Ancient Document Analysis : A Set of New Research Problems

J.-M. Ogier

Ancient Document Analysis : A Set of New Research Problems

Jean-Marc Ogier

[1] Université de la Rochelle,
UFR Sciences et Technologies, Laboratoire L3i, Avenue Michel Crépeau 17042 La Rochelle Cédex

jean-marc.ogier@univ-lr.fr

Summary *: This paper deals with the problem of ancient document analysis. The first part of the paper is dedicated to a kind of state of art concerning french community projects, the purpose of which deals with the preservation and the exploitation of heritage documents. The second part focusses on a set of open issues, which should be tackled by the document analysis community, for the management of the features of these ancient documents.*

Mots-clés *:* ancient document analysis and indexing ; preservation, document analysis, document masses.

1 Introduction

Experts plead for strong actions guaranteeing a lasting preservation of our cultural and scientific resources, which represent a living and collective memory of our societies. The evolution of our economies towards a model based on digital content has a deep impact on this preservation; the challenge is to make this impact a benefit and not a drawback. Large resources have been invested on digitization programs for the cultural heritage, including museum collections, archaelogical sites, audiovisual archives, maps, historical documents, and manuscripts. However, several factors can become a hindrance in optimizing the management of these resources.

First, the approach is often fragmented, with a lack of global management and strategic management tools and no common policy on the management of already digitized resources and on setting priorities;

hence the threat of waste in resources, efforts and investments. Digitization is also costly and

needs huge investments, often based on public funding. Some kind of “return on investment” is expected, at least from the point of view of lasting availability and usability of the digitzed resources.

But the technologies and standards chosen and used today may become quickly obsolete and inadequate.

Intellectual and industrial property rights also lead to various problems. Many partners have

obviously rights and claims on the digitized content, which need to be acknowledged and taken into account. There is a strong need for common solutions for handling these rights in the cultural domain. During the whole acquisition process—from scanning the paper and all the way to indexing the digital documents—many precautions must absolutely be taken to ensure the possibility of using automated techniques. One typical example is the fact that many institutions produce highly compressed files, e.g. using JPEG, which sometimes hinders the use of automated image processing techniques. Thus,institutions which do not consider all the constraints relative to the global “valorization process” produce more or less unusable data, from the point of view of automation. Among the fundamental constraints, let us cite the resolution of the images, that must be at least around 200 or ideally 300 dpi for a long term exploitation strategy. This highlights the necessity of having a close dialogue between different communities, from social and human sciences researchers to computer science specialists.

From the point of view of pattern recognition in general and document image analysis more specifically, we are in the presence of a classical problem involving image processing techniques as well as computing of invariants used for indexing, and database management issues. Compared to classical document image analysis problems, the main changes are due to the amount of data, which raises new research problems. This huge amount of data produces problems with respect to the organization of the feature space in which the documents are transcribed. Another important difference with classical recognition problems is the wide variability of representation of the information that can be found in ancient documents. The fact that the images are often degraded by noise adds to the difficulty. Finally, and this is probably the most important difficulty, the problem of having an exhaustive expression of the future usage of the indexed documents raises the question of how to structure the information and of the cues that have to be extracted from the images.

In this general context, the NAVIDOMASS research project, funded by the French National Research Agency, aims at designing methods for going beyond plain digitization projects for historical documents, as such projects tend to yield large databases of images with very little structure for navigating and indexing these databases. This project issues another french project, called MADONNE, held from 2003 to 2006. These projects investigate the useof document image analysismethodology for providing useful browsing and

indexing features in these large collections. The NAVIDOMASS project is thus focused on indexing, organization and incremental enrichment of heritage data, in order to provide general services to the users, including researchers in human and social sciences and to work towards interoperability of data and browsing tools.
The NAVIDOMASS project started at the begining of 2007 and will end at the end of 2009. In addition to our groups at the L3i laboratory at University of La Rochelle, the partners are from LORIA at Nancy, from the PSI lab at University of Rouen, the LI lab at University of Tours, the the CRIP5 lab at University Paris 5, and IRISA in Rennes. Most of these research partners have a strong and complementary background in document image analysis, thus allowing us to build a large set of generic services. This paper presents an overview of the main results obtained by the partners of the MADONNE/NAVIDOMASS projects, and especially by B. Couasnon, V. Eglin, L. Heutte, N. Journet, T.Paquet, R. Pareti, J.-Y. Ramel, J.-P. Salmon, S. Tabbone, S. Uttama, N. Vincent and L. Wendling.

2 Ancient Document specific Research Topics

To reach the objectives of providing structured access and browsing capabilities to large sets of cultural heritage documents, we need to index these sets using the various features which can be of interest for searching. This includes illustrations, text, styles, various kinds of symbols, handwritten annotations, etc. This leads us to the need for close cooperation between various document analysis expertise areas, as none of these areas answers the requirements on its own. In the following, we will detail some of the research themes we addressed in the MADONNE/NAVIDOMASS projects.

2.1 Collection modelling

In the context of large collections of data, one can observe a strong homogeneity in the way the information is structured, depending on the different collections. Collections modeling consists in extracting as automatically as possible the features that characterize a collection or a set of collections, in order to assist the analysis of the images, by applying adapted image processing tools. This question raises the problem of automatically discovering the similarities concerning the structures of the books, in order to construct a relevant model of the corresponding collection. In the context of the MADONNE project, Journet et al. [Jou 05] proposed a set of processes allowing to categorize the pages of a book according to the spatial organization of the data. The extraction of some features describing the layout and the structure of the document allowed them to structure the collections of books, in term of similarities between the spatial organization of their contents. For that purpose, Journet proposed a function based on autocorrelation for the extraction of the features (Fig. 1). The future of this work will consist in measuring the similarities between different books, providing the required models for the collection.

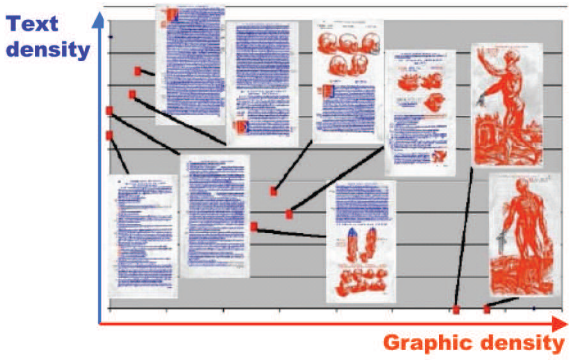

FIG. 1 Image categorization as a function of he content

2.2 Document Layout Analysis

The structure of a document is usually relative to a presentation and organization model and aims at helping the user in understanding the information provided by the document. Specific challenges appear in old collections, as the typical documents from the 15th, 16th or 17th century dealt with in our project. In a number of cases, the layout itself conveys precious information for browsing the documents.
The analysis of the elements of the layout may be an excellent guide for content based information retrieval, by using full text search of similarities measurements, applied to specific zones yielded by the layout analysis. It may also guide us in finding the information which can be made readily available to the general public, as opposed to information which is protected by privacy, confidentiality or property rules. In the context of the Madonne project, let us cite the work of Couasnon [Cou 03] on the collective annotation process of military registers from the 19th century (Fig. 2). This process goes through a very reliable analysis of the structure of the documents, based on 2D grammar techniques integrated in the DMOS system, allowing to detect each cell of the military register even if the structure of the document is degraded. Thanks to this fine detection of the cells, the system proposes a similarity measurement system allowing to browse military registers on handwritten names with textual queries without OCR. The similarity measure is based on the extraction of low level primitives, graphems, the organization of which permits to provide a measure of similarity between two handwritten models. The difficult points encountered here are relative to the overlapping of graphic layers, for which text–graphic segmentation techniques may be useful. This system has been validated on 165,000 pages. Another contribution has been the building of a system called Agora for the interactive analysis of document layout [Ram 05]. Depending on the needs (extraction of ornamental letters, of marginal notes, of titles. . .), the user can thus build scenarios allowing to label, to merge or to remove the extracted blocks. The scenarios can be stored, modified and applied to other sets of images in batch processing mode.

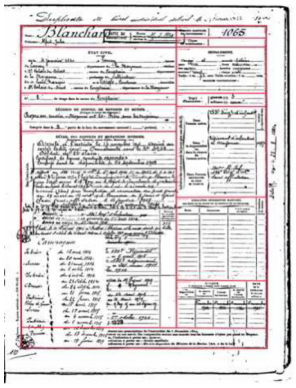

FIG. 2 Structure analysis with the DMos System

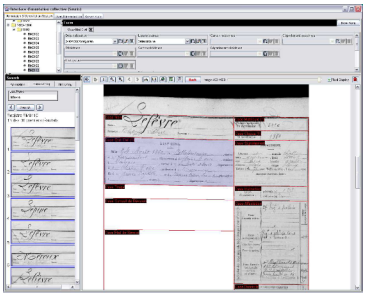

FIG. 3 Structure analysis with the DMos System [Cou 03]

2.3 Handwritten Documents

The processing of handwritten manuscripts from the cultural heritage leads to specific questions which are far away from usual handwriting recognition analysis as addressed in postal or banking applications, for instance. The aim is rarely to recognize the handwriting but rather to characterize and identify different writers [Ben 05], or to date some documents. Fig. 3 is a typical example of what we aim at working on. Indexing based on visual information features is therefore one of the main keywords for us. In specific cases (handwritten name registers for instance) global shape recognition techniques can lead to classification according to shape similarities and even to limited handwriting recognition for indexing purposes [Cou 03]. For this, lexical knowledge about the domain of use can be of considerable help.

In the context of the Madonne project, Paquet and al. have proposed a set of processes allowing to help historians to analyze Flaubert's manuscripts layouts [Nic 04]. In this context, some relevant signatures are computed in order to check that the spatial organization of the data match features characterizing Flaubert's handwriting style. In this case, Hidden Markov Models, as well as dynamic programming, are used for the segmentation and modeling process. The results highlight that the relevant features that have to be considered in such a process combine handwriting features and structural information about the spatial organization of the data. Such an analysis leads to characterization of the author's (authentication), but also to the possibility of "reconstructing" the genesis of the writing process through the successive annotations.

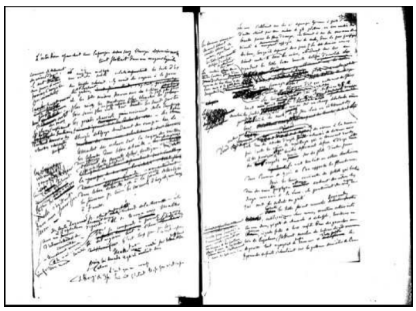

FIG. 4 An example of handwritten manuscript with authors annotations

2.4 Indexing on Graphical Features

Usually, documents are indexed mainly on text. However, heterogeneous sets of historical documents often contain features which are graphical in nature, although they represent text. This is especially the case with illustrated dropcaps associated with artwork (Fig. 5), on which we have focused a lot of work in the MADONNE/NAVIDOMASS projects [Par 06]. There is little knowhow on how to compute invariants for indexing documents on this kind of features. Actually, our experience with our CESR partners highlights the diversity of requirements that may be expressed by the users. Some historians want to detect slight differences between dropcaps in order to be able to date them, while some others are only interested in global content based retrieval problems (find similar dropcaps, see Fig. 8).

FIG. 5 An example of illustrated Dropcap

A first problem to be addressed with these features is that of the document image segmentation problem.

The images to be processed are noisy and LORIA's group has used an adaptive image-smoothing filter which is more robust regarding different noise levels than existing methods (Fig. 6).

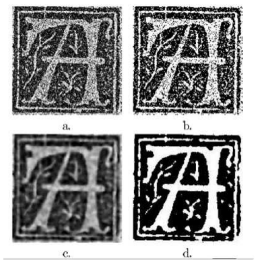

FIG. 6 Noise Filtering (LORIA's contribution)

Complementary sets of descriptors have been developed in the CRIP5 Laboratory and in the L3i laboratory.
The first is based on a statistical modeling of the distribution of the pixels within a dropcap, using the Zipf law [Par 05]. This allows us to classify the dropcaps as a function of their style, thanks to the analysis of specific ruptures according to the Zipf law (Fig. 7). The second is based on top down segmentation process allowing to provide a set of layers on each of which a signature is computed [Utt 05], in order to characterize the spatial organization of the data. This process allows us to implement a content based image retrieval system, the results of which are very encouraging in terms of recall/precision (Fig. 8).

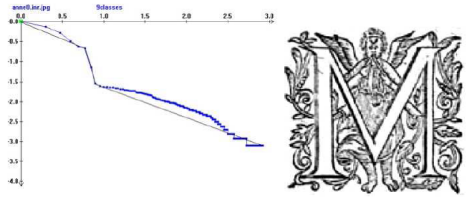

FIG. 7 Zipf Law of a specific DropCap [Par 05]

LORIA's group also defined a new method for combining shape descriptors based on a behavior study of a learning set [Sal 06]. Each descriptor is computed on several clusters of objects or symbols. For each cluster and for any descriptor, an appropriate mapping is directly carried out from the learning database.

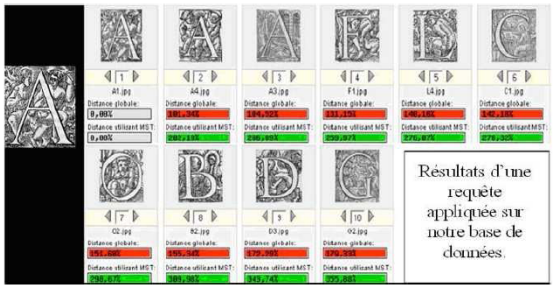

FIG. 8 Drop Cap indexing [Utt 05]

Then, existing conflicts are assessed and integrated into a map. Such a combination of descriptors improves the recognition rates and the ranking obtained on dropcaps like those in Fig. 9.

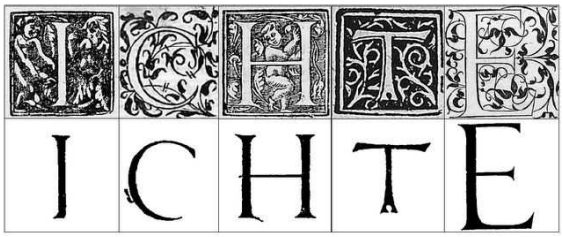

FIG. 9 Letters extracted from dropcaps and used in retrieval and ranking applications [Sal 06]

2.5 Similarity retrieval for document compression

The digitization of cultural documents also leads to difficulties in terms of storage and transmission on a bandwidth-limited network. Only lossy compression with an acceptable perceptible loss of in formation may reduce the weights of images. The existing compression formats like JPEG, DJVU or DEBORA are unfortunately not effective on handwritten documents images, due to the great complexity of handwritten shapes and to the difficulty of localizing them precisely (see Fig. 4). Within the Madonne project, a handwritten text compression methods has been proposed [Ela 05], consisting in separating the text and the background by similarity retrieval. Basically, the localization of redundancies is based on a decomposition of the handwriting text into oriented segments with invariant contours points and can be extended to any part of the image which presents distributed similarities.

3 Achievements and open problems

As one can see through these different points, the preservation of cultural heritage documents requires to combine various methods from the document image analysis field : image processing, handwriting recognition, document layout analysis, graphics recognition, etc.

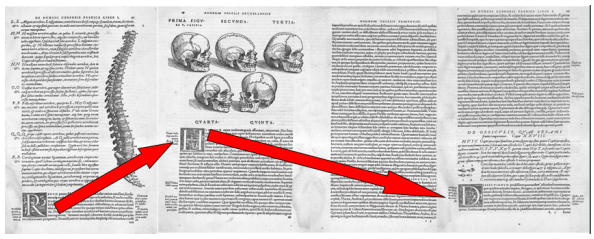

FIG. 10 Hyper text navigation [Jou 05]

However, many problems remain open, and require more works. Of course, this huge amount of data raises new problems, specifically in relation with "content based" operations. For graphics in old manuscripts, for instance, some new signatures have to be developed in order to design word spotting or graphic spotting methods (Fig. 10), similar to what can be achieved through hyper-text navigation.
These problems may be organized as follows :

1. Content characterization : the different examples illustrated in the previous parts highlight the necessity to work on the definition of relevant signatures for the indexing process. Depending on the kind of information to characterize, and depending of the user requirements, some of them can deal with structure, handwritten, or graphic indexing
   (a) . In the context of manuscript documents, as one can see in the part dealing with handwritten document, many aspects

should be considered for the authentification/transcription of a document. Some signatures integrating texture features and/or spatial based characteristics should be considered. These signatures should integrate statistical descriptors, combined in the context of structural signatures. Concerning indexing services, some new researches should be considered about word spotting techniques, in order to provide relevant shape based words descriptors, allowing to retrieve a particular word, without running any OCR system. Some interesting issues can be found in the works of (Ley 04)

(b) In the context of structure based indexing, some researches should be considered, in order to define spatial based signature for structure characterization. These signatures should be based on preliminary segmentation stages, in order to distinguish all the layers of information of the document : printed, manuscript, graphic, …This point highlights the necessity to work on image segmentation techniques, in the context of historical documents. Computer vision based techniques should be re-visited in order to analyse the approaches that may be adapted to the specific context of documents images. Some structural signatures, combining statistical descriptors should also be considered. Some interesting works issuing from [Qur 07] should be mentionned as interesting approaches for this kind of problem.

(c) In the context of graphic images, some new content based characterization techniques should be developped. Indeed, in the context of ancient documents, most of the time, one has some to consider illustations that had been printed thanks to wood stamps, ant the resulting images are generally images of strokes. As a consequence, these images have so specific features that the reuse of « classical » images processing techniques is not so obvious. Different approaches are possible for characterizing such images like segmentation based techniques (Utt 05), Zipf law based techniques for content characterisation (Par 06), keypoints detection characterization or Wavelets decomposition. Whatever classical technique is considered, this one should be re-analysed in order to see how it has to be re-configured for being adapted to ancient document images.

2. Scale resistance : this concerns mainly the problem of scaling the recognition approaches, because of the variability of representation that is one of the specific features of ancient documents. This aspect appears to be one of the most fundamental aspects since it is based on the assumption that the number of objects in the learning database is very small. On this basis, the problem is here to characterize which features are relevant for the user requirements, and which are generic enough for representing the class of object to be indexed or recognized, from a generalization point of view. Some interesting approaches dealing with this problem can be found in [Sal 06].

3. Masses management : most of the time, each object to be retrieved is « summarized » through a signature, that can be based either on a statistical or a structural description. When dealing with huge amount of document images, the problem is thus to be able te retrieve an object or a part of an object. In the context of a huge repository of images, an exhaustive and sequential comparison of the query with all the objects of the database is not reasonnable, because of obvious complexity problems. As a consequence, in order to avoid such sequential research, the problem of structuring the features space becomes a crucial problem. Depending on the nature of the signature (statistical vs structural), this problem can be tackled by using different strategies. Considering statistical approaches, the indexing and clustering techniques in the context of high dimensional features vector should be explored, in order to structure the features spaces within a hierarchical manner, so that the access could be naturally indexed. Some interesting issues can be found in [Zuw 06] In the context of structural description, the problem is often to structure graphs spaces, so that it is not necessary to run an exhaustive comparison between the query and all the graphs summarizing the images. Some interesting issues dealing with median graphs descriptions, spectral graphs and graphs probing approaches are interesting from this point. [Bun 00] (See figure 11)

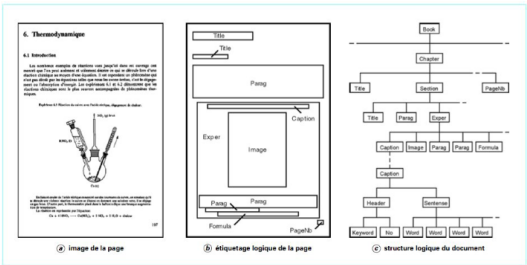

FIG. 11 Graph analysis for document image description

4. Use interaction and Knowledge modelling : Another topic is the problem of modelling the domain knowledge, in order to assist the user for producing a relevant scenario when dealing with a specific subject. Most of the time, the lack of user

requirements is a reccurrent problem. Considering this point, three aspects seem important to develop. First, the necessity to provide human interfaces allowing the user to interactively construct image analysis scenarii, according to their objectives. This kind of strategy can be implemented by proposing simple and editable scenarii that can be easily experimented by the user. Some interesting issues can be found in [Ram 06], with AGORA system. Second, in the context of information retrieval, the problem is to provide user-friendly interfaces allowing the user to interactively modify the results provided by an automatic system, and such that the system is able to « learn » the requirements of the user. This aspect raises fundamental problems related to relevance feedback and incremental learning, which are probabily the most difficult points (and one of the less significantly explored) of this research problems. Third, the problem of formal modelling of the user knowledge appears to be a fundamental aspect of ancient document indexing. Indeed, domain experts are generally able to express the criteria on which their requirements are based on. On this basis, the ontology based models should be developped in order to provide generic system, allowing to dynamically generate analysis scenarii. This research aspect, in interaction with relevance feeedback questions is a difficult point that has to be considered in the future. Some interestig issues can be found in [Dom 02]

4 Conclusion

This paper presents a short synthesis of different problems that have been tackled in the context of reserach projects funded by the French government : MADONNE and NAVIDOMASS. Some of the contributions of this research consortium have been presented and resarch issues have also been suggested. These orientations highlight this interest of working with other communities, since many of these suggestions concern new human interfaces, ontology modelling, graph clustering, ….

5 References

[Ela 05] A. El Abed, V. Eglin, and F. Lebourgeois. Frequencies decomposition and partial similarities retrieval for patrimonial handwriting documents compression. In Proceedings of 8th International Conference on Document Analysis and Recognition, Seoul (Korea), pages 996–1000, 2005.

[Ben 05] A. Bensefia, T. Paquet, and L. Heutte. A writer identification and verification system. Pattern
Recognition Letters, 26(13):2080–2092, October 2005.

[Cou 03] B. Couasnon and I. Leplumey. A Generic Recognition System for Making Archives Documents Accessible to Public. In Proceedings of 7th International Conference on Document Analysis and Recognition, Edinburgh (Scotland, UK), pages 228–232, August 2003.

[Jou 05] N. Journet, V. Eglin, J.-Y. Ramel, and R.Mullot. Text/Graphic Labeling of Ancient Printed Documents. In Proceedings of 8th International Conference on Document Analysis and Recognition, Seoul (Korea), pages 1010–1014, September 2005.

[Nic 04] S. Nicolas, T. Paquet, and L. Heutte. Enriching Historical Manuscripts: The Bovary Project. In
Proceedings of the 6th IAPR International Workshop on Document Analysis Systems, Florence, (Italy), volume 3163 of Lecture Notes in Computer Science, pages 135–146, September 2004.

[Par 06] R. Pareti, S. Uttama, J.-P. Salmon, J.-M Ogier, S. Tabbone, L. Wendling, and N. Vincent. On defining signatures for the retrieval and the classification of graphical dropcaps. In Accepted for presentation at 2nd IEEE International Conference on Document Image Analysis for Libraries, Lyon, France, April 2006.

[Par 05] R. Pareti and N. Vincent. Global Discrimination of Graphics Styles. In Proceedings of 6th IAPR International Workshop on Graphics Recognition, Hong Kong, August 2005.

[Ram 05] J.-Y. Ramel and S. Leriche. Segmentation en analyse interactives de documents anciens imprimés. Traitement du Signal, 22(3):209–222, November 2005.

[Sal 06] J.-P. Salmon, L.Wendling, and S. Tabbone. Improving the Recognition by Integrating the Combination of Descriptors. International Journal on Document Analysis and Recognition, 2006.
Accepted for publication.

[Utt 05] S. Uttama, J.-M. Ogier, and P. Loonis. Top-down segmentation of ancient graphical drop caps: lettrines. In Proceedings of 6th IAPR International Workshop on Graphics Recognition, Hong Kong, pages 87–96, August 2005.

[Ley 04] Leydier Y., Lebourgeois F., Emptoz H., Serialized Unsupervised Classifier for Adaptative Color Image Segmentation: Application to Digitized Ancient Manuscripts, *ICPR*, Cambridge, UK, 23-26 Aug 04, pp. 494-497.

[Qur 07] Qureshi, R.J., Ramel, J-Y., Cardot, H. 2007. Symbol Spotting in Graphical Documents Using Graph Representations, *Seventh IAPR International Workshop on Graphics Recognition - GREC 2007.* Curitiba, Brazil. September 20-21, 2007.

[Zuw 06] Zuwala, D. et Tabbone, S. 2006. Une Méthode de Localisation et de Reconnaissance de Symboles sans Connaissance a Priori. Dans le *Colloque International Francophone Sur l'Ecrit Et Le Document - CIFED'06,* pp. 127-131.

[Ram 06] Ramel, J., Busson, S., Demonet, M.: Agora: the interactive document image analysis tool of the BVH project. In: Conf. on Document Image Analysis for Library (2006) 145–155

[Dom 02] Dombre J., Richard N., Fernandez-Maloigne C, Use of spatial information for content-based image retrieval, Annales des Télécommunications - 2002

[Dom 00] Bunke, H. 2000. Recent Development in Graph Matching, *IEEE 15th International Conference on Pattern Recognition,* **2**, pp. 117 - 124.