PhD of Science of the University of La Rochelle
Defended by
Romain Raveaux
Graph Mining and Graph Classification : Application to cadastral map analysis.

August 25, 2010

This thesis tackles the problem of technical document interpretation applied to ancient and colored cadastral maps. This subject is on the crossroad of different fields like signal or image processing, pattern recognition, artificial intelligence, man-machine interaction and knowledge engineering. Indeed, each of these different fields can contribute to build a reliable and efficient document interpretation device. This thesis points out the necessities and importance of dedicated services oriented to historical documents and a related project named ALPAGE. Subsequently, the main focus of this work: Content-Based Map Retrieval within an ancient collection of color cadastral maps is introduced. The organization of this thesis paper is in five chapters. The interaction between chapters is illustrated in figure 1 and a short description of each chapter is put forward as follows:

1 Introduction

Chapter 1 gives the introduction to the project and provides overall concept of this thesis. We introduce a general aspect of document image analysis, the necessities and importance of historical documents and the related project named ALPAGE. Next, we focus on coloured cadastral maps and define the scope and objectives of this study.

2 Color Map Understanding: State of the art

In the present chapter, we discuss how to bring an automation of the single modules of a Raster to Vector conversion system to fullest possible extent. GIS can be categorized in two types, analytical and register GIS. Analytical GIS do not require an extremely high level of geometric exactness in the cartographic materials, whereas they do require fast processing of a large number of vector layers. An example of analytical GIS is GIS developed to solve territorial planning problems, while an example of a register GIS is GIS developed for a

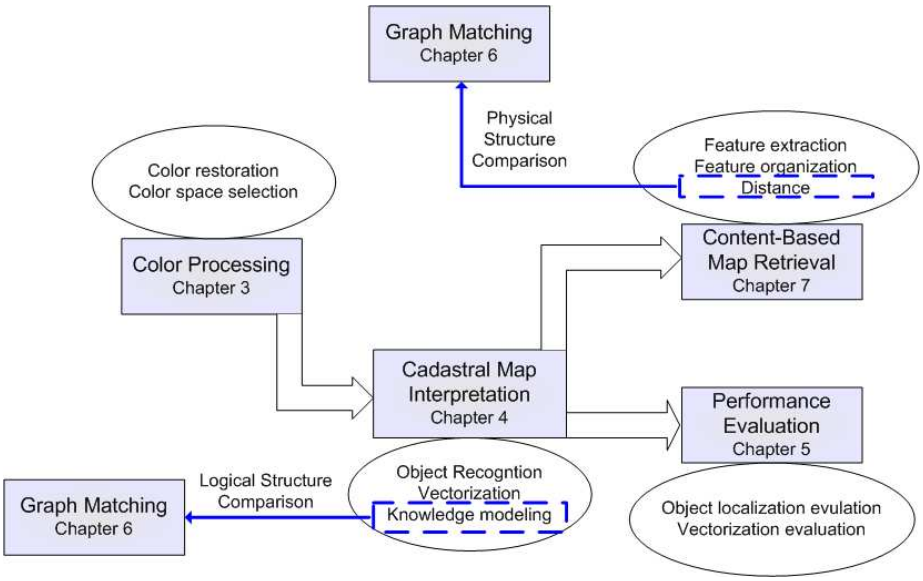

Figure 1: Thesis organization

cadastral system. Mainly, we focus on the case of register GIS with the aim is to describe and highlight the global concepts and crucial points of our problem.

3 Color Processing

The choice of a relevant color space is a crucial step when dealing with image processing tasks (segmentation, graphic recognition...). From this fact, we address in a generic way the following question: What is the best representation space for a computational task on a given image? In this chapter, a color space selection system is proposed. From a RGB image, each pixel is projected into a vector composed of 25 color primaries. This vector is then reduced to a Hybrid Color Space made up of the three most significant color primaries. Only three color components are retained to be conformed with standard image formats. Hence, the paradigm is based on two principles, feature selection methods and the assessment of a representation model. The quality of a color space is evaluated according to its capability to make color homogeneous and consequently to increase the data separability. Our framework brings an answer about the choice of a meaningful representation space dedicated to image processing applications which rely on color information. Standard color spaces are not well designed to process specific images (ie. Medical images, image of documents) so a real need has come up for a dedicated color model.

4 Cadastral map interpretation

In this chapter, an object extraction method from ancient color maps is proposed. It consists of the localization of frame, text, quarters and parcels inside a given cadastral map. Firstly, a model of cadastral map is introduced; this

knowledge representation was elaborated in collaboration with historians and architects, experts in this domain. Secondly, the color aspect is inherited from the color restoration algorithm and the selection of a relevant hybrid color space presented in chapter 3. Thereafter, dedicated image processing aim at locating the various kinds of objects laid out in the raster. These especially designed detectors can retrieve different components such as characters, streets, frame, quarters and parcels. These specific tools are run successively in the objective to identify boundaries of the different elements. In a last phase, these elements are put into a graph-based representation to be further compared with the meta-model defined by the experts. This comparison is carried out thanks to a graph matching algorithm. "Metamodeling" is the construction of a collection of "concepts" (things, terms, etc.) within a certain domain. A model is an abstraction of phenomena in the real world; a metamodel is yet another abstraction, highlighting properties of the model itself.

5 Evaluation of Cadastral Map processing

This chapter presents a benchmark for evaluating the Raster to Vector conversion systems. The benchmark is designed for evaluating the performance of graphics recognition systems on images that contain straight lines (segments) and polygons (solid) within the images. Our contribution is two-fold, an object mapping algorithm to spatially locate errors within the drawing, and then a cycle graph matching distance that indicates the accuracy of the polygonal approximation. The performance incorporates many aspects and factors based on uniform units while the method remains not rigid (threshold-less). This benchmark gives a scientific comparison at two levels of coherency and uses practical performance evaluation methods that can be applied to complete vectorization systems. It is also the opportunity to compare our unsupervised evaluation method defined in chapter 4 with a ground-truth based one. Our system dedicated to cadastral map vectorization was evaluated under this benchmark and its performance results are presented in this chapter. We hope that this benchmark will help assess the state of the art in graphics recognition and highlight the strengths and weaknesses of current vectorization technology and evaluation methods.

6 A Graph Matching method and a Graph Machting Distance based on probe assignments

During the last decade, the use of graph-based object representation has drastically increased. As a matter of fact, object representation by means of graphs has a number of advantages over feature vectors. As a consequence, methods to compare graphs have become of first interest. In this chapter, a graph matching method and a distance between attributed graphs are defined. Both approaches are based on subgraphs. In this context, subgraphs can be seen as structural features extracted from a given graph, their nature enables them to represent local information of a root node. Given two graphs G_1,G_2, the univalent mapping can be expressed as the minimum-weight subgraph matching between G_1 and G_2 with respect to a cost function. This metric between subgraphs is directly

derived from well-known graph distances. In experiments on four different data sets, the distance induced by our graph matching was applied to measure the accuracy of the graph matching. Finally, we demonstrate a substantial speed-up compared to conventional methods while keeping a relevant precision.

7 Content-Based Map Retrieval

Traditionally when facing a warehouse of natural scenes to be queried by examples; Conventional methods would jut look at the system level comparing the query images to all the images within the corpus. By system level, we mean the pixel image in its self sufficient way, pixels or a gathering of pixels. When talking about images of documents, the scenario is fairly different because we are dealing with images created by humans and dedicated to humans. This makes a huge difference and allows comparisons and an exploration at higher levels. Two more stages can be drawn : (a) Image of documents can be meaningfully vectorized, and the collection can be addressed thinking at the vector level. (b) Document images have a strong semantic and a navigation using a model representation has come true.

8 Our publications

2010

[1]

2008

[2] [3] [4] [5]

2007

[6] [7]

2006

[8] [9] [10]

2005

[11]

References

[1] Romain Raveaux, Jean-Christophe Burie, and Jean-Marc Ogier. A graph matching method and a graph matching distance based on subgraph assignments. *Pattern Recognition Letters*, 31(5):394–406, 2010.

[2] Romain Raveaux, Jean-Christophe Burie, and Jean-Marc Ogier. A Colour Space Selection Scheme dedicated to Information Retrieval Tasks. In *Pattern Recognition in Information Systems, Proceedings of the 8th International Workshop on Pattern Recognition in Information Systems, PRIS 2008, In conjunction with ICEIS 2008, Barcelona, Spain, June 2008*, pages 123–134. INSTICC PRESS, 2008.

[3] Romain Raveaux, Jean-Christophe Burie, and Jean-Marc Ogier. Object Extraction from Colour Cadastral Maps. In *DAS '08: Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems*, pages 506—-514, Washington, DC, USA, 2008. IEEE Computer Society.

[4] Romain Raveaux, Jean-Christophe Burie, and Jean-Marc Ogier. *A Segmentation Scheme Based on a Multi-graph Representation: Application to Colour Cadastral Maps*, volume 5046 of *Lecture Notes in Computer Science*, pages 202–212. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[5] Romain Raveaux, Jean-Christophe Burie, and Jean-Marc Ogier. A colour text/graphics separation based on a graph representation. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1–4, Washington, DC, USA, 2008. IEEE Computer Society.

[6] R. Raveaux, J.-C. Burie, and J.-M. Ogier. A Colour Document Interpretation: Application to Ancient Cadastral Maps. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 1128—-1132, Washington, DC, USA, 2007. IEEE Computer Society.

[7] Romain Raveaux, Barbu Eugen, Hervé Locteau, Sébastien Adam, Pierre Héroux, and Eric Trupin. *A Graph Classification Approach Using a Multi-objective Genetic Algorithm Application to Symbol Recognition*, volume 4538 of *Lecture Notes in Computer Science*, pages 361–370. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[8] Hervé Locteau, Romain Raveaux, Sébastien Adam, Yves Lecourtier, Pierre Héroux, and Éric Trupin. Approximation of Digital Curves using a Multi-Objective Genetic Algorithm. In *18th International Conference on Pattern Recognition (ICPR)*, pages 716–719, Washington, DC, USA, 2006. IEEE Computer Society.

[9] Herve Locteau, Romain Raveaux, Sebastien Adam, Yves Lecourtier, Pierre Heroux, and Eric Trupin. *Polygonal Approximation of Digital Curves Using a Multi-objective Genetic Algorithm*, volume 3926 of *Lecture Notes in Computer Science*, pages 300–311. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[10] Eugen Barbu, Romain Raveaux, Herve Locteau, Sebastien Adam, Pierre Heroux, and Eric Trupin. Graph Classification Using Genetic Algorithm and Graph Probing Application to Symbol Recognition. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 296—-299, Washington, DC, USA, 2006. IEEE Computer Society.

[11] Hervé Locteau, Romain Raveaux, Sébastien Adam, Yves Lecourtier, Pierre Héroux, and Éric Trupin. Polygonal Approximation of Digital Curves Using a Multi-objective Genetic Algorithm. In Liu Wenyin and Josep Llados, editors, *Graphics Recognition. Ten Years Review and Future Perspectives, 6th Internation Workshop, GREC 2005, Hong Kong, China, August 25-26, 2005, Revised Selected Papers*, pages 300–311. Springer, 2005.