



HAL
open science

ANIMITEX project: Image Analysis based on Textual Information

Hugo Alatrasta Salas, Eric Kergosien, Mathieu Roche, Maguelonne Teisseire

► **To cite this version:**

Hugo Alatrasta Salas, Eric Kergosien, Mathieu Roche, Maguelonne Teisseire. ANIMITEX project: Image Analysis based on Textual Information. SIMBig: Symposium on Information Management and Big Data, Sep 2014, Cusco, Peru. pp.49-52. hal-01144395

HAL Id: hal-01144395

<https://hal.science/hal-01144395>

Submitted on 21 Apr 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANIMITEX project: Image Analysis based on Textual Information

Hugo Alatrística-Salas and **Eric Kergosien** and **Mathieu Roche** and **Maguelonne Teisseire**
TETIS (Irstea, Cirad, AgroParisTech), Montpellier, France
LIRMM (CNRS, Univ. Montpellier 2), Montpellier, France
firstname.lastname@teledetection.fr

Abstract

With the amount of textual data available on the web, new methodologies of *knowledge extraction* domain are provided. Some original methods allow the users to combine different types of data in order to extract relevant information. In this context, this paper draws the main objectives of the ANIMITEX project which combines spatial and textual data. The data preprocessing step is detailed.

Keywords: Knowledge extraction, Text mining, Satellite images, Spatial feature identification

1 Aims of the ANIMITEX project

A lot of high resolution satellite data are now available. This raises the issue of fast and effective satellite image analysis as it still requires a costly human implication. In this context, remote sensing approaches enable to tackle this challenge. The exploratory and ambitious ANIMITEX project¹ aims at processing massive and heterogeneous textual data (i.e. *big data* context) in order to provide crucial information to enrich the analysis of satellite images.

The large amount of data are associated to a temporal repetitivity that increases. For instance today around ten images are available per year (e.g. SPOT, Landsat), and in 3 years, one image every 5 days (based on Sentinel-2 satellites) will be available.

¹<http://www.lirmm.fr/~mroche/ANIMITEX/> (web site in French)

The ANIMITEX project has many application areas such as image annotation (Forestier et al. 2010). For instance, identifying the precise type of culture or the function of a building is not always possible with the only use of images. Nevertheless, textual data could contain this kind of information and give additional meaning to the images. The development of approaches based on image/text matching becomes crucial in order to complete image analysis tasks (Alatrística-Salas and Béchet 2014). It also enables a better classification of data.

Moreover, image-text matching will enrich Information Retrieval (IR) methods and it will provide users a more global context of data (Sallaberry et al. 2008). This can be crucial for the decision maker in the context of land-use planning projects that have to take into account opinions of experts related to a territory (managers, scientists, associations, specialized companies, and so forth).

In the context of the ANIMITEX project, we plan to investigate two specific scenarios: (i) The construction of a road on the north of Villeveyrac (city close to Montpellier, France), (ii) A port activity area, called *Hinterland*, in Thau area (near to Sète, France). The aim of this case studies is to enrich images with information present in documents, e.g. the opinions extracted in newspapers about land-use planning.

The section 2 describes the proposed data preprocessing step. The section 3 details the partners involved in the project.

2 Data preprocessing process

The current work focuses on adapting of Natural Language Processing (NLP) techniques for recognition of Spatial Features (SF) and thematic/temporal information (Gaio et al. 2012; Maurel et al. 2011). In the proposed approach, SF appearing in a text, are composed of at least one Named Entity (NE) and one or more spatial indicators specifying its location (Lesbegueries et al. 2006). For this, a set of articles (i.e. 12000 documents) concerning Thau region between the years 2010 and 2013 has been acquired. A second part of the data set is composed of raster files (image mosaics Pleiades - spatial resolution 2x2 m - 4 spectral bands) covering all regions of the Thau lagoon (See Figure 1). Satellite images are available via the GEOSUD Equipex².

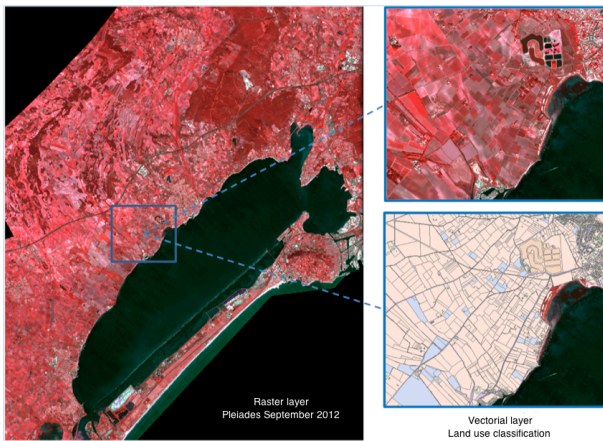


Figure 1: Mosaic images Pleiades around the Thau basin, images on the right represent the superposition of a vector classification on the raster file.

A detailed classification of the land occupation is currently in progress. It will lead to a digital vector layer where each SF (represented by a polygon) belongs to a class of specific land use. The nomenclature of this classification is organized into four hierarchical levels (See Figure 2). Moreover we investigate multi-scale information associated with different levels of classification of satellite images.

From this corpus, NLP methods have been applied in order to identify linguistic features concerning spatial, thematic, and temporal information in the documents. The combined use of lexicons

²<http://www.equipex-geosud.fr/>

and dedicated rules (Gaio et al. 2012) allows us to identify the absolute (e.g., Montpellier) and relative (e.g., south of Montpellier) Spatial Features (ASF and RSF) (Lesbegueries et al. 2006; Kergosien et al. 2014). A first framework based on sequential pattern mining (Cellier et al. 2010) has been proposed to discover relationships between SF (Alatriza-Salas and Béchet 2014). To this end, a two-step process has been defined (See Figure 3).

SF validation: for each identified ASF, we check on external resources if there is a corresponding spatial representation. In particular, we have used layers provided by the IGN³ (municipalities, roads, railways, buildings, etc.). In addition, if an ASF does not present on IGN resources, we use gazetteers (Geonames and Open Street Maps) to complete the information. Concerning the representation of RSF, we use spatial indicators of topological order associates to ASF.

Following the scopes proposed in (Sallaberry et al. 2008), the spatial indicators of topological order have been grouped in five categories:

- Proximity: different indicators can be used in relationship of proximity, such as: *near*, *around*, *beside*, *close to*, *periphery*, etc..
- Distance: the indicators used in this relationship are of the form: x km, x miles, etc.. Two representations are then proposed in our approach: 1) calcul of distance from the centroid of the ASF and construction of a circular buffer of size x from the centroid; 2) regarding the shape of the ASF and building a buffer of size x from the edge of the processed ASF .
- Inclusion: this binary operation allow us to check if an ASF is inside another taking into account indicators such as: center, in the heart, in, inside, etc.
- Orientation: This unary relationship has been broadly studied in the literature. Different approaches have been proposed to identify a cardinal points of an ASF. We have chosen to use the conical model proposed in (Frank 1991). For this, we use the centroid of ASF and we

³Institut National de l'information Gographique et forestire, i.e. National Institute of Geography

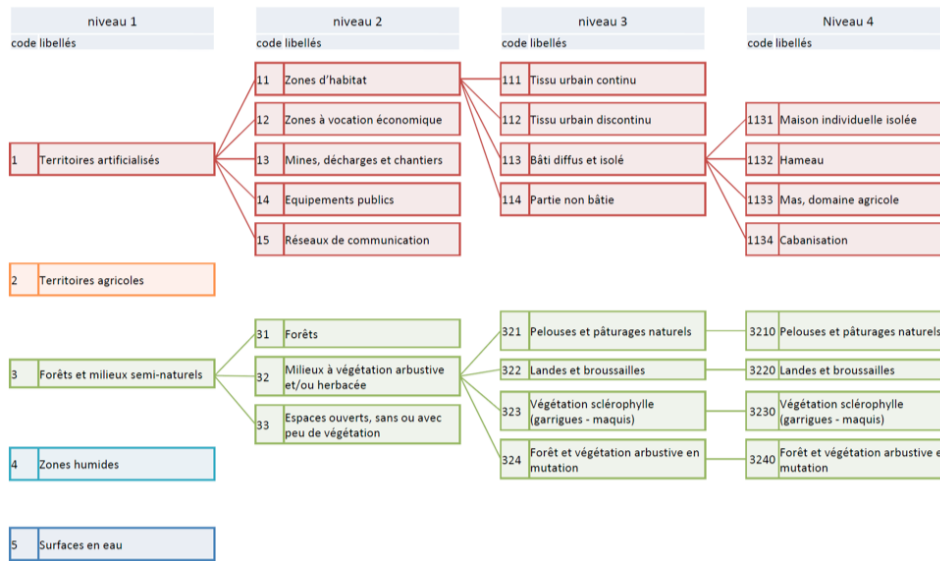


Figure 2: Nomenclature of Thau region used to image classification (in French)

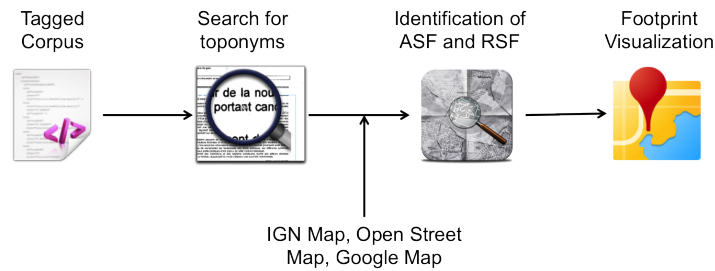


Figure 3: From document to footprint definition: the two-step process

build a buffer around. The size of this buffer will be calculated taking into account the surface of the studied ASF. Then we decompose the buffer into four equal areas (forming a "X") from the centroid. Each intersection between the buffer and cones thus formed represent the four cardinal points.

- Geometry: geometry relations are built from at least two ASF. These relationships are, for example, the union, the adjacency, the difference or a position of an ASF with respect to other ASF, for example, C between A and B (where A,B and C are ASF), etc.

Representation of the spatial footprint: after the extraction step and spatial representation of the ASF and RSF, the spatial footprint associated with

the treated document can be mapped. In this process, two main problems have been identified. The first one is the persistent ambiguity of some NE contained in SF because of some NE (e.g. *Montagnac*) corresponding to several places. To address this issue, a configurable spatial filter based on predefined scenarios has been developed. For example, to identify events related to a specific land-use planning project occurred in a part of the area of the Thau lagoon, only the SF contained in this area will be explored. The second issue is related to the use of external resources and the identification of the spatial representation appropriate to each ASF. Taking into account the spatial indicator (e.g. town, road, etc.) preceding by the toponymic name is a first answer because it allows us to specify the type of the SF and thus take into account the appropriate spatial

representation.

Thematic information is identified by semantic resources (i.e. AGROVOC thesaurus, nomenclature resulting of image classifications ...) (Buscaldi et al. 2013).

These linguistic features allow us to identify specific phenomena in documents (e.g., land-use planning, environmental change, natural disasters, etc.). The main idea is to link the phenomena identified in images with subjects found in documents during the same period. Overall, the ANIMITEX project allows the users to integrate different information sources, i.e. both types of expressions (texts vs. images). The main objective is to enrich the information conveyed by a text with images and vice versa.

3 Consortium of the project

The multidisciplinary consortium of the project involves three research domains: Computer Science, Geography and Remote Sensing. More precisely, the expertise in remote sensing and complex mining and heterogeneous spatio-temporal data, is one of the foundations of the project.

TETIS (Territories, Environment, Remote Sensing and Spatial Information, Montpellier) aims to produce and disseminate knowledge, concepts, methods, and tools to characterize and understand the dynamics of rural areas and territories, and control spatial information on these systems. LIRMM (Informatics, Robotics and Microelectronics, Montpellier) focuses on knowledge extraction. ICube (Strasbourg) is specialized in image analysis and complex data mining. ICube collaborates with geographers from LIVE laboratory (Image Laboratory, City, and Environment) and specialists in NLP (LiLPa lab – Linguistics, language, speech). These two locations (Montpellier and Strasbourg) constitute a cluster of local skills related to all major aspects of the project. LIUPPA (Pau) includes researchers specializing in Information Extraction (IE) and Information Retrieval (IR). The main work of this partner is about extraction and management of geographical information. GREYC (Caen) brings researchers in data mining (e.g. mining sequences in order to discover relationships between spatial entities) and NLP. For this aspect, a collaborations with two other labs is developed (LIPN and IRISA).

Acknowledgments

The authors thank Midi Libre (French newspaper) for its expertise on the corpus and all the partners of ANIMITEX project for their involvement. This work is partially funded by Mastodons CNRS grant and GEOSUD Equipex.

References

- Alatrasta Salas H., Béchet N. Fouille de textes : une approche séquentielle pour découvrir des relations spatiales. In Cergeo Workshop - EGC, 2014
- Buscaldi D., Bessagnet M.N., Royer A., Sallaberry C. Using the Semantics of Texts for Information Retrieval: A Concept and Domain Relation-Based Approach. Proceedings of ADBIS (2) - Advances in Databases and Information Systems, pp. 257-266, 2013.
- Cellier P., Charnois T., Plantevit M., Crémilleux B. Recursive Sequence Mining to Discover Named Entity Relations Symposium on Intelligent Data Analysis, LNCS, pp. 30-41, 2010.
- Forestier G., Puissant A., Wemmert C., Gançarski, Knowledge-based Region Labeling for Remote Sensing Image Interpretation Computers, Environment and Urban Systems, Vol. 36(5), pp. 470-480, 2012
- Frank A. U. Qualitative spatial reasoning with cardinal directions. In *Seventh Austrian Conference on Artificial Intelligence*, volume 287 of *Informatik-Fachberichte*, pages 157–167. Springer, Berlin Heidelberg, 1991.
- Gaio M., Sallaberry C., and Nguyen V.T. Ty-page de noms toponymiques à des fins d'indexation géographique. *TAL*, 53(2):143–176, 2012.
- Kergosien E., Laval B., Roche M., Teisseire M. Are opinions expressed in land-use planning documents? *International Journal of Geographical Information Science*, Vol. 28(4), pp.739-762, 2014.
- Lesbegueries J., Gaio M., and Loustau P. Geographical information access for non-structured data. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, pages 83–89, New York, NY, USA, 2006.
- Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., and Nouvel D. Casen: a transducer cascade to recognize french named entities. *TAL*, 52(1):69–96, 2011.
- Sallaberry C., Gaio M., and Lesbegueries J. Fuzzifying gis topological functions for gir needs. In Jones C. B. and Purves R., editors, *5th ACM Workshop On Geographic Information Retrieval*, pages 1–8, 2008.