ISSN 2047-3338

Architecture of a Character Recognition System Based on Ontology

Aicha Eutamene[1], Mohamed Khireddine Kholladi[2] and Hacene Belhadef[3]

***Abstract*—The current revolution in the World Wide Web by using the notion of ontology has influenced in many domains of scientific research such that the domains that try bring a semantic to data and treatment; character recognition is one of these domains. In this paper we present new architecture of character recognition system that is characterized by using a domain ontology created by an expert; this ontology plays an intermediary role (bridge) between the low-level data and ones of high-level. The purpose behind this role is to bridge the semantic gap between these two types of data.**

***Index Terms*—Character Recognition, Grapheme, Ontology and Semantic Gap**

I. Introduction

ONTOLOGIES play an important role in the field of knowledge representation and sharing of information, they can be modified, adapted and reused in different applications and domains. In this work, we tried to devise a new approach to character recognition (handwritten or printed) assisted by a domain ontology. The main idea is to model the domain of character recognition (CR) by a domain ontology, whose concepts are graphemes (defined hereinafter) extracted from a segmentation step and extraction primitive, these graphemes are interconnected by spatial relationships showing their inter-location in the document and having both intrinsic and extrinsic properties describing their forms.

The use of ontology in such a process is justified by all the benefits of the latter, in order to improve the quality of results by introducing the concept of semantic and inference on facts that already exists. Our primary objective by given this proposition is to reduce and bridge the semantic gap between low-level knowledge provided by the image in the form of pixel and the high-level knowledge extracted and enriching the image. In the domain that we studied: the lower-level consists of several segments or primitives, characterized by low-level descriptors. The top level includes a generic model of knowledge "ontology" and an instantiated model of knowledge, valid for the vocabulary of the language written in the image being processed. So our approach is based on an ontological modelling of all graphemes can constitute the Latin script and spatial relationships that may exist between these graphemes to build letters of vocabulary, our studies focused primarily on the Latin printed capital-letter alphabet to test the feasibility of our approach, but it can be generalized to other forms of writing in adopting new forms of graphemes and creating new spatial relations, specific of features vocabularies may exist between concepts.

The reliability of our approach depends on a crucial step is to segment the text document to a set of graphemes, so a good segmentation is already an important step for good recognition, because it is always problematic segmentation and that is why we preferred to manual segmentation by an expert guided to maximize the percentage of expected results. After segmentation and feature extraction, we proposed a normalization step of morphological primitives that based on a classification (each class represents a type of graphemes), this classification will allow us to unify and to appoint all primitives similar under the name of single grapheme, this last going to facilitate us the instantiation of ontology concepts (a concept represents a class of primitives).

At the end, we can summarize our objectives and trends in the following points:

- Analyze handwritten and printed documents.
- Modelling of writing.
- Expressing explicitly the semantics contained in these documents by metadata.
- Create a generic process for the character recognition based and supported by an ontology.
- Apply the results obtained in the field of image analysis and restoration of cultural heritage.
- Use external resources such as WordNet to enhance the recognition process.

[1]Aicha Eutamene is PhD student at NTIC Faculty, University of Constantine 2-Algeria (aicha.eutamene@yahoo.fr)

[2]Mohamed Khireddine Kholladi, Professor at NTIC Faculty, University of Constantine 2-Algeria, President of MISC Laboratory (kholladi@yahoo.fr)

[3]Hacene Belhadef, Associate professor at NTIC Faculty, University of Constantine 2-Algeria, Member of MISC Laboratory (corresponding author email: hacene_belhadef@yahoo.fr)

II. Related Work

Ontologies have been applied in many scientific fields such as biology, ecology and hydrology. Specifically, the application of ontologies for the interpretation of digital images comes in varied areas, photographs of landscapes to remotesensing; each domain has a particular characteristics. Certain works propose to use ontologies from the segmentation step.

In [1], the proposed ontologies contain the parameters of the segmentation algorithm and labels potential regions. The description concepts of ontology are mainly based on geometric features. After an initial segmentation, segments are adjusted in order to be closer to their description in the ontology.

In the classification stage [2], the authors have conceptualized fuzzy spatial relations in ontologies to identify organs in images from the medical field. This emphasizes the importance of using spatial information in this area, as a body is largely identifiable by its shape and position. However, satellite images, spatial relationships are insufficient because they do not know a priori the provision of geographic objects in images.

The authors in [3] have constructed an ontology for satellite imagery. They developed process (semi-) automatic facilitate indexing and retrieval content of these images. They present their results on the knowledge available through automatic analysis process in terms of concepts and relations between concepts and also offer some thoughts on how to organize their storage and operation, taking into account the scope of these images.

The contribution in the thesis of [4] is based on the complex analysis of ancient documents applied to Lettrine images. The author is interested in graphic images that are as complex objects composed of different layers of information (consisting of lines and shapes). The purpose behind the use of ontologies is to identify an image looking like those in the database (search by example) and look for very specific items in drop caps, to help identify concepts in images (image decorative image composed of characters , ...).

In the article [5], they propose a new approach to semantic annotation of movements based on OWL (Ontology Web Language). This model uses the movement notation of Benesh that consists of ontology concepts of movements and their relationships to achieve their annotations, they developed semantic rules (SWRL: Semantic Web Rules Language). The results show the effectiveness of the proposed system produces a consistent set of annotations.

The authors of the article [6] present an image analysis of old documents, especially images of Lettrine, which aim to reduce the semantic gap between image processing and keywords a key area of expertise. They offer information extraction methods adapted to the case of graphic images of ancient documents, and a formal framework based on ontology representation of these knowledge's. They also propose an ontology that allows representing knowledge from the field of expertise Images: historians. The link between these two areas is then based on inference rules that create a bridge between the keywords field and low-level features.

III. Ontologies

The term ontology originates from philosophy. In that context, it is used as the name of a subfield of philosophy, namely, the study of the nature of existence (the literal translation of the Greek word Οντολογια), the branch of metaphysics concerned with identifying, in the most general terms, the kinds of things that actually exist, and how to describe them. For example, the observation that the world is made up of specific objects that can be grouped into abstract classes based on shared properties is a typical ontological commitment.

However, in more recent years, ontology has become one of the many words hijacked by computer science and given a specific technical meaning that is rather different from the original one. Instead of "ontology" we now speak of "an ontology". For our purposes, we will uses T.R.Gruber's [7] definition, later refined by R.Studer, An ontology is an explicit and formal specification of a conceptualization.

In general, an ontology describes formally a domain of discourse. Typically, an ontology consists of a finite list of terms and the relationships between these terms. The terms denote important concepts (classes et objects)of the domain. For example, in a university setting, staff members, students courses, lecture theaters, and disciplines are some important concepts.

The relationships typically include hierarchies of classes. A hierarchy specifies a class C to be a subclass of another class C if every object in C is also included in C.

Apart from subclass relationships, ontologies may include information such as:

- Properties
- Value restrictions
- Disjointness statements
- Specification of logical relationships between objects

In the context of the Web, ontologies provide a shared understanding of a do main. Such a shared understanding is necessary to overcome differences in terminology. One application's zip code may be the same as another application's area code. Another problem is that two applications may use the same term with different meanings.

Ontologies are useful for the organization and navigation of Web sites. Many Web sites today expose on the left-hand side of the page the top levels of a concept hierarchy of terms. The user may click on one of them to expand the subcategories.

Also, ontologies are useful for improving the accuracy of Web searches. The search engines can look for pages that refer to a precise concept in an ontology instead of collecting all pages in which certain, generally ambiguous, keywords occur. In this way, differences in terminology between Web pages

and the queries can be overcome.

In addition, Web searches can exploit generalization/specialization information. If a query fails to find any relevant documents, the search engine may suggest to the user a more general query. It is even conceivable for the engine to run such queries proactively to reduce the reaction time in case the

User adopts a suggestion. Or if too many answers are retrieved, the search engine may suggest to the user some specializations.

In Artificial Intelligence (AI) there is a long tradition of developing and using ontology languages. It is a foundation Semantic Web research can build upon. At present, the most important ontology languages for the Web are the following:

• XML provides a surface syntax for structured documents but imposes no semantic constraints on the meaning of these documents.

• XML Schema is a language for restricting the structure of XML documents.

• RDF is a data model for objects ("resources") and relations between them; it provides a simple semantics for this data model; and these data models can be represented in XML syntax.

• RDF Schema is a vocabulary description language for describing properties and classes of RDF resources, with a semantics for generalization hierarchies of such properties and classes.

• OWL is a richer vocabulary description language for describing properties and classes, such as relations between classes (e.g., disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g., symmetry), and enumerated classes [8].

IV. Grapheme

The grapheme is the fundamental unit of a writing data; it is the smallest unit of meaning graph whose variation changes the value of the sign in writing[1]. For ideographic scripts, it can represent a concept. In phonographic writing, it represents an element of achieving sound (syllable, consonant, and letter). So in alphabetic writing, the grapheme is commonly referred letter[2]. In our work, we show the interest of using the graphemes as features for describing the individual properties of Handwriting (Part of character).

Each grapheme is generated by segmentation of manipulated document content. At the end of this task, the document can be regarded as the concatenation of some consecutive graphemes (see example section). So, a document D will be described by a set of graphemes X_i

$$D= \{X_i / i: \text{from 1 to } n\} \quad (1)$$

A subset of successive graphemes, may construct a word W_j or a single character C_k:

$$W_j= \{X_i/ i: \text{from 1 to } m \text{ and } m<n\} \quad (2)$$

$$C_k= \{X_i/ i : \text{from 1 to } p \text{ and } p<m<n\} \quad (3)$$

[1] wikipedia

[2] http://alis.isoc.org/glossaire/grapheme.fr.htm

V. Semantic And Sensorial Gap

Visual similarity does not necessarily a semantic similarity, Example: two regions in an image of the same color does not mean they represent the same region or the same object. This difference between the conceptual level between machine, which knows only the pixelique data (bridge, line, curve, etc..) And the user who can interpret them (house, bridge, etc.) is called semantic gap. It is the recognition system to bridge this gap by proposing a high-level interpretation of low-level data.

The sensory gap is defined as "the gap between the objects in the real world and the information contained in a description (computer) derived from recording the scene." It is the projection of a reality, 3D and often continues in a 2D discrete computer representation. This gap is to be accepted by researchers working on 2D images, or repelled by researchers working on stereoscopic or 3D images (see figure 1).

The semantic gap is the most difficult to treat. For several years, researchers have revolved around this gap without actually naming it, what is done today. The semantic gap is defined as "the lack of concordance between the information that can be extracted from visual data and the interpretation of these data for a user in a given situation." This gap is more or less the same problem as linking lowlevel treatments and highlevel treatments, except that now it is clearly seen as a problem of information management and not only as a control problem [9].

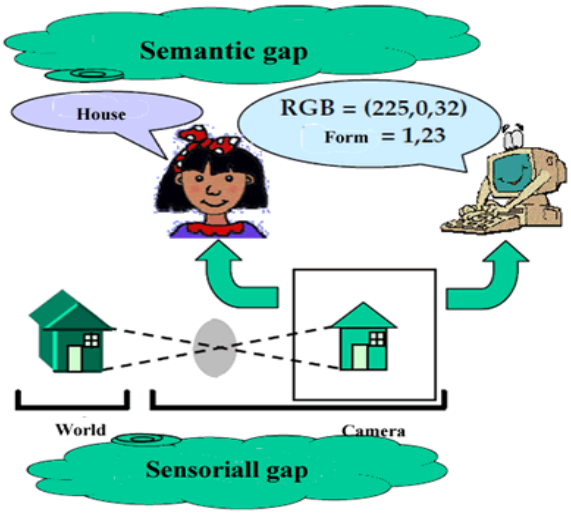

Fig. 1. The sensorial gap between the observed world and the acquired image. The semantic gap between computer analysis and human interpretation of the image [10]

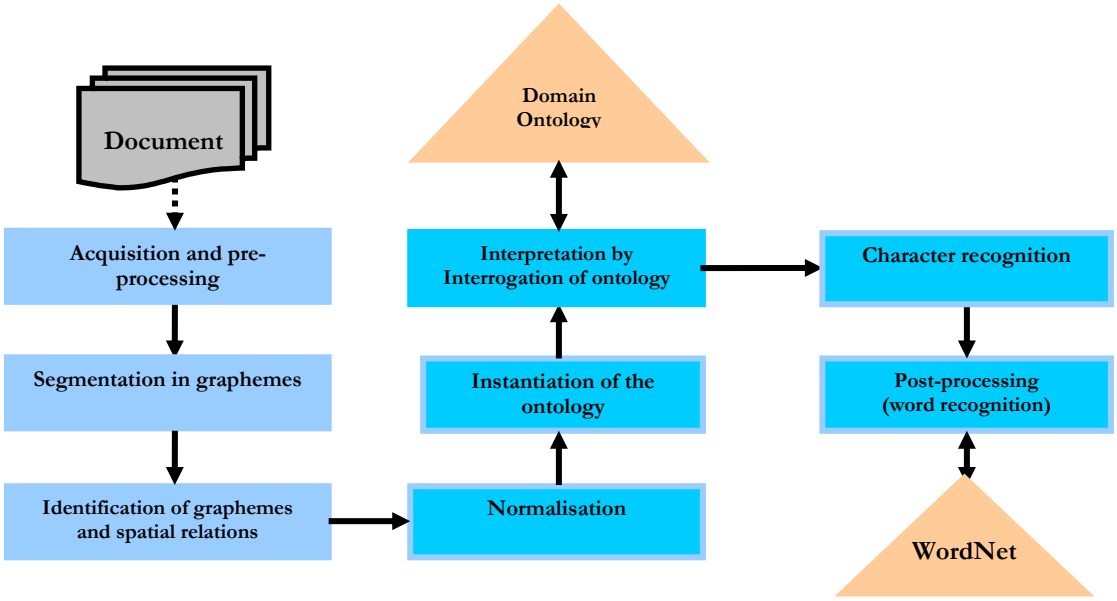

Fig.2. Architecture of our system

In recent years, the gap between semantic concepts and lowlevel numerical features retained much attention from the scientific community, and remains one of the major challenges in the field of computer vision. To reduce this semantic gap, we propose a character recognition system based on an ontology and a segmentation of the document processed into graphemes.

VI. A Proposed Architecture Of Our System

The operation of our system is based on a direct instantiation of domain ontology representing the alphabet of the written language of the processed document, in our case, we treated the Latin alphabet printed, but we can tailor and extend our ontology to support other cases (see Figure 2).

In this paper we describe an original idea presented in [11]. Our approach is based on two main stages, the analysis and interpretation, regarding the analysis of the document, we have adopted a bottom syntactic-structural approach, and for the interpretation we have exploited the spatial relations that chain the graphemes, to identify and recognize the characters. The structural analysis is to extract spatial information about segments, to name it by graphemes and also find the spatial relationships between these graphemes. The result of this analysis is a structural description of segments resulting from a segmentation step. This result will lead the stage of interpretation and thereafter will feed syntactic parsing to find an interpretation of words constructed.

A. *Build of Domain Ontology*

The primary purpose of an ontology is to model a set of knowledge in a given field, which can be real or imaginary. In our case, the domain treated is the written content in Latin (**A.. Z**) Printed image documents.

The build of ontology is an important step in the process of our system; it describes the vocabulary of processed documents and the relationships that exist between the elements of this vocabulary. Our ontology is composed of two taxonomies, taxonomy of concepts and taxonomy of spatial relations (classes hierarchy and properties hierarchy) there are two types of properties: simple data type properties and objects that represent semantic relationships between classes. This is cross these taxonomies to bring out the horizontal relationships between terms.

1) *Identification of Concepts*

The concepts of our ontology are of two types, the letters of the Latin alphabet and segments extracted by segmentation also called graphemes or spatial entity, these concepts are organized into a hierarchy of semantic objects according to the specialization relation is-a (subsumption). This taxonomy allows to discretize the space of semantic representation in large classes (see Figure 4).

2) *Identification of Relationships*

To model reality, it is not sufficient to define spatial entities. It should also define the spatial relationships between these entities. They are important, especially for knowledge representation and spatial reasoning. They are mainly used to describe the structural relationships between spatial objects.

A spatial relationship allows describing the relative positions between two symbols. For example, given the expression ***ab,*** the symbols ***a*** and ***b*** are connected by the relation left/right. While they are connected by the relation (top-right/bottom-left) in the expression $\boldsymbol{a^b}$. In most systems proposed a relationship bears more structural sense a semantic meaning (logical relation). For example, the structural relationship in $\boldsymbol{a^b}$ means mathematically that ***a*** to the power ***b***.

The types of relationships that we addressed in our study are the following relations:

• adjacency between two segments or two objects of interest or two semantic objects (distinguishing adjacencies top, bottom, right and left): "*a **Bar** is adjacent to a **Trunk** and is located to the right of it* (the case of letter '**L**'). "

• Neighbourhood between two segments or two objects of interest or two semantic objects without contact between them (distinguishing cases north, south, east, west) "*a **Point** is located north of **Trunk*** (the case of the letter '**i**'). "

B. Segmentation of the text document as a set of graphemes

A good segmentation of an expression is the key to good recognition and interpretation. The segmentation is an essential step in the recognition process. It affects strongly its robustness and determines a priori its recognition approach. We distinguish thus two approaches: global and analytical approach. To avoid the problem of segmentation of the word into letters, the overall approach does not rely on segmentation and therefore considers the trace of the word as a whole to recognize the form (the case of WordSpotting). For cons, the analytical approach we have adopted is based on the segmentation and seeks to isolate and identify meaningful units drawing a priori word corresponding to the letters composing it. So a good segmentation can introduce the system to recognize the characters in good conditions, against a bad segmentation, in turn, will lead to a fall in the rate of recognition.

The segmentation is an essential step for handwriting recognition. Whatever the source (online or offline) and nature (handwritten or printed) of signal writing; a good segmentation of the symbols is essential for proper recognition. Segmentation allows both to identify the basic elements that make up the signal, and also to introduce spatial information necessary steps in structural analysis and interpretation.

C. Extraction of Graphemes

Feature extraction is a crucial step and very important in the recognition process because subsequent treatments will no longer manipulate the original image but the results provided by this step. Its role is to locate graphemes, name and identify the spatial relationships between these different graphemes them.

Indeed, instead of cutting the writing in characters, which requires complete recognition of the text, we segment the writing script into graphical units called graphemes which generally correspond to character pieces. Figure 3, shows the four types of graphemes that can be found in a document written in Latin, according to their morphological structure and the classification of philipe Coueignoux [12] (see Figure 3).

According to the classification of philipe Coueignoux, we created taxonomy of graphemes, where each of them is

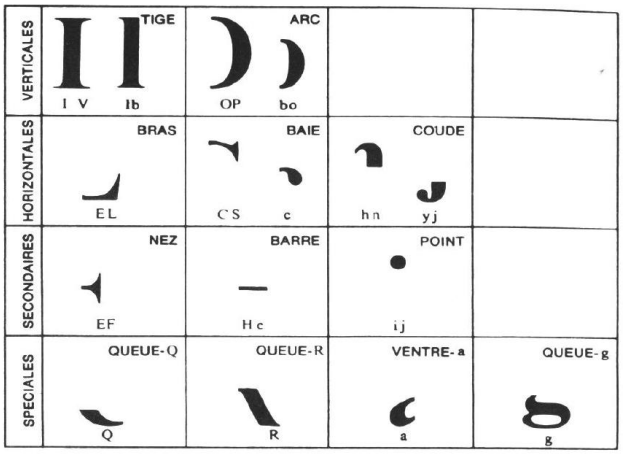

Fig. 3. Classification of graphemes by philipe Coueignoux

represented by a concept bearing his name. Figure 4 shows an excerpt of this taxonomy published under the ontologies editor: Protege2000.

The ontology of Figure 4 shows the morphological characteristics of the Latin alphabet, but we can enrich it and reuse it for other types of alphabet (Arabic, Chinese, etc.). While expanding their vocabulary by new concepts and new relationships.

D. Normalization of graphemes

As we said earlier that the type of documents processed by our system and the type printed where the characters are well formed and uniforms, but in the case of handwritten document, this constraint is not always true, that is why, we proposed a normalization step of the structure of graphemes following a classification of forms of the same semantic type. This normalization allows us to ensure a transition or transformation of handwritten script towards the printed script in order to unify the instantiation task of the ontology concepts.

E. Instantiation of the ontology

After be segmented and extract all graphemes of the document, we begin by instantiating each concept of the ontology for the corresponding grapheme and the spatial relationships identified between graphemes. Figure 5 shows some examples of segmented letters as graphemes, such as the letter "𝓛" is composed of two graphemes such as "**Tall-Trunk**" and "**Bar**", which are respectively numbered by the number 3 and 4, the relationship that connects them is " **below**". In this example, we can observe that a given grapheme can be figured in several characters, same for relationships. For example, the grapheme "**Tall-Trunk**" included in the four characters (𝓓, 𝓛, 𝓟, 𝓣) as well as the relationship "**atRight**» that fugues in the characters (𝓓 and 𝓛). So to remove this ambiguity, we have created an attribute "***number***" that indicates the position of the grapheme in the processed document, this number is sequential and unique.

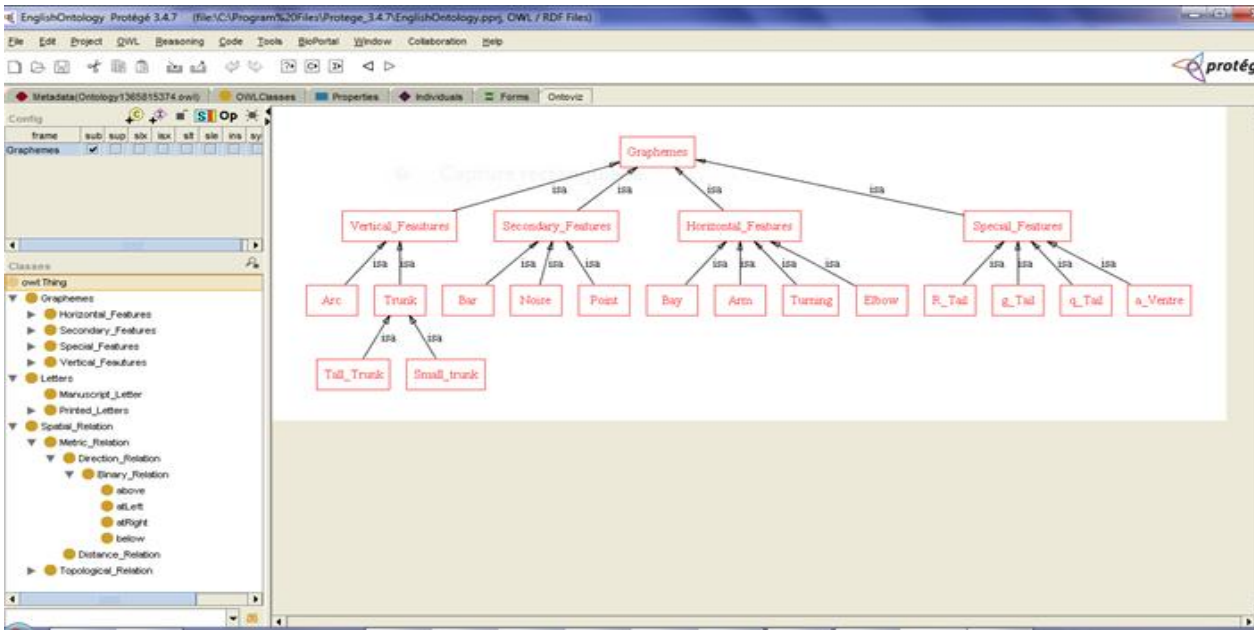

Fig. 4. Taxonomy of graphemes

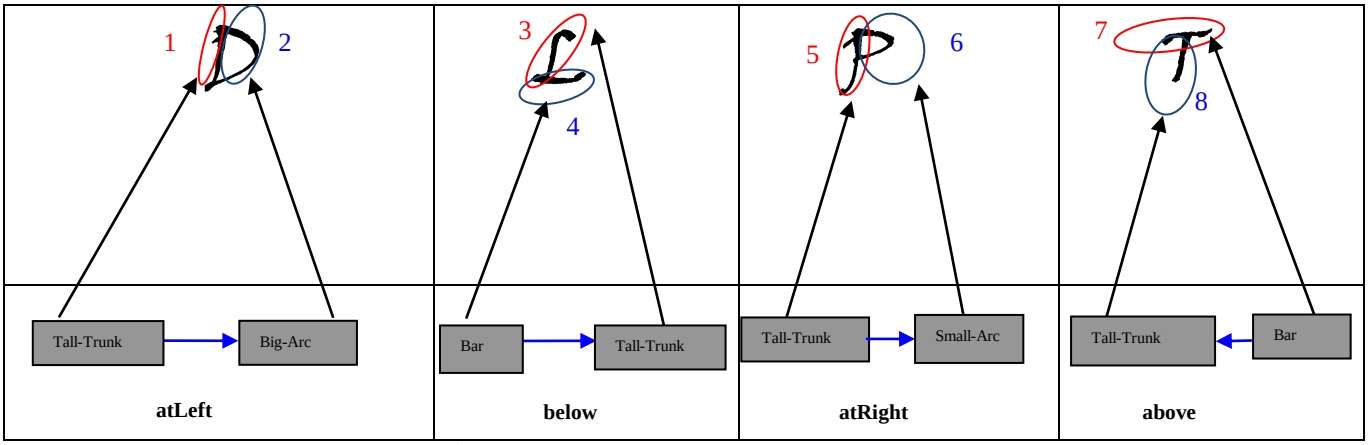

Fig. 5. Segmentation of letters in graphemes

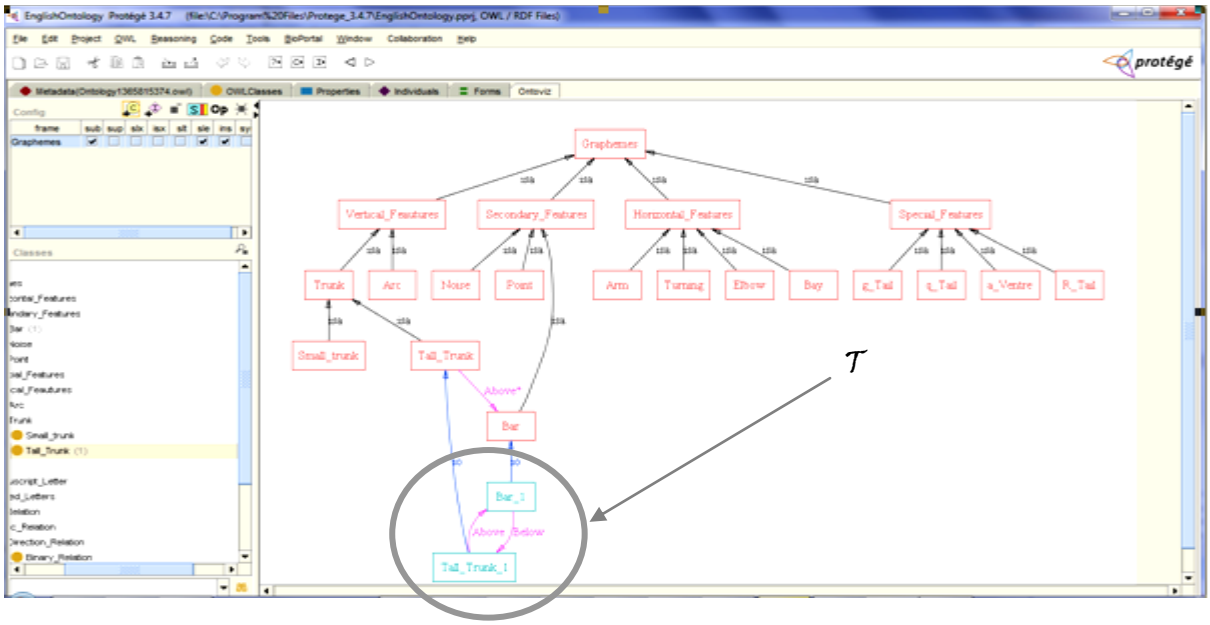

Fig. 6. Instantiation of the letter ‘ $\mathcal{T}$ ’

Figure 6, shows via the plugin Ontoviz of the editor Protege2000 an example of an instantiation of concepts "Tall-Trunc" and "Bar" and the relationship "above" to build the character 'T'. We can also observe in this figure the inverse relation “below” existing between the same concepts.

A. Word recognition and post-processing

A document is composed of a set of characters. Therefore, the recognition of these characters involves the recognition of the full text. However, often, for various reasons, the system makes a mistake or fails to identify certain characters, which makes some words invalid. The main purpose of this phase is to improve the recognition rate of words (as opposed to character recognition rate) by morphological or spelling corrections using higher levels of information (syntactic, lexical, semantic ...) using reference tools such as dictionaries, thesauri, ontologies contextual ... etc.. The identification of tokens in the document is not only the primary concern of a recognition system. The majority of the recognition methods are limited to the analysis of the physical structure and logical recognition remains to be developed, for which we propose in our system to add other steps, such as checking the syntactic and semantic content of the documents , using external resources such as WordNet and WOLF.

In the field of heritage restoration, especially ancient documents that have been preserved in non-adequate conditions, these documents may experience problems such as missing important parts or deleting words or paragraphs constituting the document. One solution of this problem is the use of external ontologies to check the semantic of syntactic and lexical units. Currently we are developing similarity measures dedicated to this type of problem.

VII. CONCLUSION

The work we have presented in this article represents a new and original vision, to improve the process of character recognition, our idea is built around a modelling based on ontology that is to describe the field of Latin script printed, and this ontology can be instantiated by the contents of the document treated after a segmentation step extracting and graphemes.

The ontology in question is composed of two taxonomies, one representing a hierarchy of concepts, or for each type of grapheme there is a concept that represents it, the second taxonomy represents the hierarchy of properties and spatial relationships that may exist between graphemes.

Our objective behind this idea is not limited to the representation of knowledge contained in the documents processed, but the exploitation of this semantic content, while taking advantage of all the benefits of the notion of ontology, including the formulation of local queries that are needed to make intelligent decisions as to the interpretation and identification of characters (ditto for words), or creating Web services that can share the solutions developed, so the ontologies in this area can be used to provide support for the automation of image analysis and effective use of modern methods and techniques of pattern recognition. Several points are also being developed for future works to improve our approach.

REFERENCES

[1] Hassan S., Hetroy F., Palombi O., ≪ Segmentation de maillages guidee par une ontologie ≫, *22èmes journées de l’association francophone d’informatique graphique*, Arles, 2009.

[2] Hudelot C., Atif J., Bloch I., ≪ Fuzzy spatial relation ontology for image interpretation ≫, 2008.

[3] Marine Campedel, Marie Lienou, Ivan Kyrgyzov, Henri Maître Vers la construction d’une ontologie appliquée à l’imagerie satellitaire**, ???**

[4] Mickaël Coustaty, thèse doctorat « Contribution à l’analyse complexe de documents anciens Application aux lettrines », 2011

[5] Sawsan Saad,Dominique De Beul, Saïd Mahmoudi, Pierre Manneback « Annotation sémantique des mouvements humains par une approche ontologique», 2011

[6] Mickaël Coustaty, Norbert Tsopze, Alain Bouju, Karell Bertet, Georges Louis, Jean-Marc Ogier **«** Ontologies et documents anciens: application aux lettrines », CIFED, Bordeaux : France 2012

[7] T.R. Gruber**,** « A Translation approach to Portable Ontology Specification Knowledge Acquisition », An International Journal of Knowledge Acquisition for Knowledge-based systems, Volume 5, no. 2, June 1993.

[8] Grigoris Antoniou and Frank van Harmelen, “A Semantic Web Primer”, TLFeBOOK, The MIT Press, Cambridge, Massachusetts, London, England, ISBN 0-262-01210-3, 2004.

[9] A.W.M. Smeulders, M. Worring, S.Santini, A.Gupta and R.Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 22, no 12, pp. 1349-1380, Dec.2000.

[10] Alain Boucher and Thi-Lan Le, Comment extraire la sémantique d’une image ?, SETIT 2005, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 27-31, 2005 – TUNISIA

[11] Hacene Belhadef and Aicha Eutamene, Ontology and character recognition : New axis of research, International journal of academic research Vol. 4. No. 2. April, 2012

[12] Ph. Coueignoux Approche structurelle de la lettre In: Langue française. N°59, 1983. pp. 45-67.