Ontology based New Approach for Character Recognition

Aicha Eutamene
University Mentouri of Constantine (Algeria)

Hacene Belhadef
University Mentouri of Constantine (Algeria)

M. K. Kholladi
University Mentouri of Constantine (Algeria)

ABSTRACT

The domain of pattern recognition and especially the character recognition is an area rich in term of scientific production, despite the abundance of work in this area but there are always anomalies. In this paper, we present a new approach for character recognition based on ontology, this last is carefully created with a domain expert, it contains a set of concepts and relationships, where each concept represents a grapheme, it is a feature extracted by an extraction module of a recognition system. Relationships between concepts are type spatial; describing the different possible relationships can be used between the graphemes, thus the characters in a document written in Latin alphabet. Our ontology is generic and can support other languages by enriching it by new specific spatial relationships.

Keywords

Characters recognition, grapheme, ontology, spatial relation, typographical features.

1. INTRODUCTION

The general goal of the recognition of documents; whether printed or manuscript, is to transform them into understandable and usable representation by machine. The process of recognition is not always easy as long as the content of documents can have multiple representations. In the case of printed documents, size, style (Bold, Italic... etc.), the cast of characters and other factors play a crucial role in this process. As for handwritten documents, the conditions of safeguarding are not often adequate. In our days, a large number of books and old manuscripts are preserved in museums and archives are in danger of disappearing due to several factors such as moisture, acidity. This requires digitization of these documents in order to preserve the heritage and exploit it more effectively. The digitization of documents is the most effective and speedy remedy, it consist to convert a paper document in the form of a digital image. Transcription is another solution but it is less used and limited to manuscripts documents not-long.

The result image of such operation of digitization is used as raw material in the recognition process, to decorticate the content and extract the necessary primitives for the identification and characters recognition, also the entire contents of the document, to use it in a lot of area such as the restoration of national heritage or world, classification, indexing and archiving.

No matter the rich content of the documents, but this wealth is insufficient to help the process of character recognition. All OCR systems interested in the content of documents is usually based on the step of classification forms but they are not interested in what's behind this, as meta-information or information semantics. In this paper we realized that a step annotation of the document is necessary to add additional information to help this process to accomplish its task. The annotation of an image through the construction of ontologies is the main tool for associating semantics to an image and allows the use of more powerful search methods and able to answer complex queries. The association between data and ontologies then allows software agents to take advantage of the knowledge represented in ontologies to better use of images.

The present paper is organized as follows. In the first part, we present an overview on the two concepts closely linked to our process, as the ontology and grapheme. In the following paragraphs, we present a description of the steps of a classical recognition process, dice digitization until the last stage which is the post-treatment, and in the third section we illustrate our contribution in such process and we end by an illustrative example to show how our approach works by using ontologies.

2. ONTOLOGY

All The term "ontology" comes from the field of philosophy that is concerned with the study of being or existence. In philosophy, one can talk about ontology as a theory of the nature of existence. In the context of computer and information sciences, ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)**. [2]**

The preceding definition leads to a set of definitions that can be used as a basis for algebraic formulation of the term Ontology and its components [1]:

Definition 1. **A term** is a triple $\tau = [\eta, \delta, A]$, $\tau \in T$, where η is a string of characters containing the name of the term, δ is a string of characters containing its definition and A is a set of attribute domains A_1, A_2, ..., A_n, each associated to a value set V_i.

Definition 2. **A relation** ϕ : $T \rightarrow T$, $\phi \in \Phi$, : is a function from T to T such that for every term $\tau_1 \in T$, there is a term $\tau_1 = \phi(\tau_1), \tau_2 \in T$.

Definition 3. **A semantic relation σ** between two terms is a relation that belongs to the set of semantic relations Σ = {Hypernymy, Hyponymy (is-a), Mereonomy (part-of), Synonymy }, $\Sigma \subset \Phi$.

Definition 4. **A spatial relation** ρ between two terms is a relation that belongs to the set of spatial relations P =

{adjacency, spatial containment, proximity, connectedness}, $P \subset \Phi$.

Definition 5. An ontology is a pair $\Theta=[T, \Phi]$, where $T=\{\tau_{1,}\tau_1, ...,\tau_n\}$ is a set of terms, and $\Phi=\{\Phi_{1,}\Phi_{2,......,}\Phi_n\}$, and $\exists \phi_i \in (\sum \cup K)$.

3. 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 produced by the segmentation module [3] of a recognition system. At the end of a segmentation step, the document can be viewed as the concatenation of some consecutive graphemes (**see example section**). The handwritten document D is thus described by the set of graphemes X_i

$$D=\{ X_i , i:1 \text{ to } n\}$$

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

$$Wj=\{X_i, i:1 \text{ to } m, m<n\}$$

$$C_k=\{X_i, i:1 \text{ to } p, p<m<n\}$$

4. CLASSICAL CHARACTER RECOGNITION PROCESS

The main steps of a recognition process are **[5]** (see figure 1):

- The acquisition enables the conversion of paper document in the form of a digital image (bitmap). This step is important because it is concerned with preparing documents to be seized, the choice and parameterization of hardware input (scanner), and the format for storing images.

-Pre-processing whose role is to prepare the document image processing. Pre-processing operations are related to the recovery of the image, remove noise and redundant information, and finally the selection of appropriate treatment areas.

- Recognition of the content that often leads to the text recognition and extraction of logical structure. These treatments usually accompanied by preparatory operations of block segmentation and classification of media (graphics, tables, pictures, etc...).

-Post-processing or correction of recognition results, to validate the digitization process. This can be done either automatically by the use of dictionaries and linguistic methods of correction, or manually through dedicated interfaces.

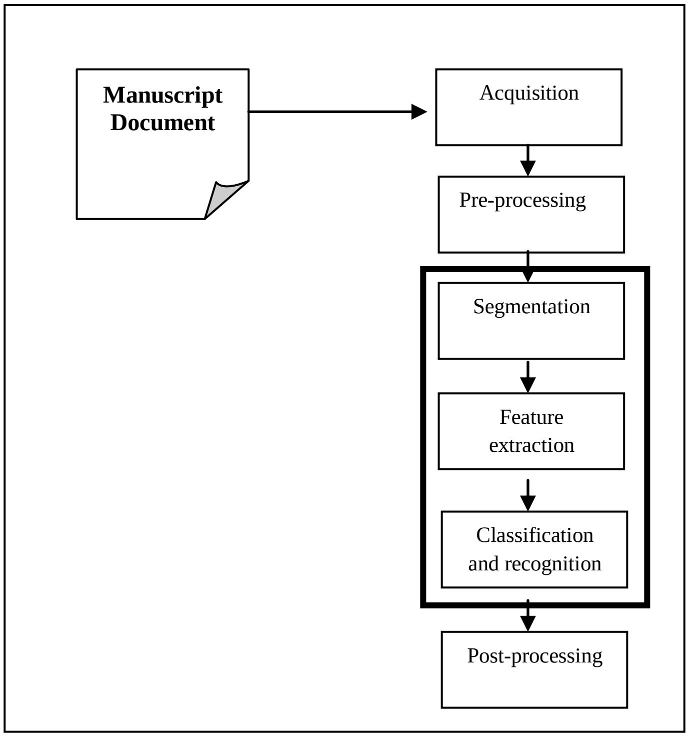


Fig 1: Architecture of classical system of character recognition

5. CHARACTER RECOGNITION APPROACHES

In this section we present four different approaches in this area that are most used.

5.1 Bayesian approach

The Bayesian approach is based essentially on the properties of the typographic character. It consists in choosing among a set of characters, one for which the following primitive has the highest probability with respect to characters previously learned.

5.2 Structural approach

Like the Bayesian approach, the structural approach is also based on the physical structure of characters. Indeed, it is to describe the nature of relations linking different topological primitives (a loop, arc, ...). These relationships can be primitive relative position compared to another type of features (vertical or horizontal), the size of a primitive compared to another, ... There are variants of the structural approach are methods based on different principles. The most commonly used are:

5.2.1 Syntactic methods

These methods use the notion of language for recognition. Each character is described or represented by a phrase in a language where the vocabulary consists of primitives. Thus, recognition of a character is whether the sentence of the representative character may be generated by the grammar.

[1] wikipedia

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

5.2.2 *The method of graphs*

This method is based on graph theory. It is to build graphs, where nodes represent the primitives and arcs, the relationships between these primitives. In the learning phase, graphs representing characters of reference are established. Which gives for each character reference or model, which includes a graph representing the different primitives describing (node) and all relations between them (arc).

5.3 STOCHASTIC APPROACH

A stochastic process implements specific probabilistic models in order to manage uncertainty and lack of information that plague the characters to recognize. Among the methods, we find the method of Freeman code, hidden Markov models for modeling of characters (strings of primitives) ... etc.

5.4 NEURONAL APPROACH

Connectionist or neuronal approach is based on the behavior and the human brain, to achieve entirely new models. His philosophy is reflected in the design of systems, called neural networks, based on a strong interconnection of a large number of elementary processors or pseudoneurones. The connections between pseudo-neurons (variable weight) and contain all knowledge.

6. OUR APPROACH

The main goal of image analysis ontology (IAO) development is formal description of knowledge on the processing, analysis, and recognition of images accumulated and used by experts **[7]**.

Our approach is based on the architecture of a new system of character recognition manuscript, which was described in [6].

Our contribution is between the step of feature extraction and recognition step, it is expressed by instantiating the ontology already created by a domain expert (see fig 2). In this figure we show just an excerpt of the global ontology representing all the Latin alphabet. In this ontology, we implemented all possible relationships, which can be found between the different graphemes that build the characters, these relationships are of spatial type (Just-below, Next-left, etc. ..). For example, the grapheme Tal-trunk can have multiple relationships with graphemes: Bar, Arc, etc.... to form respectively the characters **L** and **P**.

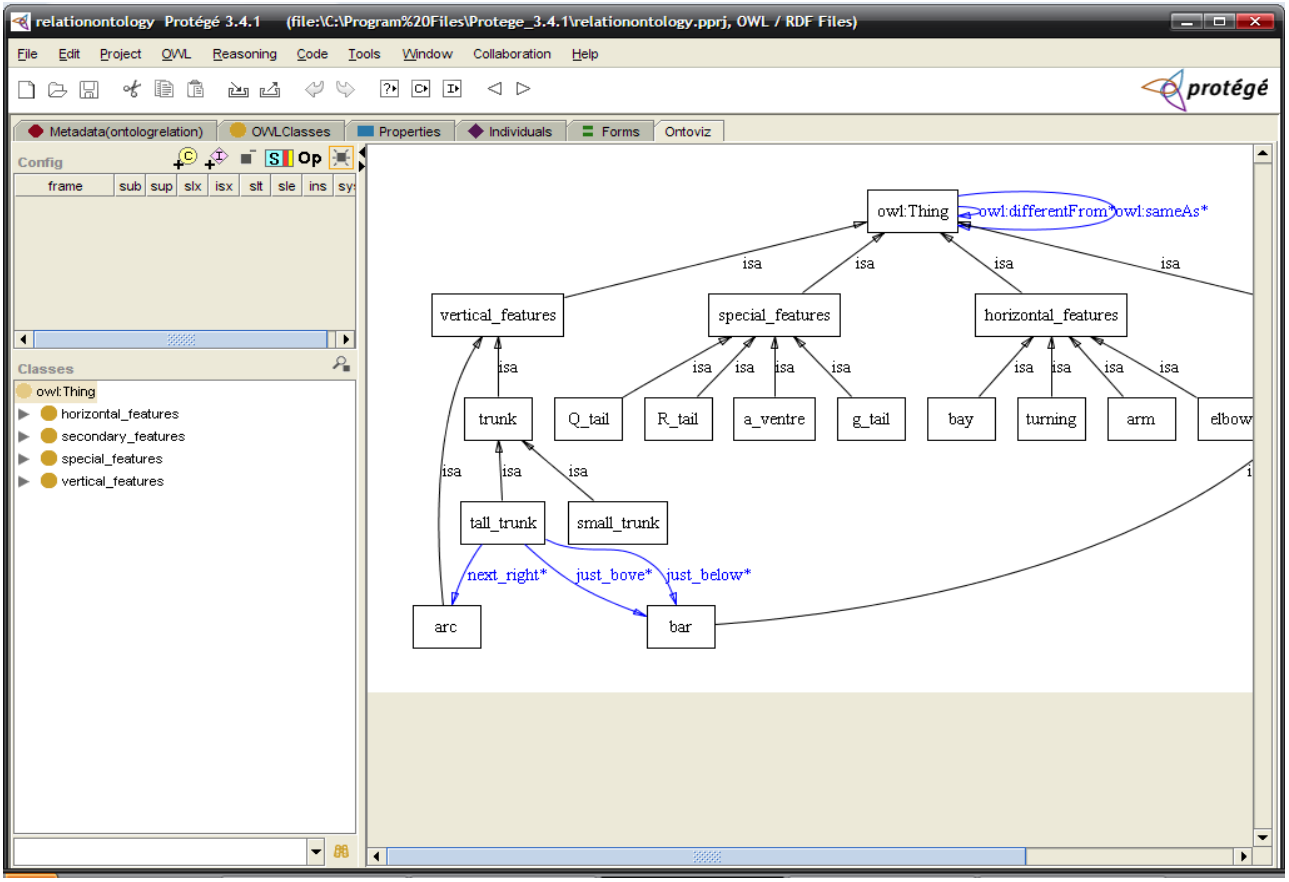


Fig 2: Ontology of the Latin alphabet.

Figure 2, presents our ontology that created under protégé editor[3] . "protégé" is the best known editor and most widely used, it is open source, developed by Stanford University, it has evolved since its first versions (Protégé-2000) to integrate from the 2003 web standards including OWL semantics. It offers many optional components: reasoners and graphical interfaces.

In the figure 3, we show an example of the different graphemes that form the word “LIRE” and specially the character “L” that is composed by two graphemes such as Tall-trunk and Bar, respectively numbered by number 1 and 2.

The relationship between the grapheme number 1 and 2, is clearly depicted in figure 4, and the corresponding code of this relationship is shown below.

RDF[4] description of data type property “ next_left “ betwen the classes: “tall_trunk" and” bar”.

```
 <rdf:Description rdf:about="#next_left">
   <rdfs:domain rdf:resource="#tall_trunk"/>
   <rdf:type rdf:resource=
"http://www.w3.org/2002/07/owl#ObjectProperty"/>
   <rdfs:range rdf:resource="#bar"/>
  </rdf:Description>
```

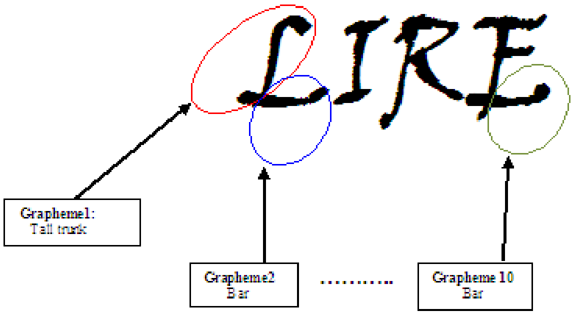
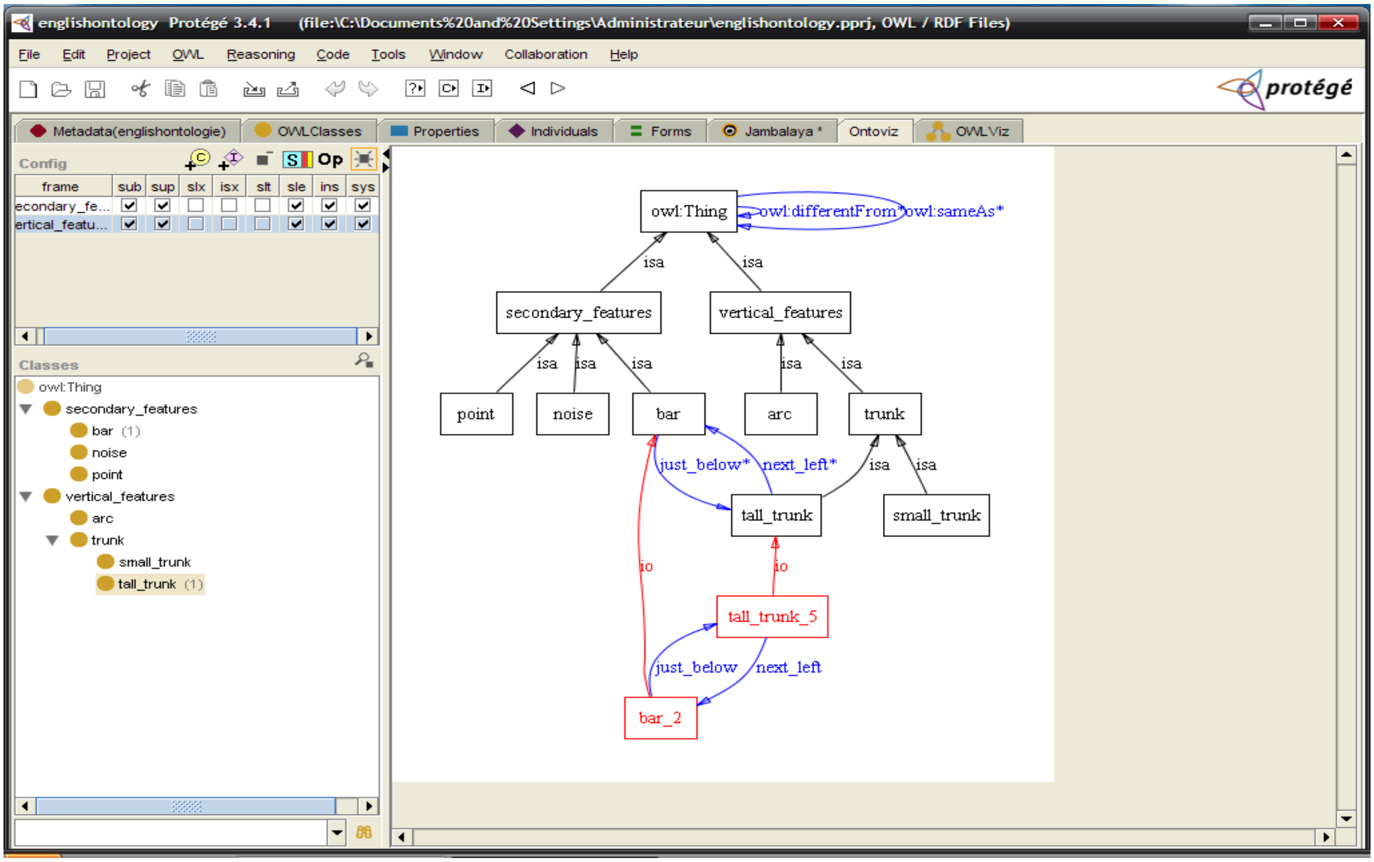


Fig 3: Segmentation of the word “LIRE” in graphemes features.


englishontology Protégé 3.4.1 (file:\C:\Documents%20and%20Settings\Administrateur\englishontology.pprj, OWL / RDF Files)
File Edit Project OWL Reasoning Code Tools Window Collaboration Help
protégé
Metadata(englishontologie) OWLClasses Properties Individuals Forms Jambalaya * Ontoviz OWLViz
Config
Op

frame	sub	sup	slx	isx	slt	sle	ins	sys
econdary_fe...	☑	☑	□	□	□	☑	☑	☑
ertical_featu...	☑	☑	□	□	□	☑	☑	☑

Classes
owl:Thing
secondary_features
bar (1)
noise
point
vertical_features
arc
trunk
small_trunk
tall_trunk (1)
owl:Thing
owl:differentFrom*owl:sameAs*
isa
isa
secondary_features
vertical_features
isa
isa
isa
isa
isa
point
noise
bar
arc
trunk
just_below*
next_left*
isa
isa
tall_trunk
small_trunk
io
io
tall_trunk_5
just_below
next_left
bar_2

Fig 4: Spatial relationship between graphemes

[3] http://protege.stanford.edu/
[4] Resource Description Framework

7. CONCLUSION

The work presented in this paper presents a new and original vision, to solve the problem of character recognition, our idea is expressed through the creation of an ontology that represents the contents of a document written manually or automatically (by seizure). This ontology transforms the image-version of document to a representation of concepts and spatial relationships. The idea here is not limited to the representation of the content but rather the exploitation of side semantic of this content, taking advantage of all the benefits of ontologies, including the formulation of local queries which are necessary for making intelligent decisions, or creating the web services that can work above. The development of ontologies in this area can be used to provide image analysis automation support and efficient use of modern methods and techniques for image analysis and pattern recognition.

In a future work, we plan to use the matching operations with external resources such as WordNet, using dedicated similarity measures.

8. REFERENCES

[1] Frederico Fonseca, Clodoveu Davis, Gilberto Câmara, 2003. Bridging Ontologies and Conceptual Schemas Geographic Information Integration. Geoinformatica, Volume 7 Issue 4, 355-378.

[2] Thomas R. Gruber. A Translation Approach to Portable Ontology Specifications. 1993. Knowledge Acquisition, 5(2):199-220.

[3] A. Bensefia and T. Paquet and L. Heutte. 2003. Information retrieval based writer identification. 7th International Conference on Document Analysis and Recognition (ICDAR2003). 946-950.

[4] Document pour l'école Jeunes Chercheurs CNRS. 1997 . INTERACTION HOMME-MACHINE, Luminy.

[5] Belaïd, Abdel. 2001. Reconnaissance automatique de l'écriture. Pour la science.

[6] Aicha Eutamene, Hacene Belhadef, M. khireddine Kholladi. 2010. Thinking about a new process of handwritten characters recognition based on an ontology. SNIB'2010, 7th National Seminar in Computer Science. Biskra-Algeria

[7] I. B. Gurevicha , O. Salvettib , and Yu. O. Trusovaa. 2009.“ Fundamental Concepts and Elements of Image Analysis Ontology, Pattern Recognition and Image Analysis. Vol. 19, No. 4. 603–611