, n° 6, 2002, pp. 838–844.

[TER 03] TERRADES O. R., VALVENY E., Radon transform for linear symbol representation, *Proc. 7th ICDAR*, 2003, pp. 700–704.

[TER 05] TERASAWA K., NAGASAKI T., KAWASHIMA T., Eigenspace method for text retrieval in historical documents, *Proc. 8th ICDAR*, 2005, pp. 437-441.

[TZA 02] TZACHEVA A., EL-SONBATY Y., EL-KWAE E. A., Document Image Matching Using a Maximal Grid Approach, *Proceedings of the SPIE Document Recognition and Retrieval IX*, 2002, pp. 121-128.

[WIL 00] WILLIAMS W., ZALUBAS E., HERO A., Word spotting in bitmapped fax documents, *Information Retrieval*, vol. 2, n° 2/3, 2000, pp. 207–226.

[WOR 01] WORRING M., WIELINGA B., ANJEWIERDEN A., VERSTER F., TODORAN L., KABEL S., Automatic indexing of text and graphics in technical manuals, *ICME 01*, 2001, pp. 241–244.

[ZHA 04] ZHANG B., SRIHARI S., HUANG C., Word image retrieval using binary features, *SPIE, Document Recognition and Retrieval XI*, 2004, pp. 45–53.