Ontology-based management of ancient "lettrines"

Alain Bouju [a,*] and Mickaël Coustaty [a]
[a] *L3i Laboratory, La Rochelle University, Avenue Michel Crepeau, 17042 La Rochelle, France*
E-mail: alain.bouju@univ-lr.fr

Abstract. A large amount of historical documents have been digitized over the years. Browsing into these Cultural Heritage data can be done using query by keywords or query by example systems. Going from one kind of query to another raises the problem of the semantic gap. Identify region inside image could be needed. In order to deal with this problem, this paper presents an ontology-based approach to the resolution of the semantic gap problem that uses a semantic web approach with historical images. To do this, historians' knowledge and knowledge from the document processing domain were modeled using dedicated ontologies. Then, links between the regions of interest from the computer vision algorithms on the one hand, and their meaning on the other hand, were created. These links will subsequently be used to help historians to search, query, analyze and enrich images dataset. Based on the three ontologies defined or reused and combined in this approach, we have defined rules to automatically annotate an image (to define the background for example), or a part of an image (to identify a letter, a body-part, ...).

Keywords: Semantic, Ontology, Document, Image, Initials

1. Introduction

Historical documents constitute the memory of our societies, and a reminder of our Cultural Heritage. With the evolution of digitization techniques, there has been a drive all over the world to preserve historical documents. This huge number of documents needs to be processed and indexed so that historians can retrieve them and analyze their content (year of production, where they were printed, historical and social studies etc.).

Among the various elements present in a historical document, we are particularly interested in lettrines (also known as drop caps). In this article, we will use the French word "lettrines" (specific to French documents) rather than the term drop caps, to represent an enlarged letter at the beginning of a paragraph. These images are of particular importance to historians as they can be used to identify many different types of information (they were printed using wood stamps belonging to a specific printer at a defined period). Thus, retrieving similar lettrines provides information about the era and social life at that time.

Information retrieval is a very active field of research and, more specifically, image retrieval is becoming an increasingly difficult task that involves handling textual data and the image content. To overcome this problem, a very active research area [23, 30, 36, 41], generally known as Content-Based Image Retrieval, is devoted to describing and summarizing the content of images by computing descriptors. These descriptors extract low-level features of images in or-

[*] Corresponding author. E-mail: alain.bouju@univ-lr.fr.

der to summarize the information (low-level features generally based on color, shape, texture and nowadays computed using convolutional neural network). Starting from these features, searching for an image consists in finding images with similar features. The advantage of image descriptors is that they enable comparisons to be made between images, but they cannot be used to make a direct comparison with a text request from a user. This is the well-known semantic gap problem.

Dealing with the semantic gap requires a search engine that can query by example (when the end-user gives an example of what he is looking for) and by keyword (when the end-user gives a description of what he is looking for using specific keywords). This paper focuses on the use of ontologies to define explicit semantics between concepts from historical images which adds semantics to low level features and words from natural languages. The aim is to create links between keywords and visual concepts. Thus, two ontologies were defined: the first is dedicated to keywords from historians (experts on the Renaissance); the second is associated with image processing (extraction and description of a region of interest). These ontologies are fed using historians' knowledge and the results of image processing. Our approach allows to identify region inside an image with their descriptor. Then, these two ontologies and a spatial ontology are linked into a final ontology that is enriched using inference rules. Our goal is to provide relations between image regions and historian's keywords. In a first work [11], we have found some rules to provide relations but with some performance issues. In this article we propose a new approach with SPARQL UPDATE and a spatial triplestore.

This article will first give a detailed presentation of the semantic gap problem, followed by a description of the proposed model for which navigation tools were developed to help retrieve historical document images. In the fourth section, a case study used to validate our approach is presented, and finally the results and significance are discussed in the last section.

2. Semantic gap, ontologies and images

2.1. Semantic gap and annotations

A huge amount of digitized documents that have now been created needs to be managed efficiently. Efficient management can be achieved through indexing processes and semantic annotations that provide dedicated search tools to meet the needs of the end users. Effective tools must therefore be developed that enable a precise semantic description of images. To meet the rapid growth in multimedia content, several studies have addressed the well-known problem of the semantic gap [3] by attempting to provide an analysis and semantic interpretation of images. This semantic gap corresponds to the difference between the users' representation of an image and a description based on low-level features.

Several categories of techniques have been identified to solve the semantic gap problem [26]. Different methods of annotation can be identified, ranging from the use of ontologies to learning techniques and automatic annotation: *free* where no annotation vocabulary is predefined in advance and the user uses his own knowledge to annotate ; *annotation by keywords* where the user uses a predefined vocabulary (no relation between words) to annotate images. In this case, a hierarchy between keywords exists but is limited to a one-to-one relation between keywords (taxonomy); *annotation by ontology* where a hierarchy exists between the keywords suggested to the user and this hierarchy emerges from an analysis process (concepts are connected by many kinds of relations, not all of which are taxonomic).

Automatic image annotation is a major challenge. They were introduced in the early 2000's and the first studies dealt with machine learning based on statistics and probabilities. These approaches provide powerful and efficient tools to create links between visual features and semantic concepts like the approaches developed by [4, 14, 23].

[28] were the first to use the information from the context (perceptual context) in an image annotation process. They proposed a generative statistical model that computes a joint probability to associate keywords with image regions. An image is then described by all its automatically extracted regions (the results of computer vision processes) and its keywords. The relationship of extracted regions represents the context, while the association of keywords to the image regions represents the semantic. However, this method does not really capture the semantics of images, but instead applies statistics to the context to improve the description of the image.

The work done by [40] firstly described the process of feature extraction and representation, and then listed a set of annotation methods in the second step. These methods, based on machine learning (SVM, decision trees, neural networks, Bayesian networks), are used to classify the images' features based on the concepts previously learned. However, these methods have to deal with the problems of the learning step (a time-consuming problem; the number of classes must be small and manually defined).

[37] emphasized the role of the user, the purpose and the context in which the image's annotation process is performed. The generality, accuracy and choice of vocabulary keywords are dependent on the application and on user knowledge. Different degrees of abstraction can also be considered in the image's annotation process: contextual, cultural, emotional, technical, etc.

A method to automatically construct a hierarchy of concepts was proposed by [17] that combined visual information, conceptual information and contextual information.

Still seeking models that could help to link low-level to high-level features, some approaches were based on contextualized knowledge. These approaches assume that real world objects are always associated with their context, and the representation of that context is essential for image analysis and understanding. As contextualized knowledge can come from multiple sources, the heterogeneity of these sources creates a complexity that provides a more accurate description. In a specific context, the introduction of this knowledge helps reasoning and improves image annotation [3, 17, 32]. This improvement may rely on the use of semantic hierarchies, or incorporate *a priori* knowledge.

A study proposed by [7] attempted to reduce the sensory gap and the semantic gap. The *sensory gap* may be seen as the gap between the real scene and the acquired image, while the *semantic gap* corresponds to the difference between the user's representation of the image and a description based on low-level features.

More recently, many works used deep neural networks [2, 42] to find a common representation space between images and text. The idea behind those architectures is to learn links between some visual clues and the keywords using some image and textual embeddings. Some promising results start to appear (especially for the Visual Question Answering problem [5]) but the black box effect associated to these approaches is an impediment to its use by scholars and experts.

It is therefore important to use explicit and formal methods to represent knowledge. In this way it is possible to take into account knowledge associated with both the general context and the specific context. Moreover, this improves image understanding without being linked to the implementation used. Knowledge from the general context is then considered as the knowledge of the domain, and knowledge from the specific context is generally seen as the knowledge from images.

2.2. Ontologies

Ontologies approach have many advantages to represent knowledge for a given domain. They provide an *explicit and formal framework for a shared conceptualization* [16]. *Formal* means that they are machine-readable, and both humans and machines can apply reason to their content using the model. *Explicit* means that the type of concept used, and the constraints on its use, are explicitly defined. The idea of *shared* refers to the knowledge that is communally owned. Lastly, *conceptualization* refers to models obtained by an abstraction of phenomena that exist in the real world, and to the identification of the relevant concepts of these phenomena. Thus, ontologies capture the knowledge of a relevant domain, provide a common understanding of knowledge in the domain, determine the vocabulary of the domain, give an explicit definition of this vocabulary and the relationships between the terms of the vocabulary, and all of this is achieved using formal models.

[32], image content was modeled using Description Logic and Ontologies, demonstrating its importance in the interpretation of scenes. The authors noted that some errors can occur with an intuitive construction of knowledge and inference. Formal logic can avoid these errors. The authors then proposed a formal model for scene interpretation and emphasized the importance of spatial and temporal contexts in the task of interpretation.

The image processing ontology developed by [7] counted 279 concepts, 42 roles and 192 restrictions. In this ontology, the concepts were ordered into different levels: a physical level to process information from acquisition to storage of the image; a perception level to process visual features of the image; a semantic level to establish the generalization relationships and inclusion between concepts; a task level process for feature extraction and feature detection; and a stress level to express the efficiency and robustness of the system. However, no correspondence was established between the concepts outlined and the semantic level meta-data. The link between the user's knowledge and the regions extracted from the image was not represented as it was context dependent.

The problem of ground truth (or learning database for machine learning based methods) and annotation quality measurement arises in the work mentioned above. The interpretation of an image depends on the user's knowledge, background and knowledge of the degree of granularity. The use of an ontology has different goals: a unified description of image features, a visual characterization of the relationship between features (lines, regions, etc.), the use of contextual information, and finally to make a connection between the visual level and the semantic level.

Several categories of ontologies can be distinguished in the image annotation process as described by[19]:

High-level ontology (or thesaurus) images are described using metadata independent from the image content (*e.g.* date, author name);

Low-level ontology images are described using metadata based on low-level features (*e.g.* texture, color). The ontology then represents the different methods of image regions analysis;

Linked-level ontology this ontology links the elements from the low-level features (texture, color) to those with a high-level of semantics in the image (car, building). Regions of the image are described by their low-level features, and the ontology provides meaning for certain regions. This category reduces the semantic gap.

The first category of ontologies simply associates a vocabulary (list of keywords) to an image without any processing, and the words are associated with the entire image. In the last two categories, the image processing algorithms are first applied. The ontology is then used to annotate the results of these algorithms, and the keywords are associated with the image content or the image's regions.

There are many work to properly annotate an image for retrieval [31, 43]. In this work, we wish to manage region inside an image.

The challenge we are confronting in this paper (*i.e.* the problem of content-based historical image retrieval) imposes two major restrictions. First, we must design a system that is able to model and structure knowledge from the domain of history (semantic concepts) and those from the image processing domain (low-level features). On the other hand, this system must address the problem of semantic gap reduction between low-level features and semantic concepts used in historians' queries. This challenge cannot be overcome with only one kind of ontology, as defined in [19]. We therefore propose to use a combination of ontologies to have not just a single representation of an image, but different levels of representation. In our case, we decided to represent historians' knowledge with a high-level ontology (images are described with keywords), while knowledge related to image processing results are represented with a low-level ontology (images are described with radiometric features). So, we have at least two kind of expertise with different models. Finally, a linked-level ontology was used to connect the two previous ontologies and to attach different levels of description to the same image. Semantic gap reduction was achieved using inference rules to automatically create links between low-level and high-level concepts.

3. Proposed semantic gap reduction model

3.1. Proposed model

As mentioned earlier, an image can be described using keywords (high-level ontology) and pixel-based features (low-level ontology). Combining these two ontologies links two heterogeneous descriptions to the same image. The main contribution of this work is the use of inference rules to match concepts from these two different ontologies, and thus automatically annotate images (*i.e.* associate keywords to an image region with specific radiometric properties). This model is generic enough to be applied to a large number of domains by simply adapting each ontology to the new case study. Figure 1 presents an overview of the proposed model. Each domain was modeled with a specific ontology. By adding a link between them, we were able to mix heterogeneous data. Finally, the mapping between the various concepts derived from these two domains was performed using inference rules. We will present these ontologies and some examples of inference rules in the following sections.

3.2. Ontology of experts' domain

The description of images using an expert's vocabulary is based on our approach to a domain ontology. This consists of a content annotation which can be performed by a user in different ways: a free annotation without any predefined annotation vocabulary, where

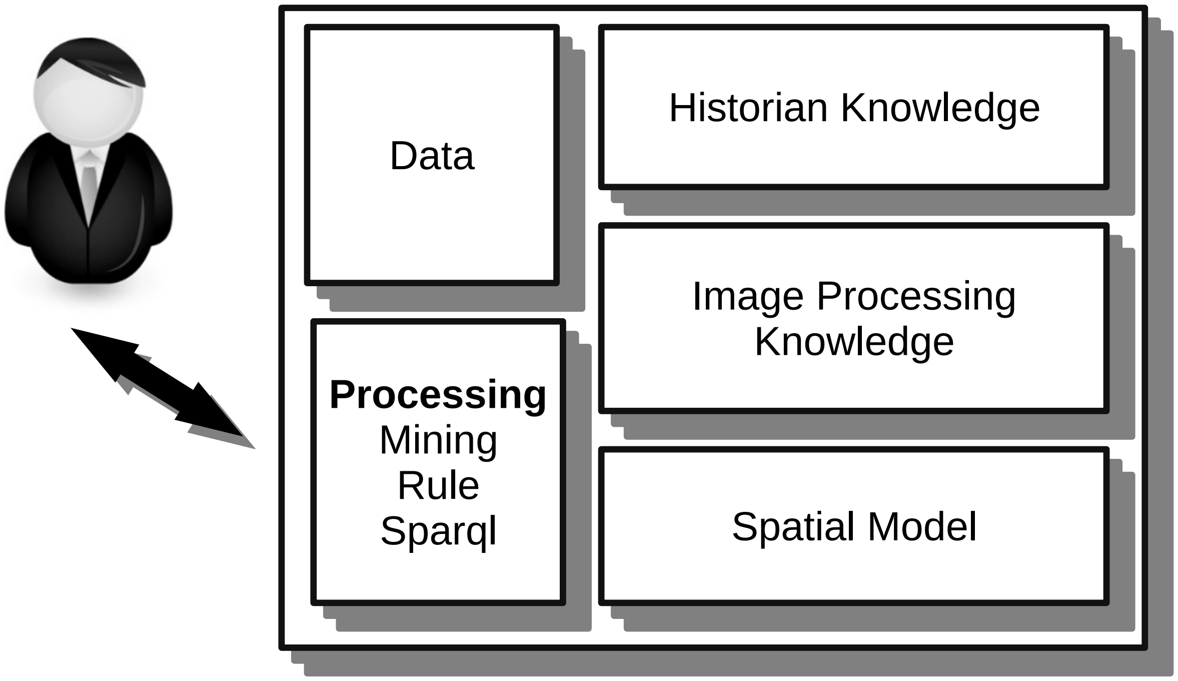


Fig. 1. Overview of the proposed system, with the combination of two expert knowledge and a spatial model into a single ontology : the ontology of "lettrines". The main contribution corresponds to rules that enable the creation of links between the different levels of information and thus reduce the semantic gap

the user uses his own knowledge; annotation via a predefined set of keywords (without hierarchy) ; an annotation via vocabulary defined by a hierarchy of words. In the third annotation case, the hierarchical vocabulary is used to identify semantic "target" elements, and the semantic gap reduction problem is simplified. The latter case corresponds to our work on the images of lettrines since a single thesaurus semantically structured by historians is provided.

3.3. Image processing knowledge

A large amount of work has been done in the image processing domain. One common point in this work is that it tends to segment an image into uniform subparts and to provide dedicated features related to these regions called "Regions of Interest" (RoI). The description of an image is then composed of two levels of information: a general one and a local one. Local information corresponds to a description of each region of interest on the image, while the general description contains information related to the original image and information related to the spatial organization of the regions of interest. We propose to model this in two ontologies, the first for regions of interest, while the second is devoted to the representation of spatial relations.

3.3.1. Regions Of Interest ontology

Regions of interest are the results of the image segmentation process which separates an image into homogeneous regions. In order to be able to compare regions and to identify similarities between them, each region is described using a feature vector computed from its pixel values. We have introduced the concept of *image* to represent the original image and *regionSet* to represent the set of regions of interest in an image. Each region of interest can be specialized (see Figure 2) by using local features specific for that region. This representation integrates features either for the initial image, for derivative images or for regions of interest. We thus recover the notions of general and local signatures of an image defining an image analysis domain.

As our aim was to produce a generic model for image processing results, we also added two other concepts: *DerivativeImage* for the regions not extracted from the original image but from an image obtained after a pre-processing technique; *ImageReferenceSystem* to get meta-information related to the image reference system (e.g. color spaces, image format).

3.3.2. Spatial ontology

Bearing in mind the generic vision of our model, it is important to be able to locate regions inside an image in relation to each other. This can be obtained with spatial relations that include topology (RCC8 defined by Cohn et al. [8]), distance (near or distant) and orientation (N, S, E, W). In particular, a ground truth indicating the spatial positioning of information in the image can be considered as a manual extractor. A comparison between a ground truth and automatically extracted regions provides information about the robustness of an extraction method and could be extended to large-scale experiments. This notion of extractor, under the terms of interpretation, was introduced in Lamiroy and Lopresti [27] to specify the result of an image processing algorithm.

Now that we have presented the general framework of our system, we will give a detailed example of how we used it for the automatic annotation of lettrines.

4. Case study: an ontology for lettrines

In order to validate our model, we defined an ontological representation for the images of lettrines. This ontological representation corresponds to the implementation of a general system that covers all aspects of the automatic image annotation. This ontology provides a standard representation for the heterogeneous and complex data that describe these images. The data consist of numerical characteristics computed from the image or from regions of interest in the image, spatial relations among image regions, and semantic data from the historian's knowledge of the domain. This ontology enabled us to reduce the semantic gap between pixel-level descriptions and semantic descriptions of the image thanks to automatic annotation of some images or image regions using ontological properties. This ontology is partly depicted in Figure 4 (the Image Processing Ontology is accessible on http://pageperso.univ-lr.fr/alain.bouju/ImageProcessingOntology/). Figure 3 describes the visual vocabulary we used to represent ontologies in the sequel to this paper.

We defined this ontology in several phases, and each is described in separate sections below:

- **The knowledge of historians (section 4.1):** Knowledge of historians is described by a thesaurus proposed by [24] that was reformulated as an ontology with Lettrines as the main class. A fund of 4288 lettrines, manually annotated by historians, was used to populate this ontology.
- **The ontology of regions of interest (RoI) (section 4.2):** This ontology, with Image as the main class, consists of regions of interest extracted from the image of lettrines with their low-level characteristics. This ontology is populated by 909 lettrine images (only 909 images of the 4288 lettrines were available), 5588 regions corresponding to shapes, and 451711 regions containing strokes that were extracted using algorithms defined in [10, 33]. This ontology was then enriched by partitioning the image and introducing spatial relations in order to locate each region in the center or at the edge of the image.
- **The ontology of lettrines (section 4.3):** The historians' ontology provided a high-level representation of lettrines, while the RoIs ontology provided a lower-level one. We combined them in a single ontology by linking their respective main classes (Lettrine and Image). In this way, we obtained a homogeneous representation and querying of all the data available on the lettrines. The main contribution of this paper was to attempt to reduce the semantic gap between the two levels by enriching the ontology. This was done using an inference mechanism that was able to bridge the gap between historians' keywords and the image processing features. In order to assess this contribution, we will now provide more details on two inference processes:

1. **Inferences on the RoIs extracted from shapes (section 4.4):** we added logical descriptions of properties that enabled us:
   - ∗ to identify and annotate a region as the letter in the lettrine image (isLetter property)
   - ∗ to identify and annotate certain regions as parts of figures in the background of the lettrine (isBody property)
2. **Inferences on the RoIs extracted from textures (section 4.5):** the image consists of a set of regions and it is these that provide the relevant information. We used the characteristics of the set to determine whether the lettrine had a hashed background (isHashed property)

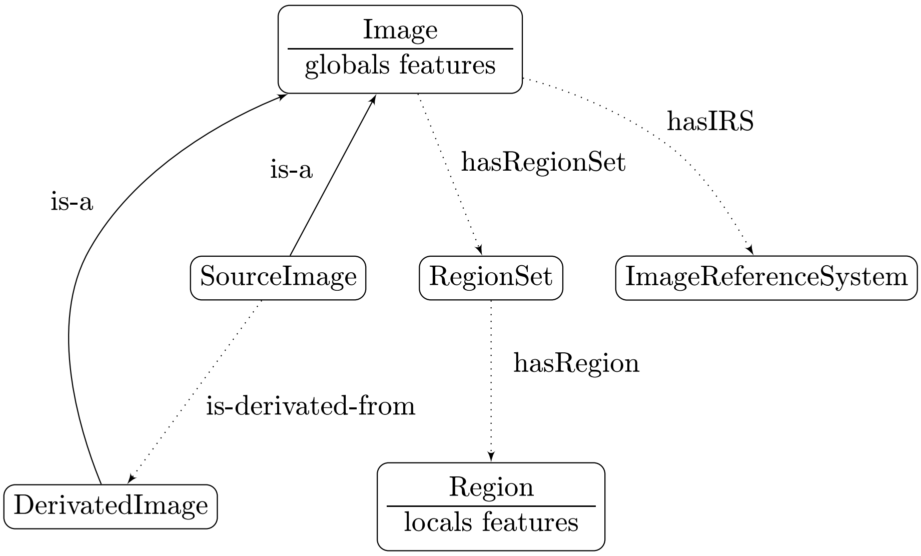
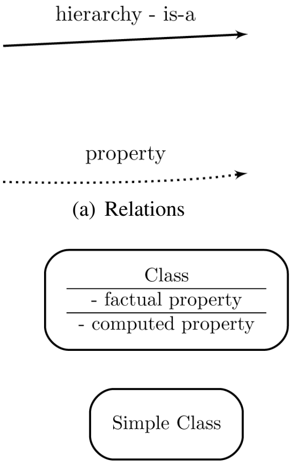


Fig. 2. Ontology of generic regions



(a) Relations

(b) Classes

Fig. 3. Symbols used to describe ontologies

4.1. Ontology provided by historians

Images of lettrines are graphic images found in ancient documents from the XV^{th} and the XVI^{th} centuries. Figure 5 gives a few examples. Lettrines are decorated capital letters at the beginning of a paragraph that are common in books of that time. They were obtained from wooden hand sculpted stamps. Their main component is a letter, but they are also characterized by a background that can be ornamental or represent social scenes of the time (figurative scenes). Nuances and shadows were obtained by parallel strokes.

The *Centre d'Etudes Superieures de la Renaissance* (Higher Education Center for Renaissance Studies - CESR) of Tours, France, works on lettrines. The lettrines provide historians with information that situate documents in time and they also study the social scenes in the figurative backgrounds. The stamps were used many times, and the ageing process makes it possible to derive a chronology of the documents with respect to each other. Furthermore, stamps are often characteristic of a particular sculptor.

CESR historians proposed a semantic description of the images of lettrines based on the work of [24]. Starting from this work, a lettrine can be decomposed into four layers. Each layer provides specific information (see Figure 6):

Letter: placed at the center of the image, the letter layer characterizes in particular the letter it contains (*e.g.* A, B), its color (black or white), the alphabet (Latin, Greek, Hebraic), and the font (roman, gothic)

Pattern: consists of ornamental forms that can be decorative or figurative

Background: can be uniform (black or white), hatched, or honeycombed

Frame: corresponds to the edges of the typographic stamp. It can consist of zero, one or two lines

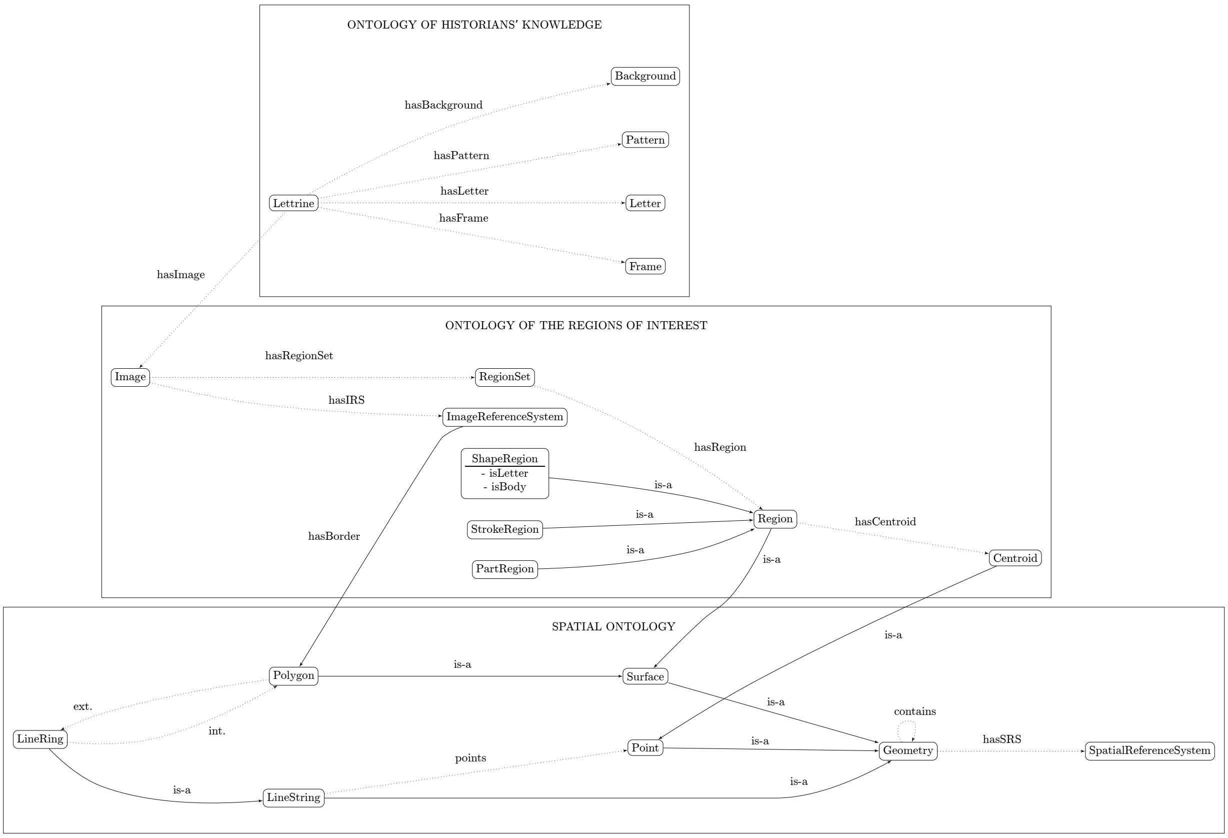


Fig. 4. Excerpt from the general ontology of lettrines where only computed properties of class are presented. The ontology of historian's knowledge is fully detailed in Figure 7

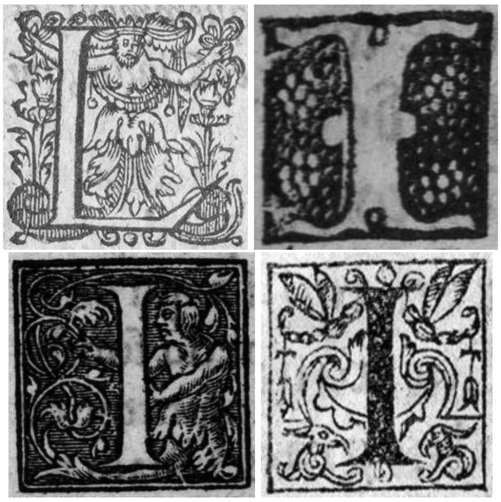
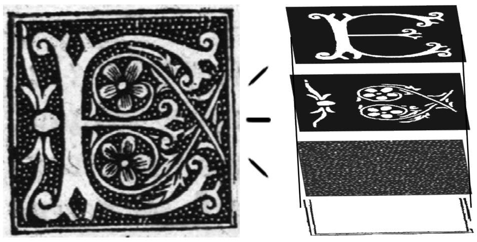


Fig. 5. Exemples of lettrines

We transcribed this description of lettrines into an ontology (ontology of the historians). Figure 7 describes the T-Box of that ontology, which consists of the main Lettrine class with a link to each semantic layer (Letter class, Background, Pattern and Frame). The *hasLetter*, *hasBackground*, *hasPattern* and *hasFrame* properties were introduced for these links. Class Letter itself was linked by properties to the ColorLetter, TypeFont and Alphabet classes. The IdentificationLetter property specifies a given letter. We then



(a) Original image (c) Various layers

Fig. 6. Layer decomposition by historians

populated the A-Box of the ontology with 4288 lettrines which were hand-annotated by historians.

4.2. *Ontologies of the regions of interest*

The images of lettrines are particularly difficult images to process due to their specific nature and to deterioration over time (yellowing of the paper, spoiled and stained pages, binary images composed of strokes). Although some standard texture-based approaches have been defined for natural and/or color images, they are not efficient with this kind of image. Techniques that are not sensitive to these deteriorations and are adapted to these features have been developed, and a specific image characterization process was developed in [10, 12]. We used it to define the ontology of the regions of interest.

The method used relies on a series of processes. First, the image is decomposed into two layers (see Figure 8). The decomposition method, described in [13], results from a sequence of projections presented in [18] that are especially relevant for the analysis of lettrine images. We extracted regions of interest from each of these two layers: the shape layer and the texture layer.

4.2.1. *Regions of interest from the shape layer*

The letter itself as well as the natural scenes in lettrines with a figurative background correspond to shapes in the image. For, this reason we set up a method adapted to the shape layer, which extracts some regions of interest. The method is described in [10]. The RoI's are obtained in three steps

1. First, the Zipf law [34] is used for its robustness to grey level variations and the absence of influence of the components' color. It can be used to segment the shape layer into connected components
2. Next, the connected components with an area greater than a given threshold of the image (1% in our experiments) are retained, the smaller ones were judged to be less pertinent. The remaining connected components form the RoI's of the shape layer
3. Finally, characteristics describing their shape are associated with each region:
   - The eccentricity *Ecc* of the region is defined as the ratio between the minor radius r_m and the major radius r_M of the minimal ellipse encompassing the region [35]: $Ecc = \frac{r_M - r_m}{r_M + r_m}$
   - The mean of the grey levels (*GreyMean*) and their standard deviation (*GreySTD*) serve as an estimate of the color of the region and its regularity
   - The Euler number E_n of the region as introduced in [35] provides an estimate of its compactness. In fact, it computes the number of holes H contained in the region, and $E_n = 1 - H$ gives an approximation of the compactness of each region

4.2.2. *Regions of interest from the texture layer*

The texture layer essentially consists of the strokes in the image. In the image of a lettrine, a set of strokes with similar visual characteristics (length, orientation, thickness, curvature) corresponds either to a hatched background or to shadows.

Based on the processing developed in [33], we set up texture layer processing to extract areas consisting of similar strokes. These regions are obtained in four steps:

1. We first applied a binarization process followed by noise removal to retain only the strokes
2. Strokes can have various widths. We thus used a distance transform algorithm [6] to obtain strokes with the same thickness
3. We then defined the characteristics of the strokes which are significant for human vision, namely length, thickness, homogeneity of orientation and curvature. A non-supervised classification based on these characteristics was used to create classes of similar strokes
4. Finally, stroke regions were obtained by grouping strokes with similar characteristics which are close together in the image. Some characteristics were then recomputed for the set of strokes of each region: the number of strokes (*StrokesNumber*) ; the average length of the strokes (*StrokesLenght*) ; the average thickness of the strokes (*StrokesWidth*) ; the orientation of the strokes (*StrokesOrientation*) ; the homogeneity (*StrokesHomogeneity*) ; the curvature of the strokes (*StrokesCurvature*)

4.2.3. *Enrichment of the ontology of the regions of interest*

The T-Box of the ontology of the regions of interest has a main class: Image (see Figure 11). General information about the image is provided by the *ImageReferenceSystem*, *hasCentroid*, *hasLength*, and *hasWidth* properties.

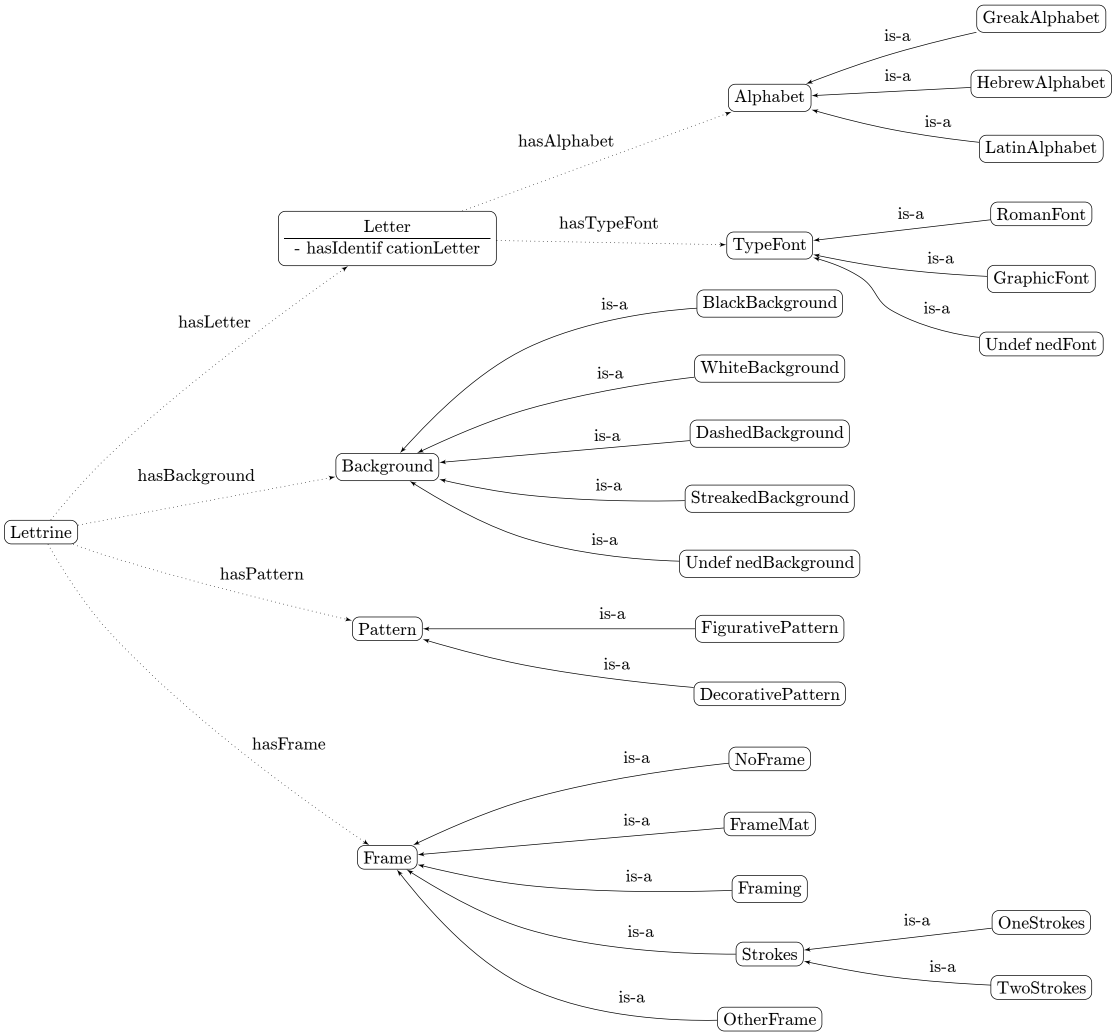


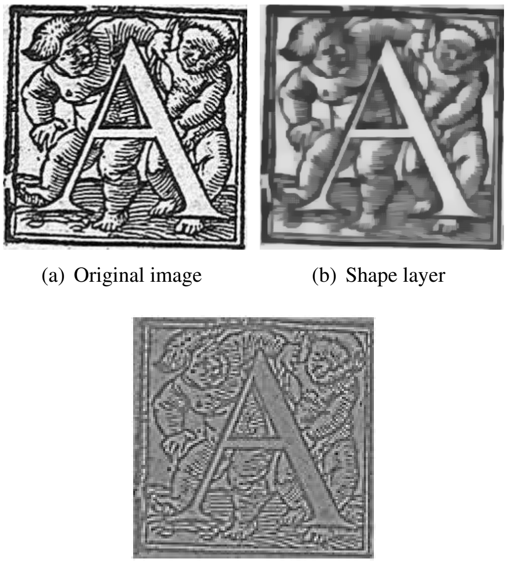
Fig. 7. Excerpt from the ontology of the historians (T-Box). This Lettrine class is related to the historian's knowledges which must be linked to the image concept.

The shape and texture layers are images which are derived from the initial image. Accordingly, the initial images belong to the *SourceImage* subclass, while regions of interest belong to the *DerivedImage* subclass.

RoI (Region of Interest) Classes and *ROISet* associate a set of regions with an image. The area of a region as well as the coordinates of its center of gravity are introduced by the *hasArea* and *hasCentroid* properties, respectively, to locate the region inside the image. Shape and stroke regions belong to their respective subclasses: *ShapeROI* and *StrokesROI*. Of course, they inherit the properties mentioned above.

We then enriched the ontology of the regions by adding some features linked to the partitioning of a lettrine into 9 areas (Figure 9). These features belong to the PartROI subclass of the RoI class, and spatial relations were then used to locate a RoI relative to an area of the partition. On the recommendation of the ISO standard in [1] on spatial data, we chose to use RCC8 algebra [8], which defines eight types of spatial relations between two spatial objects, as well as the SFS format (Simple Feature Specification) to represent spatial objects. SFS is a standard derived from the spatial specifications for SQL, recommended by the OGC consortium (OpenGIS Consortium) in [20] (see Figure 10).

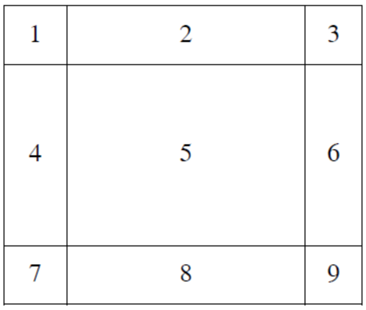
RoI Class and its subclasses inherit from abstract Geometry class from the SFS standard. RCC8 spatial



(a) Original image (b) Shape layer

(c) Texture layer

Fig. 8. Decomposition of a lettrine into layers: results obtained from the process proposed in [9]



1	2	3
4	5	6
7	8	9

Fig. 9. Image partitioning used to locate RoIs

relations are computed properties that can be used to compare regions. In particular, they locate regions in *ROIShape* and *ROIStrokes* with respect to the nine areas in *ROIPart*.

We selected 909 images of lettrines. Of these, 5588 regions were extracted from the shape layer, and 451711 regions from the texture layer. The A-Box of the ontology of regions of interest was populated by these images and these regions.

4.3. *Ontology of the lettrines*

Each of the two ontologies carries information on the lettrine images: the ontology of the historians provides semantic information, while the ontology of the regions of interest provides low-level information. That is why we created the *ontology of lettrines*, which combines them. The combination is performed by linking the main *Lettrine* class of the ontology of the historians to the main *Image* class of the RoIs ontology. The T-Box of the new ontology brings together all the classes and properties of the combined ontologies.

Meanwhile, the A Box is populated with the images (instances), that are classified according to the classes defined in the two Tbox, and which link together the lettrines and the images. Consequently, each image from the RoIs ontology is described not only by the extracted regions of interest and their low-level properties, but also by the manual annotations of historians. However, not all annotated images are present in the ontology of the regions of interest. Hence, some lettrines from the ontology of the historians are not described by regions of interest. This particularity offers an advantage in that our system does not require fully described images, it can be run on images with partial descriptions.

Integrating the data set describing a lettrine image into a single ontology enables us to envisage a reduction in the semantic gap between the low-level characteristics in the ontology of the regions of interest and their manual annotations in the ontology of the historians. The data were modeled and certain elements, required by historians in their historical image retrieval process, needed to be automatically extracted and annotated with historians' keywords. This represents a mechanism for semantic annotation of regions in images.

Our contribution to semantic gap reduction is based on an enrichment of the ontology. New properties were added to the T-Box in order to semantically annotate images or regions of interest. These properties were computed using a formula or rules of logic, which can include the set of data available for the image. These properties are presented in Figure 11 (*isHashed*, *hasCentroid*, *isLetter* and *isBody* properties).

We propose to do this using inference processes that are presented in the following two sections.

Actually, these methods are based on instance classification. The goal is to determine for each region of interest if it is a letter, a body, or if it has a hashed background. The first method is based on production rules while the second on a decision tree. We also eval-

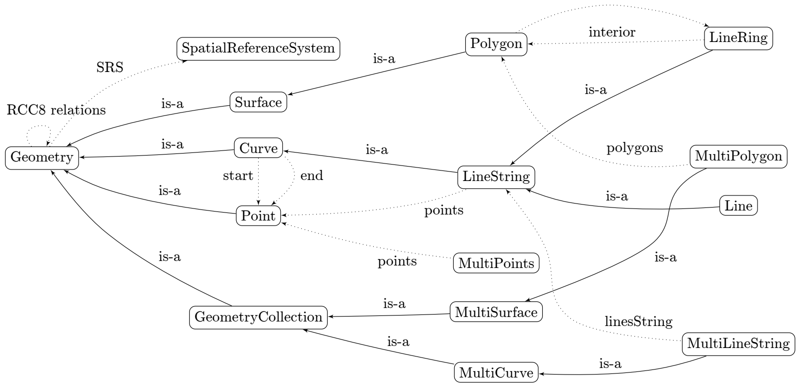


Fig. 10. Spatial ontology of the SFS standard from the OGC

uated the performance of these rules to assess the efficiency of our approach. It is very difficult to compare these results as no methods were found in the literature. Finally, we would like to mention the fact that this system enhances the readability of the results. The end user has an insight into the thought process, which is not possible with standard statistical approaches (often seen as black boxes).

4.4. Inference on the regions of interest of the shape layer

The letter, as well as elements of the natural scenes in lettrines with a figurative background, correspond to shapes in the image. Hence, we tried to identify the region of interest in the shape layer that corresponded to the letter (with the computed *isLetter* property) or to the body parts of the characters (with the computed *isBody* property). These two properties are based on low-level properties of the RoIs of the shape layer, and on the spatial features of images (see Figure 12). As for the *isBody* property, it has a specificity as it is expressed as a function of *isLetter* (the letter cannot be a part of the body of a character).

***isLetter* property: description and experimentation** This computed property is defined for the *ShapeROI* class and indicates whether a region is identified as the letter in a lettrine. It is deduced from four properties of the regions in the shape layer and also spatial information with respect to image partitioning:

1. Maximal area region: in the ontology of regions, the area is indicated for each region (*hasArea* property in the ontology of regions). Of all the regions that verify the properties given below, the one with maximal area is selected and labelled by the *isLetter* property.
2. Region located at the center of the image: check that the region is contained in the central area of the partition (spatial property *contains*) and does not intersect with the edge areas. Since the partition is as shown in Figure 9, this guarantees that the letter is inside the central area (the one numbered 5).
3. Region with few holes: the Euler number, computed for each region of the shape layer (*hasEuler* property in the ontology of regions), is used here. Based on our knowledge, regions with few holes were taken as those with a Euler number between -2 et $+2$ (the Euler number represents the number of white connected components in a black shape or vice-versa).
4. Containing the center of the image of the lettrine in its smallest surrounding rectangle: this property specifies that the center of the image (*hasCentroid* property in the ontology of regions) must lie inside the smallest rectangle surrounding the selected region, within a margin of 15 pixels.

The *isLetter* property was computed on the 909 lettrines regions that populate the ontology of regions. After manual verification, 816 regions were correctly identified as letters, and 103 were not. The error rate was thus 11%. Of these 103 regions, some correspond to a sub-part of the letter (see Figure 14). By using the knowledge from the historians' ontology (the *hasIdentificationLetter* property), we could extend the search not only by keeping the biggest shape, but many big shapes around the biggest one, in order to aggregate them. This search could then be guided by the letter identification (comparison between the knowledge and the result of an OCR applied to the shapes) using topo-

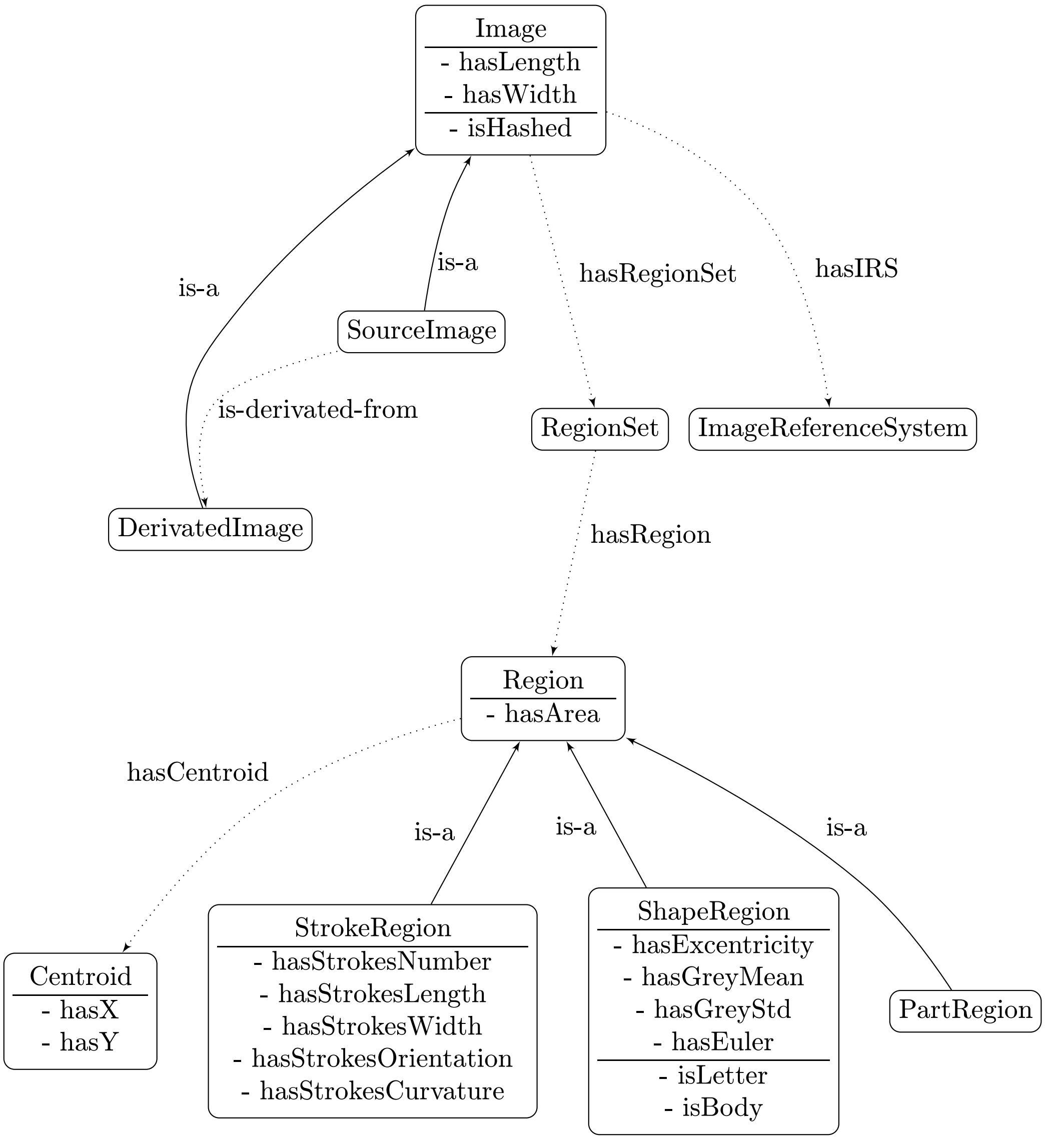


Fig. 11. Enriched ontology for the RoIs (T-Box part)

logical constraints related to the way letters are produced.

The *isBody* property: description and experimentation This property, defined for the *ShapeROI* class, indicates that a region has been identified as a part of a character in the background of the lettrine. It is computed from five properties that include properties of the shape regions as well as spatial relations with the partitioning and information from the ontology of the historians:

1. Region in a lettrine with figurative background: only regions in lettrines with a figurative background (the *hasPattern* property in the ontology of the historians) are considered as possible candidates.
2. Region located at the center of the image: as for the *isLetter* property, this is tested with the spatial *contains* property.
3. Region with few holes: as with the *isLetter* property, this test is based on the Euler number (the *hasEuler* property in the ontology of regions).

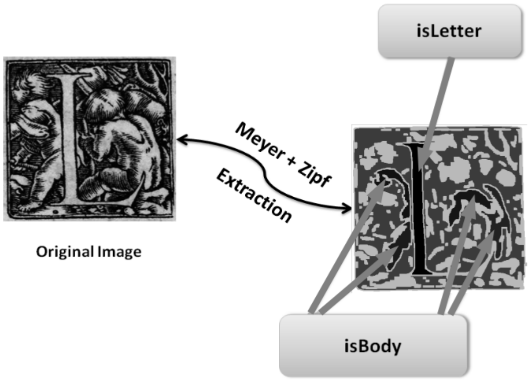


Fig. 12. Extraction of an area of interest from a lettrine

4. Region with light grey color: the image grey level goes from 0 (black) to 255 (white). Light grey was defined as greater than 90. The mean grey level (the *hasGreyMean* property in the ontology of regions) is compared to that threshold.
5. Region that is not a letter: among all the regions that verify the four preceding properties, only those that are not identified as letters are retained.

Validating annotated regions is a manual process. That is why 45 images of lettrines were randomly selected to apply the *isBody* rule. Of these, 27 had an ornamental background, the other 18 had a figurative background. 112 shape regions were extracted in total. The isBody property correctly labelled shape regions from 17 out of the 18 images with a figurative background, giving an error rate of about 2%.

4.5. *Inference on the regions of interest in the texture layer*

An inference mechanism for the regions of interest in the texture layer was set up to reduce the semantic gap between the ontology of the regions of interest and the ontology of the historians.

A lettrine with a hatched background contains regions of strokes with a total area large enough to cover a good part of the image. Hence, significant information is carried by the properties of the set of stroke regions of the image that could be used to decide if it has a hatched background. We tried to identify the images with a hatched background by introducing the computed *isHashed* property.

To that end, we first enriched the image description with properties that are related to its set of stroke regions.

The computed *isHashed* property was used to decide if a lettrine had a hatched background on the basis of low-level features. The result was checked by comparing it with the *hasBackground* property in the knowledge base of the historians. That is why we were able to use the standard supervised classification approach ($C4.5$ decision tree) to fix the thresholds of the low-level properties used to make the decision. Figure 13 shows a simplified decision tree, obtained from algorithm $C4.5$, where:

YX are tree nodes and *X* corresponds to each of the 6 properties specific to the regions in *StrokesROI*, namely *hasStrokesNumber*, *hasStrokesLength*, *hasStrokesWidth*, *hasStrokesOrientation*, and *hasStrokesHomogeneity*; *Y* corresponds to the minimal, maximal or average value, or to the standard deviation (Min, Max, Avg and Std) of each property of the set of all stroke regions in the image;

SV are labels for the branches of the tree leading to the class (in this case *isHashed*). S is a comparison operator: $<$, $\leqslant$, $>$, or $\geqslant$; V is the threshold to be compared with the attribute value.

(isHashed): comparison of the minimal, maximal or average values, or of the standard deviation of the properties of stroke regions to thresholds (7 leaves) obtained by supervised learning.

This rule was applied to the set of 909 lettrines populating the ontology of regions. Of those, 140 had a hatched background. 126 images had the *isHashed* property. An automatic verification was then possible by querying the ontology of the historians: 123 of the 126 had a hatched background. Hence the recall was $\frac{123}{140} = 87.8\%$, and the precision $\frac{123}{126} = 97.8\%$.

5. Discussion

Content based image retrieval for historical document images raises two major issues. First, the knowledge both from the domain of history (semantic concepts), and from the domain of image processing (low-level features) has to be modeled and structured in the same way. Secondly, in order to provide better answers to historians' queries, the semantic gap problem, *i.e.* bridging the gap between low-level features and semantic concepts, has to be taken into account.

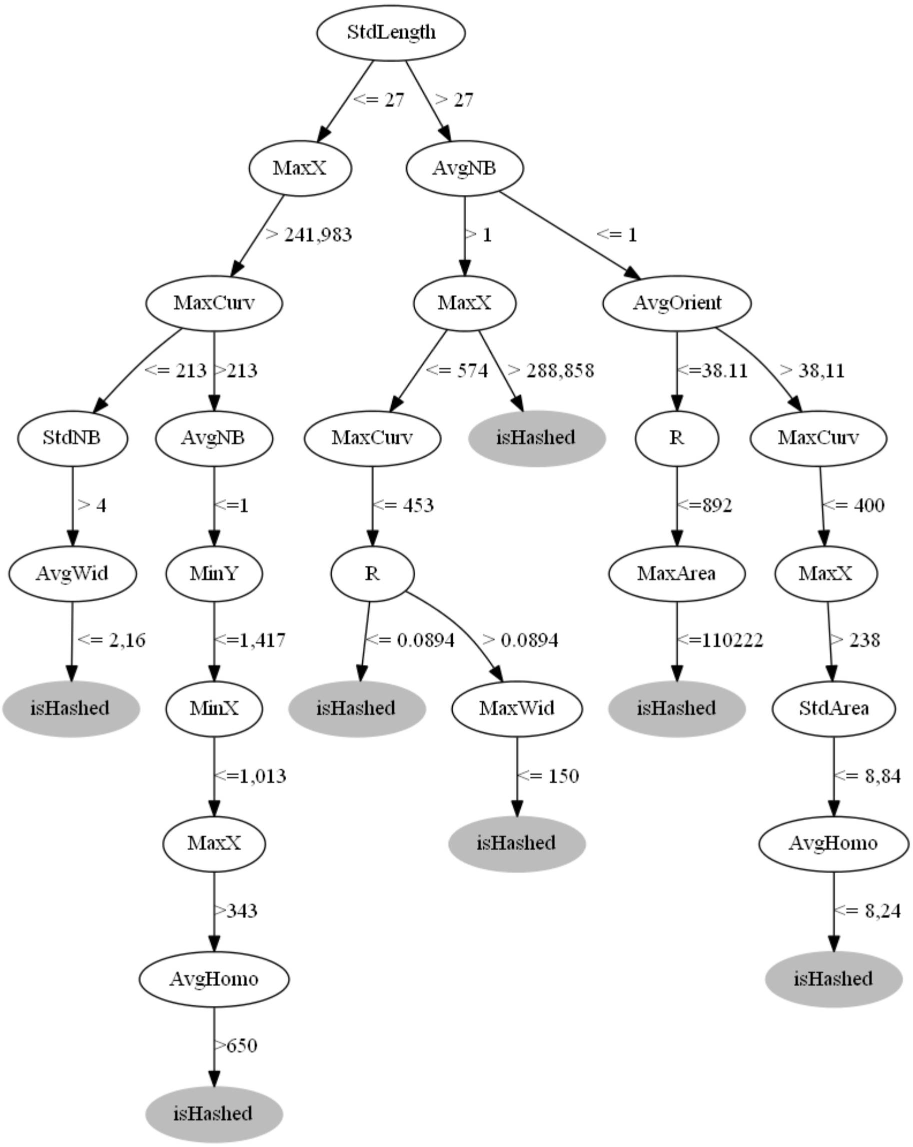


Fig. 13. Simplified decision tree for finding decision rules

5.1. Data structuration

Building an application that combines image processing features (low-level) and symbolic processing (high-level) needs efficient handling of large and heterogeneous data. The results of image processing identify regions of the image with their properties (*e.g.* region, connectivity), while symbolic processing relies on keywords from experts in the field.

In our approach, some rules were applied to identify some of these regions as semantic components of the document (*e.g.* this region is the letter in the lettrines, this region is a part of a body). All these data must be collected in a consistent manner to allow further deduction processes, navigating, extracting results as images, etc.

Figure 14 presents two examples of results obtained by a sequence of image processing algorithms and deducting processes. Such images are produced automatically starting from the recorded low-level features. They are used to evaluate our processes by comparing the results with the ground truth. In the first example (Figure 14), each image contains the initial lettrine, and the region that has been identified as the letter is superimposed. The first image describes a successful identification, the second a partial failure.

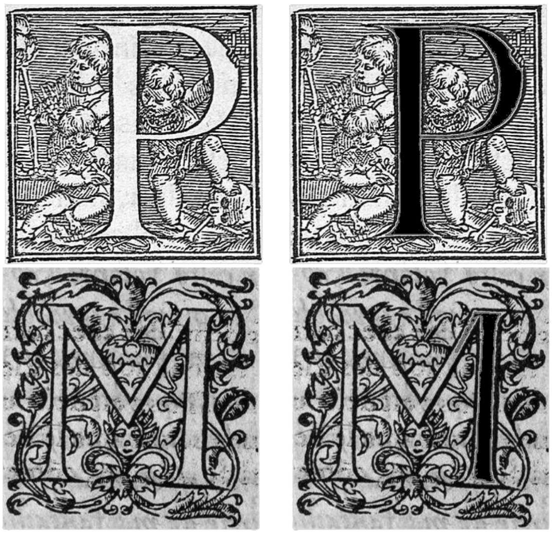
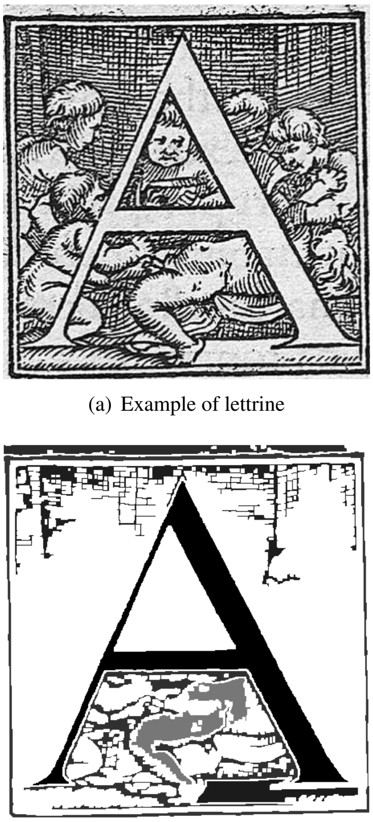


Fig. 14. Final result of a sequence of image processing based on a segmentation and deducting process.

The second example (Figure 15) represents a figurative lettrine (i.e having a figurative pattern) whose regions are not properly annotated by the *isBody* property. In the second image, all regions extracted from the shape layer are superimposed. The black region is correctly tagged with the *isLetter* property, while the light grey one, which corresponds to a character subpart, is not annotated by the *isBody* property (because its Euler number is too large). The use of a more complex or finer partitioning of the image, or other statistical features to describe regions, could help to correctly annotate regions. Similarly, a more flexible management of the classification (*e.g.* using rejection, confidence or a fuzzy classifier) may help to provide the user with candidate regions, but for which the system cannot take firm decisions.

5.2. Tools to model and their constraints

Alongside the issue of representation, the question of the choice of a relevant data representation that supports processing, deductions and calculations is critical.



(a) Example of lettrine

(b) Regions extracted form the lettrine shape's layer

Fig. 15. The image where regions are not properly identified as a part of body by *isBody*

The modeling stage could be carried out using UML. In general, the UML model is converted into another data model such as the relational model. However, a UML class diagram or database schema is located at the logical level and has the disadvantage of being static. Each new image processing, each new deduction would require a modification of the UML model or the relational schema. Although a UML model can represent a hierarchy of concepts, a relational schema does not represent it directly or explicitly. Despite the high level of effectiveness of relational databases systems, this possibility was therefore discarded. However, we used a UML model to define the spatial part of the ontology which was transformed into an ontology in [21] using the eclipseuml2owl tool.

The formal ontologies based on description logics provide a very flexible approach for structuring data. The T-Box, meta-level, describes the structuring of data that are contained in the A-Box. This approach is much less static than UML or the relational schema. Introducing a new concept or a new role does not require special precautions (except to check the consistency of the resulting model).

However, for large volumes of data, computation time using ontologies can become prohibitive, and this led us to use other tools such as SWRL rules [1] or DLV presented by [29] in our first work [11] and now we use a spatial triplestore. DLV is based on deductive logic programming, is simple to use, with ease of query expression, and a relatively low response time. Nevertheless, we used an ontological approach as a guide, even when we used these approaches that do not fit into this framework.

In order to validate our approach, we used a number of tools: PROTÉGÉ [25], an ontology editor to build the ontology, and the JENA API [22] [2] to populate it. We also used DLV[3], which is an implementation of Datalog [15], as a deductive database. DLV was used in command-line mode to define and test the inference rules presented in the subsequent sections. Based on logic programming, DLV is easy to use, makes the composition of queries easier, and has a rather short response time for middle size dataset.

To improve our system, we have chosen to use a spatial triplestore. Currently, several triplestores support storing and querying spatial data using the GeoSPARQL[4] or stSPARQL[5] query language. Many use PostgreSQL/PostGIS has backend for efficient spatial request. We have chosen Apache Marmotta an Open Platform for Linked Data [6] and his KiWi Triplestore over PostgreSQL/PostGIS. Now, SPARQL 1.1 [7] proposes CONSTRUCT and UPDATE operations. So, we rewrite our SWRL rules with SPARQL. YASQE[8] was used as editor. Probably we could have used a SPARQL Inferencing Notation (SPIN)[9], but SPARQL is supported by almost all triplestore. Marmotta/KiWi was chosen also because it proposes a KiWi Reasoner. It is a rule-based reasoner that can be used on top of a KiWi Triple Store. We are developing in our lab an interface for spatiotemporal data analysis "STRDFMining"[10] to exploit spatiotemporal heterogeneous data in spatial triplestore [38, 39].

References

[1] 13249-3:2002, I. (2002). *Information technology-Database languages ; SQL Multimedia and Application Packages ; Part 3: Spatial*.

[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086. IEEE Computer Society.

[3] Bannour, H. and Hudelot, C. (2011). Towards ontologies for image interpretation and annotation. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 211 –216.

[4] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135.

[5] Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Valveny, E., Jawahar, C. V., and Karatzas, D. (2019). Scene text visual question answering.

[6] Breu, H., Gil, J., Kirkpatrick, D., and Werman, M. (1995). Linear time euclidean distance transform algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(5):529–533.

[7] Clouard, R., Renouf, A., and Marinette, R. (2010). An ontology-based model for representing image processing application objectives. *International Journal of Pattern-Recognition and Artificial Intelligence*, 24:1181–1208.

[8] Cohn, A., Bennet, B., Gooday, J., and Gotts, N. (1997). Representing and reasoning with qualitative spatial relations about regions. In *Spatial and temporal reasoning*, pages 97–134. Kluwer.

[9] Coustaty, M. (2011). *Contribution à l'analyse complexe de documents anciens. Application aux lettrines*. PhD thesis, Université de La Rochelle, La Rochelle, France.

[10] Coustaty, M., Pareti, R., Vincent, N., and Ogier, J.-M. (2011). Towards historical document indexing: extraction of drop cap letters. *IJDAR*, 14(3):243–254.

[11] Coustaty, M., Tsopze, N., Bouju, A., Bertet, K., and Louis, G. (2015). Towards ontology-based retrieval of historical images. *Applied Ontology*, 10(2):147–167.

[12] Coustaty, M., Uttama, S., and Ogier, J.-M. (2012). Extraction of light and specific features for historical image indexing and matching. In *21st International Conference on Pattern Recognition, Tsukuba, Japan*, pages 1326–1329.

[13] Dubois, S., Lugiez, M., Péteri, R., and Ménard, M. (2008). Adding a noise component to a color decomposition model for improving color texture extraction. In *proc. of the 4th European Conference on Colour in Graphics, Imaging, and Vision (CGIV'08)*, pages 394–398.

[14] Fan, J., Gao, Y., and Luo, H. (2008). Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *Image Processing, IEEE Transactions on*, 17(3):407 –426.

[1] SWRL: A Semantic Web Rule Language

[2] Jena: https://jena.apache.org/

[3] DLV: http://www.dlvsystem.com/dlv/

[4] GeoSPARQL: https://www.opengeospatial.org/standards/geosparql

[5] stSPARQL: http://www.strabon.di.uoa.gr/

[6] Marmotta: http://marmotta.apache.org/

[7] SPARQL 1.1: https://www.w3.org/TR/sparql11-query/

[8] YASQE: https://yasqe.yasgui.org/

[9] SPIN: https://www.spinrdf.org/

[10] STRDFMining: https://gitlab.univ-lr.fr/abouju/STRDFMining

[15] Gallaire, H. and Minker, J., editors (1978). *Logic and Data Bases, Symposium on Logic and Data Bases, Centre d'études et de recherches de Toulouse, 1977*, Advances in Data Base Theory. Plemum Press.

[16] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43:907–928.

[17] Gudewar, A. D. and Ragha, L. R. (2012). Article: Ontology to improve cbir system. *International Journal of Computer Applications*, 52(21):23–30. Published by Foundation of Computer Science, New York, USA.

[18] Hamidi, A. E., Ménard, M., Lugiez, M., and Ghannam, C. (2010). Weighted and extended total variation for image restoration and decomposition. *Pattern Recognition*, 43(4):1564 – 1576.

[19] Hanbury, A. (2008). A survey of methods for image annotation. *Journal of Visual Languages and Computing*, 19:617–627.

[20] Inc., O. G. C. (1999). Opengis simple features specification for sql. *OpenGIs Project Document 99-049*.

[21] J., M., W., M., and A., B. (2009). Une approche ontologique pour la modélisation et le raisonnement sur les trajectoires. prise en compte des règles métiers, spatiales et temporelles. In *JFO 2009 3ème édition des journées Francofones sur les Ontologies, Poitiers,*, pages 157–168.

[22] Jena (2011). Jena - a semantic web framework for java. {http://jena.sourceforge.net/}, date visite: Mars 2011.

[23] Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 119–126, New York, NY, USA. ACM.

[24] Jimenes, R. (2008). Les bibliothèques virtuelles humanistes et l'étude du matériel typographique. Technical report, Centre d'Etude Superieur de la Renaissance.

[25] Knublauch, H., Fergerson, R., Noy, N., and Musen, M. (2004). The protégé owl plugin: An open development environment for semantic web applications. In *International Semantic Web Conference (ISWC)*, volume 3298, pages 229–243.

[26] Kompatsiaris, Yiannis; Hobson, P., editor (2008). *Semantic Multimedia and Ontologies*, volume 1 of *Theory and Applications*. Springer.

[27] Lamiroy, B. and Lopresti, D. (2011). An Open Architecture for End-to-End Document Analysis Benchmarking. In *11th International Conference on Document Analysis and Recognition - ICDAR 2011*, pages 42–47, Beijing, China. International Association for Pattern Recognition, IEEE Computer Society. ISBN: 978-1-4577-1350-7.

[28] Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A model for learning the semantics of pictures. In *NIPS'03*.

[29] Leone, N., Pfeifer, G., and Faber, W. (2005). The DLV Project - A Disjunctive Datalog System (and more). *http://www.dbai.tuwien.ac.at/proj/dlv/*.

[30] Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262 – 282.

[31] Mainz, D., Weller, K., and Mainz, J. (2008). Semantic image annotation and retrieval with iken. In Bizer, C. and Joshi, A., editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[32] Neumann, B. and Möller, R. (2008). On scene interpretation with description logics. *Image Vision Comput.*, 26:82–101.

[33] Nguyen, G., Coustaty, M., and Ogier, J. (2010). Stroke feature extraction for lettrine indexing. In *Image Processing Theory Tools and Applications (IPTA), 2010 2nd International Conference on*, pages 355 –360.

[34] Pareti, R. and Vincent, N. (2006). Ancient initial letters indexing. In *Proc. of the 18th International Conference on Pattern Recognition (ICPR'08)*, pages 756–759, Hong Kong, China. IEEE Computer Society.

[35] Pratt, W. (2007). *Digital Image Processing: PIKS Scientific Inside*. Wiley-Interscience, 4 edition.

[36] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1349–1380.

[37] Tousch, A.-M., Herbin, S., and Audibert, J.-Y. (2012). Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45:333–345.

[38] Tran, B., Bouju, A., Plumejeaud-Perreau, C., and Bretagnolle, V. (2016). Towards a semantic framework for exploiting heterogeneous environmental data. *IJMSO*, 11(3):191–205.

[39] Tran, B., Plumejeaud-Perreau, C., and Bouju, A. (2018). A web interface for exploiting spatio-temporal heterogeneous data. In *Web and Wireless Geographical Information Systems - 16th International Symposium, W2GIS 2018, A Coruña, Spain, May 21-22, 2018, Proceedings*, pages 118–129.

[40] Zhang, D., Islam, M. M., and Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition*, 45:346–362.

[41] Zhou, W., Li, H., and Tian, Q. (2017). Recent advance in content-based image retrieval: A literature survey. *CoRR*, abs/1706.06064.

[42] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J. M. R., Bradley, A., Papa, J. P., Belagiannis, V., Nascimento, J. C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A., editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham. Springer International Publishing.

[43] Zomahoun, D. E. and Yétongnon, K. (2014). EMERGSEM: emergent semantic and recommendation system for image retrieval. In *Tenth International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014, Marrakech, Morocco, November 23-27, 2014*, pages 256–263.