Content-Based Old Documents Indexing

Mickaël Coustaty, N. Sidère, Jean-Marc Ogier, Pierre Héroux, Jean-Yves Ramel, Chouaib Hassan, Nicole Vincent, Salim Jouili, Salvatore Tabbone

Content-Based Old Documents Indexing

Mickael Coustaty[1], Nicolas Sidere[3], Jean-Marc Ogier[1], Pierre Heroux[2], and Jean-Yves Ramel[3]

[1]L3i Laboratory, Avenue Michel Crepeau, 17042 La Rochelle, France
{mcoustat, jmogier}@univ-lr.fr
[2]LITIS Laboratory, Avenue de l'Universit, 076 800 Saint-Etienne-du-Rouvray - France
Pierre.Heroux@univ-rouen.fr
[3]LI Laboratory, 64 avenue Jean Portalis, 37200 Tours, France
{ramel, nicolas.sidere}@univ-tours.fr

Abstract. A huge amount of printed documents were published and distributed during the 15th century. In order to protect this inheritance, a digitalization campaign is held on these documents. The mass of documents generated by digitization create a problem to retrieve and to index them. This paper presents a french collaborative project that brings together seven laboratories around ancient documents indexing by content. This project is composed of two steps: 1) Extraction of information from images and 2) Retrieving documents with similar semantic interpretation for user. Two actors of this project present in this paper how they combine their works and their goal.

Key words: Old documents, CBIR, Indexing, Graph

1 Introduction

With the improvement of printing technology since the 15th century, there are a huge amount of printed documents published and distributed. The printed book quickly becomes a regular object in the world. By 1501 there were 1000 printing shops in Europe, which had produced 35,000 titles and 20 million copies. Since that time, a vast amount of books have been falling into decay and degrading. This means not only the books themselves are disappearing, but also the knowledge of our ancestors. Therefore, there are a lot of attempts to keep, organize and restore ancient printed documents. With the better digital technology, one of the preservation methods of these old documents is the digitization. However, digitized documents will be less beneficial without the ability to retrieve and extract the information from them which could be done by using techniques of document analysis and recognition.

2 NaviDoMass and Ancient Document Indexing challenge

NAVIDOMASS (Navigation into Documents Masses) is a french collaborative project. With the collaboration of seven laboratories in France, the global objective of this project is to build a framework to derive benefit from historical documents. It aims to preserve and provide public accessibility to this national heritage. NAVIDOMASS is a three-year project expected to complete in 2010. It is established on four principles: anywhere (global access), anyone (public and multilingual), anytime and any media (accessible through various channels such as world wide web, smartphone, etc.). The focus of NAVIDOMASS is on five studies: (1) user requirement, participative design and ground truthing, (2) document layout analysis and structure based indexing, (3) information spotting, (4) structuring the feature space [HSO$^+$08,JT08] and (5) interactive extraction and relevance feedback. As a part of NAVIDOMASS project, this paper focuses on the graphics part : graphics indexing and CBIR. However, the main interest of this study is based on specific graphics called drop caps and is inspired by [PV06] and [ULDO05]. Even if the use case of this paper is reduced to drop caps indexing (also called dropcaps or ornemental letters), the aim of this study is to develop complex graphic indexing techniques. By complex graphic, we mean here any graphic document which does not respect a rigourous structure, as one can find in the domain of symbol recognition for instance.

3 Old images difficulties

Old documents images are particularly difficult images. First of all, they are only composed of lines due to stamps that were applied on paper. Next, the second difficulty rely on the degradation of paper and documents. Finally, many of images contained in old documents are composed of multiplicity of informations and their content is a mixed of simple elements.

4 Information Extraction

Due to the difficulties above-cited, traditional algorithms to extract region of interest are not efficient. Thus, we chose to develop a new region of interest method to extract information. This new method, which relies on a Aujol and Chambol decomposition ([AAFC05]), simplifiy images to disjoin simple elements. Then a Zipf Law ([PV06]) is applied to separate foreground from background. These two decomposition permit to obtain connected components and we finally select some of them using their topologic attributes (area, eccentricity, location, ...). Each connected component extracted with its description is associated to a node. All the node are linked to build a graph of each image. Finally, this graph is used to structure the description space and to retrieve similar images.

5 Indexation

Indexation of structured data is highly correlated with the use of graphs. However, the indexing of structural data makes an intensive use of graph distance which is a problem known for its high computational cost. We choose to develop a new graph embedding method, with the encapsulation of the topology of the graph and the associated labels in a numerical matrix, which reduce the distance complexity to linear. Our approach relies on the non-isomorphic graphs network which includes every graphs with maximum N edges. These graphs, called patterns, builds a lexicon. The topology of the graph is embedded by the count of the occurences of the patterns from the lexicon. The attributes are listed and also counted to take care of the label information. The table 1 shows an example of our matricial representation of the graph in figure 1. This representation allows the use of classical distance computation, ie euclidean distance, to measure the dissimilarity between two matrix, ie two graphs.

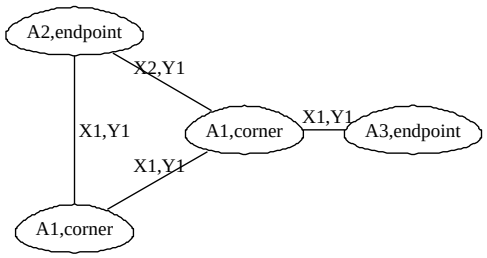

Fig. 1. A simple labelled graph

Pattern						
Freq.	4	4	4	2	1	1
A1, corner	2	4	7	4	2	2
A2, corner	0	0	0	0	0	0
A3, corner	0	0	0	0	0	0
A1, endpoint	0	0	0	0	0	0
A2, endpoint	1	2	3	2	1	1
A3, endpoint	1	1	2	2	0	1
X1, Y1	0	3	6	5	2	2
X1, Y2	0	0	0	0	0	0
X2, Y1	0	1	2	1	1	1
X2, Y2	0	0	0	0	0	0

Table 1. Co-occurence matrix of the graph in fig.1

6 Conclusion

This paper presents a french collaborative project which proposes a full indexing system to navigate into a large database of old documents by contents. The first step extract region of interest specifically to

old documents images attributes. Once these region are extracted, they are fully describe by selected features and indexed using a low computational cost graph matching algorithm. This one relies on a new graph embedding method which encapsulate the topology of graphs and the associated labels in a numerical matrix. This work is actually in development and results are expected for the end of this project. In perspective of this project, an interactive process that include user will be efficient to guide extraction of information and indexing process. The mass of information contained by old documents make retrieving system difficul and results expected depends on user mainly. A relevance feedback permit to give a significant information to adjust different process parameters.

References

[AAFC05] J. F. Aujol, G. Aubert, L. Blanc Feraud, and A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision*, 22(1):71–88, January 2005.

[HSO+08] H.Chouaib, S.Tabbone, O.Ramos, F. Cloppet, and N.Vincent. Feature selection combining genetic algorithm and adaboost classifiers. In *ICPR'08*, Florida, 2008.

[JT08] Salim Jouili and Salvatore Tabbone. Applications des graphes en traitement d'images. In *ROGICS'08*, pages 434–442, Mahdia Tunisia, 2008. University of Ottawa, Canada and University of Sfax, Tunisia.

[PV06] Rudolf Pareti and Nicole Vincent. Ancient initial letters indexing. In *ICPR '06*, pages 756–759, Washington, DC, USA, 2006. IEEE Computer Society.

[ULDO05] Surapong Uttama, Pierre Loonis, Mathieu Delalandre, and Jean-Marc Ogier. Segmentation and retrieval of ancient graphic documents. In *GREC*, pages 88–98, 2005.