Mixing Low-Level and Semantic Features for Image Interpretation

A Framework and a Simple Case Study

Ivan Donadello[1,2](✉) and Luciano Serafini[1]

[1] Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[2] Department of Information and Communication Technology, University of Trento,
Via Sommarive 14, 38123 Trento, Italy
{donadello,serafini}@fbk.eu

Abstract. Semantic Content-Based Image Retrieval (SCBIR) allows users to retrieve images via complex expressions of some ontological language describing a domain of interest. SCBIR adds some flexibility to the state-of-the-art methods for image retrieval, which support query either by keywords or by image examples. The price for this additional flexibility is the generation of a semantically rich description of the image content reflecting the ontology constraints. Generating these semantic interpretations is an open research problem. This paper contributes to this research line by proposing an approach for SCBIR based on the somehow natural idea that the interpretation of a picture is an (onto) logical model of an ontology that describes the domain of the picture. We implement this idea in an unsupervised method that jointly exploits the ontological constraints and the low-level features of the image. The preliminary evaluation, presented in the paper, shows promising results.

Keywords: Computer vision · Ontologies · Semantic image interpretation

1 Introduction

In recent years internet has seen a terrific increase of digital images. Thus the need of searching for images on the basis of human understandable descriptions, as in the case of textual documents, is emerging. For this reason, sites as YouTube, Facebook, Flickr, Grooveshark allow the tagging of the media and support searching by keywords and by examples. Tags associated to media constitute a simple human understandable representation of the media content. Tagging activity is very stressful and often is not well done by users. For this reason, methods for automatically generate a description of the image content, as in textual document understanding, become a real necessity. There are many approaches to image understanding which try to generate a high level description of an image by analysing low-level information (or features), such as colours, texture and contours, thus providing such a high level description in terms of

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 283–298, 2015.
DOI: 10.1007/978-3-319-16181-5_20

semantic concepts. This would allow a person to search, for instance, for an image containing "a man is riding an horse". The difficulty to find the correspondence between the low-level features and the human concepts is the main problem in content-based image retrieval. It is the so-called *semantic gap* [17]. It's widely recognised that, to understand the content of an image, contextual information (aka background knowledge) is necessary [21]. Background knowledge, relevant to the context of an image, can be expressed in terms of logical languages in an ontology [6]. Ontologies can play two main roles in image processing. First, they allow to express a set of constraints on the possible interpretations of an image and the satisfaction of such constraints can be checked via logical reasoning. Second, the terminology introduced by the ontology can be used as formal language to describe the image content. This will enable semantic image retrieval using queries expressed in the language introduced by the ontology. These two roles can be obtained by designing ontologies that formalize human understandable concepts (aka object types) and relations that can be found in the set of considered pictures (e.g., rides, part-of, nearby, is-talking-to, etc.). Furthermore, the background knowledge encoded in ontologies provides constraints on types of objects and relations, e.g. a vehicle has at least two wheels or horses can be ridden by men. The advantage of having the tags as concepts coming from a background knowledge allows to reason over the image. For example the tag "horse" enables to infer the presence of an animal.

In the present work we adopt the natural idea, envisaged in [19,23], that the interpretation of an image, in the context of an ontology, is a (partial) model of the ontology, which expresses the state of affairs of the world in the precise moment in which the picture has been taken. We propose to formalize the notion of image interpretation, w.r.t. an ontology, as *a segmented image, whose segments are associated with a set of objects of a partial model of the ontology.* To cope with the fact that a picture reports only partial information on the state of affairs we use the notion of partial model of a logical theory [30]; to cope with the possibility of having multiple alternative interpretations of a picture we introduce the notion of *most plausible partial model* of an image. The most plausible partial model for a picture is a partial model that maximizes a given scoring function, which depends from the low-level features of the image.

To have a preliminary evaluation of the above idea, we implemented this framework for a specific and limited case. We developed a fully unsupervised method to generate image interpretations able to infer the presence of complex objects from the parts present in the picture, thus inferring the relative "part-whole" structure. The method jointly exploits the constraints on the part-whole relation given by the ontology, and the low-level features of the objects available in the image. This work should be considered preliminary. Nevertheless, the evaluation shows promising results.

The paper is organized as follows. In Section 2, we present an overview on semantic image interpretation (SII). Section 3 describes our formal framework for SII. Section 4 shows how we adapt our general framework to the specific task of interpreting part-whole relation. Finally Section 5 describes the preliminary evaluation.

2 Related Work

The pure logical approach to image interpretation considers the information coming from a knowledge base for generating a semantic interpretation of an image. It is the most popular and satisfactory method. The first work that faced the problem in a logical approach is described in [23]. The authors propose a framework, based on first-order logic (FOL), for the depiction and interpretation of images. They address the image interpretation problem as finding the set of logical models of a knowledge base under the closed world assumption (CWA). The framework is presented with the example of interpreting hand drawn geographical maps, but it can be applied to other domains. The uncertainty is treated adding assertions on the specific case. A possible drawback is that an interpretation based on a total segmentation of the image using the CWA is unreasonable. This critique was described in [26] where the authors further explore the notion of logic-based approach to image interpretation. They introduce the notion of partial model for finding an image interpretation. Moreover, they propose a DL language with a calculus system for computing such a partial model. Uncertainty is not addressed. The growing interest in DL led to the first DL framework for computer vision [18]. In this work the authors investigate reasoning about spatial information in order to understand objects in a scene. The output are simple assertions on the objects and uncertainty is not handled. Following the DL-based approach, the authors of [20] explore a framework for the general high-level scene understanding task. The main interest of the work is in the conceptual structure for describing the basic components of a scene: the aggregates. An aggregate is a set of parts that compose a concept in a scene with some constraints. For example, an aggregate can be the concept of laying a table, its parts are physical objects as the table cover, actions as the transport of a dish and temporal constraints: the tablecloth has to be put before the dishes. Thus, the task of scene interpretation is the instantiation of aggregates driven by the evidence. The output of the framework is a partial model and uncertainty is not handled. This work has been extended in [19], where the authors propose a DL framework for knowledge-based high-level scene understanding. The framework remarks the necessity of a partial model and, finally, it introduces the notion of the most plausible partial model. Indeed, more interpretations can arise, so the construction of a partial model has to be guided for selecting the most probable one using a probabilistic approach. Uncertainty is not addressed. Another approach for selecting the most plausible partial model, or explanation, for a multimedia is given in [22]. Here the authors propose a DL framework for the multimedia interpretation based on abduction. The abductive reasoning [13] infers a possible explanation from a set of facts, or evidence. In this work, the evidence coming from the media analysis is the input for the abduction process that computes a plausible high-level interpretation (a partial model) of a knowledge base. The preferred explanation for the media is the one that contains more evidence and less hypotheses. This method requires a set of DL rules for defining what is abducible and uncertainty is not handled. A recent method for performing abduction, for scene understanding problem, is given by the algebraic erosion over the concept lattice of a background knowledge [4]. A survey on logical approaches to multimedia interpretation can be found in [9].

The above-mentioned works assume that the information coming from the low-level image analysis is certain and without errors. But it is possible that this information, such as the labels or the spatial relations between regions, can be incomplete, vague and contradictory. We can have regions without labels, or more weighted labels or even contradictory labels. Fuzzy DL [31] is an appropriate formalism to reason in presence of imprecision. Fuzzy DL can reduces the semantic gap as in [14] where the authors propose a fuzzy DL ontology of spatial relations. The goal is to recognize objects exploiting the spatial information extracted from the image. A fuzzy DL framework for handling the vagueness and the inconsistency of the semantic features is proposed in [7]. The presented system enriches the image with new labels taken from an ontology.

Alternative approaches rely on Gestalt theory, attribute grammars and machine learning techniques. In [32] a generic framework for scene understanding that integrates domain knowledge with Gestalt theory [28] is proposed. The framework exploits the Gestalt laws of grouping such as similarity, closure and continuty with domain knowledge to perform the semantic segmentation of images. The work described in [12] uses an attribute graph grammar and a top-down/bottom-up inference algorithm for bulding the parse tree of man-made scenes such as buildings, hallways, kitchens ect. The algorithm maximizes a Bayesian posterior probability. In [29] the authors train a recursive neural network for parsing natural scene images. They recover the intrinsic structure of the natural scene by individualizing objects and capturing part-whole and proximity relations among them. The work in [3] detects structured objects, building façades, using a hierarchical approach based on layers. Every layer detects and classifies structures in the image for the next layer that computes higher level semantic structures. Every layer selects the best interpretation of the image using an ad hoc similarity distance between graphs. Uncertainty is addressed using this similarity distance. This method is generalized in [2] using a kernel function for the graph similarity. The above methods perform the parsing of the scene starting from low-level information of the image, but the structures they build lack of a formal semantics as the logic approaches provide.

Probabilistic approaches are alternatives to fuzzy DL for handling the vagueness but also for driving the construction of the most plausible model. A well-known formalism that combines FOL knowledge bases and probabilistic graphical models in a unique representation is given by Markov Logic Networks [24]. Another significant approach is given by combining FOL with kernel machines [8].

3 Problem Formulation

We start by introducing some assumptions and definitions which constitute the basic elements of the proposed framework.

Background knowledge. We suppose that background knowledge is contained in a knowledge base expressed in a logic of the family of Description Logics (DLs) [5]. In the following we briefly introduce DL formalism. Given three disjoint sets of symbols $\Sigma = \Sigma_C \uplus \Sigma_R \uplus \Sigma_I$, denoting concepts, relations (or roles) and individuals respectively, a $\mathcal{SHIQ}$ concept is defined by the following grammar:

$$C, D := A \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid \exists R.C \mid \forall R.C \mid (\geq n)R.C \mid (\leq n)R.C$$

where $A \in \Sigma_C$, and $R \in \Sigma_R$. Furthermore, we suppose that Σ_R is closed under inverse role, i.e., if $R \in \Sigma_R$ then R^- (the inverse of R) is in Σ_R. Axioms are expressions of the following forms:

Axioms of the T-box	**Axioms of the A-box**
$C \sqsubseteq D$, concept inclusion axiom	$C(a)$, object class assertion
$R \sqsubseteq S$, role inclusion axiom	$R(a, b)$, role assertion

An interpretation $\mathcal{I}$ of the signature Σ is a pair $\langle \Delta^\mathcal{I}, \cdot^\mathcal{I} \rangle$, where $\Delta^\mathcal{I}$ is a non empty set called the interpretation domain of $\mathcal{I}$. The symbol $\cdot^\mathcal{I}$ is a function from Σ to the subsets, the relations and the elements of $\Delta^\mathcal{I}$ satisfying the following constraints: $\cdot^\mathcal{I} : \Sigma_C \longrightarrow 2^{\Delta^\mathcal{I}}$, concept names are interpreted as subsets of the domain; $\cdot^\mathcal{I} : \Sigma_R \longrightarrow 2^{\Delta^\mathcal{I} \times \Delta^\mathcal{I}}$, role names are interpreted as binary relations; and $\cdot^\mathcal{I} : \Sigma_I \longrightarrow \Delta^\mathcal{I}$, individual names are interpreted as elements of the domain. The function $\cdot^\mathcal{I}$ can be extended to all the concept expressions as follows:

$(\neg C)^\mathcal{I} = \Delta^\mathcal{I} \setminus C^\mathcal{I}$	$(C \sqcap D)^\mathcal{I} = C^\mathcal{I} \cap D^\mathcal{I}$	$(C \sqcup D)^\mathcal{I} = C^\mathcal{I} \cup D^\mathcal{I}$
$(\exists R.C)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \text{for some } (d, d') \in R^\mathcal{I},\ d' \in C^\mathcal{I}\}$		
$(\forall R.C)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \text{for all } (d, d') \in R^\mathcal{I},\ d' \in C^\mathcal{I}\}$		
$((\geq n)R.C)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \#(\{d' \in C^\mathcal{I} \mid (d, d') \in R^\mathcal{I}\}) \geq n\}$		
$((\leq n)R.C)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \#(\{d' \in C^\mathcal{I} \mid (d, d') \in R^\mathcal{I}\}) \leq n\}$		

where $\#(A)$ is the cardinality of the set A. A knowledge base $\mathcal{KB}$ is a set of axioms. $\mathcal{I}$ is a *model* of a knowledge base $\mathcal{KB}$ if it satisfies all the axioms in $\mathcal{KB}$, i.e. $\mathcal{I} \models \phi$ for all $\phi \in \mathcal{KB}$, where the satisfiability relation is defined as follows:

Axioms of the T-box	**Axioms of the A-box**
$\mathcal{I} \models C \sqsubseteq D$, iff $C^\mathcal{I} \subseteq D^\mathcal{I}$	$\mathcal{I} \models C(a)$, iff $a^\mathcal{I} \in C^\mathcal{I}$
$\mathcal{I} \models R \sqsubseteq S$, iff $R^\mathcal{I} \subseteq S^\mathcal{I}$	$\mathcal{I} \models R(a, b)$, iff $(a^\mathcal{I}, b^\mathcal{I}) \in R^\mathcal{I}$

An interpretation that satisfies $\mathcal{KB}$, namely a model of $\mathcal{KB}$, is a *complete* representation (at a certain level of abstraction) of a possible state of affairs of the real world. The knowledge base, by means of its axioms, imposes constraints on possible states. The states of affairs corresponding to interpretations that do not satisfy $\mathcal{KB}$ are considered impossible. So, for instance, the axiom $\mathsf{House} \sqsubseteq \exists \mathsf{hasPart.Door}$ imposes that the state of affairs where a house has no door will never be the case.

$$\begin{tikzcd}
\text{Real World} \arrow[r, "\text{formalized by}"] & \text{Model } \mathcal{I} \text{ of } \mathcal{KB} \\
\text{Picture} \arrow[u, "\text{partial view}"] \arrow[r, "\text{formalized by}"'] & \text{Partial Model } \mathcal{I}_p \text{ of } \mathcal{KB} \arrow[u, "\subseteq"']
\end{tikzcd}$$

Fig. 1. The world is formalized by a model of $\mathcal{KB}$ and the partial view of the world contained in the picture is formalized by a partial model

Partial models. An image is a partial view of the world. Therefore, a formal representation of the content of an image should be a *partial view* of a model of $\mathcal{KB}$. This view can be considered as an interpretation of the language of $\mathcal{KB}$, but it does not necessarily satisfy all the axioms of $\mathcal{KB}$. The intuition is represented in Figure 1. For example, in a picture we can see a car with only two wheels, the others could be not visible due to the perspective of the view. The claim that a car has four wheels is not satisfied in the picture but it is satisfied in the real world supposing to be in a normal situation. Thus, if we formalize the world as a model of our knowledge base $\mathcal{KB}$ we formalize the picture with the notion of *partial model* $\mathcal{I}_p$. A *partial model* for a knowledge base $\mathcal{KB}$ is an interpretation $\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle$ of the knowledge base, such that there is a model $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ of $\mathcal{KB}$, called *the completion of* $\mathcal{I}_p$ such that:

1. $\Delta^{\mathcal{I}_p} \subseteq \Delta^{\mathcal{I}}$
2. $a^{\mathcal{I}_p} = a^{\mathcal{I}}$ for all $a \in \Sigma_I$
3. $A^{\mathcal{I}_p} = A^{\mathcal{I}} \cap \Delta^{\mathcal{I}_p}$ for all $A \in \Sigma_C$
4. $R^{\mathcal{I}_p} = R^{\mathcal{I}} \cap \Delta^{\mathcal{I}_p} \times \Delta^{\mathcal{I}_p}$ for all $R \in \Sigma_R$.

Labelled picture. Our starting point is a segmented picture where every segment is associated with a set of labels paired with a confidence level. Labels are symbols taken from the alphabet of a knowledge base which is used to describe the real world from which the picture is taken. Given the current states of image processing software this seems a realistic assumption. We assume therefore that an image is divided into regions where every region has a set of weighted labels. Labels are taken from the signature Σ of the knowledge base. An example of labels and weights of a region is $\{(\mathsf{Duck}, 0.8), (\mathsf{DonaldDuck}, 0.7), (\mathsf{isArguingWith}, 0.4)\}$. We now provide a formal definition of labelled segment with the notion of *patch.*

A *labelled picture* $\mathcal{P}$ is a finite set of *labelled patches* $\mathcal{P} = \{p_1, \dots, p_n\}$. A *labelled patch* p is a pair $p = \langle P, L \rangle$ where:

- P is a set of adjacent pixels $(i, j) \in \mathbb{N}^2$ of the labelled image $\mathcal{P}$. The pair (i, j) is the coordinates of the pixel in the image.
- L is a set of weighted labels of the patch and it is defined as $L \subseteq \Sigma \times \mathbb{R}$.

The function $Labels : \mathcal{P} \to \Sigma$ returns the set of labels (without weights). Namely for every $p = \langle P, \{\langle l_1, w_1 \rangle, \dots, \langle l_n, w_n \rangle\} \rangle$, $Labels(p) = \{l_1, l_2, ...l_n\}$.

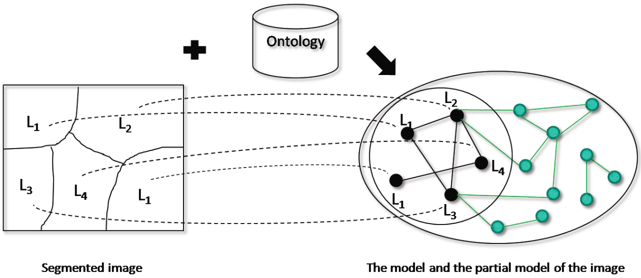


Fig. 2. Alignment between a labelled picture and its semantic interpretation

Problem definition. Following the intuition about partial models we define the semantic image interpretation as computing a partial model $\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle$ of the knowledge base. Thus, the solution is to find a method for creating the individuals (the nodes) of $\Delta^{\mathcal{I}_p}$, typing them and linking together (the arcs) according to $\cdot^{\mathcal{I}_p}$, in order to create the structured information representing the semantic content of the image. Having this graph describing the image content is not enough. We need also the information about the segmentation, e.g. in an information retrieval system it could be also necessary returning the single patches. So, we need a link between the individuals of our partial model and their corresponding segments, see Fig. 2. This consideration leads to the following formal definition of the semantic interpretation task.

Definition 1 (Semantic interpretation of a labelled image). *Given a knowledge base $\mathcal{KB}$ with signature Σ and a labelled picture $\mathcal{P}$, a* semantic interpretation of a labelled image *is a couple $(\mathcal{I}_p, cf)$ where:*

- *$\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle \subseteq \mathcal{I}$ is a partial model for $\mathcal{KB}$;*
- *$cf : \mathcal{P} \rightarrow \Delta^{\mathcal{I}_p}$ is called* conceptualization function *from the set of patches $\mathcal{P}$ to individuals, that is:*

$$\begin{aligned} cf(p) = i \in \Delta^{\mathcal{I}_p} : \exists l \in Labels(p) :& \\ i = l^{\mathcal{I}_p},\ with\ l \in \Sigma_I,& \\ i \in l^{\mathcal{I}_p},\ with\ l \in \Sigma_C,& \\ \exists j \in \Delta^{\mathcal{I}} : (i,j) \in l^{\mathcal{I}},\ with\ l \in \Sigma_R\ .& \end{aligned} \tag{1}$$

Preference relation between (partial) models. In general there are many possible explanations of the content of a picture. Formally this means that there are many partial models. On the other hand the interpretation of a picture should be unique, we have therefore to select one among a set of possible partial models. To face this problem, we introduce a scoring function $\mathcal{S}$ that assigns a score to a partial model based on its adherence to the image content, the highest the adherence the highest the score. Our problem turns to construct a partial model $\mathcal{I}_p^*$ that maximizes $\mathcal{S}$. In symbols:

$$\mathcal{I}_p^* = \operatorname*{argmax}_{\mathcal{I}_p \in \mathbb{M}_p} \mathcal{S}(\mathcal{I}_p) \tag{2}$$

where $\mathbb{M}_p$ is the set of all possible partial models. This function can not be addressed in a purely logical manner but in a statistical framework that mixes low-level features with the logical constraints between concepts (the axioms of the knowledge base). There will be the necessity of a dataset for learning the correlation between objects and relations.

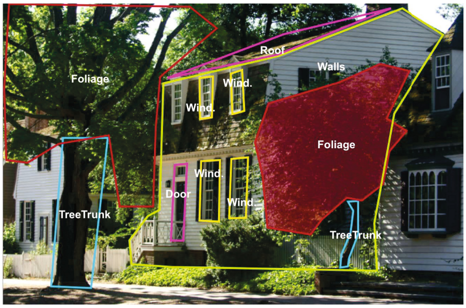
Issues in constructing an image interpretation. To construct the partial model $\mathcal{I}_p$ we have to determine its elements $\Delta^{\mathcal{I}_p}$, their types, their relations, and to search for a completion $\mathcal{I} \supseteq \mathcal{I}_p$ which satisfies all the axioms of $\mathcal{KB}$. There are several problems to face. Decide which are the elements of $\Delta^{\mathcal{I}}$ and $\Delta^{\mathcal{I}_p}$ that correspond to the picture patches, for example two regions labelled with car can be assigned to the same individual due to occlusions in the image. There can be also elements in $\Delta^{\mathcal{I}_p}$ which correspond to the composition of a set of patches. For instance, an individual of type House corresponds to the region obtained by joining the regions labelled with Window, Door, Roof, and Wall.

We also have to decide which are the types of the elements of $\Delta^{\mathcal{I}_p}$, this can be done using the labels contained in the corresponding patch as well as the axioms in the ontology. In general labels are not unique and weights need to be taken into consideration.

Another problem is to decide which are the relations between the elements of $\Delta^{\mathcal{I}_p}$. This can be achieved mixing visual and semantic features. For instance, by clustering with respect to the position of the patches, we can instantiate new individuals and linking them according to the part-whole relation. These inferences strongly depends on the type of relation we are considering.

4 Recognizing Complex Objects from their Parts

In this section we apply our framework to a specific subtask of semantic image interpretation: inferring the presence of complex objects from the presence of their parts. We consider the simplified scenario of a segmented image where patches can be labelled with at most one (non weighted) label. The background knowledge (and constraints) about part-whole relation is described by a simple ontology. Preference relation between partial models is inspired by a general principle of the mereology: the parts of the same object are topologically close in the space. Thus, we will prefer models where close parts in the image are considered parts of the same complex object. But we have to consider that sometimes close parts are not always parts of the same complex object. Therefore, to compute this preference, we need to take into account low-level features, such as the topological distance between patches, as well as semantic features, in order to prefer models that group together parts close in the space belonging to the same object. To compute the best partial model (i.e., the best grouping of parts in wholes) we use clustering techniques.



$$\begin{aligned}
\mathsf{House} &\sqsubseteq \exists \mathsf{hasParts.Door}\\
&\sqcap\ \exists \mathsf{hasParts.Window}\\
&\sqcap\ \exists \mathsf{hasParts.Roof}\\
&\sqcap\ \exists \mathsf{hasParts.Chimney}\\
&\sqcap\ \exists \mathsf{hasParts.Walls}\\
\mathsf{Tree} &\sqsubseteq \exists \mathsf{hasParts.Foliage}\\
&\sqcap\ \exists \mathsf{hasParts.TreeTrunk}\\
\mathsf{Car} &\sqsubseteq \mathsf{StreetVehicle}\\
\mathsf{Motorbike} &\sqsubseteq \mathsf{StreetVehicle}\\
\mathsf{isPartOf} &\equiv \mathsf{hasParts}^{-}
\end{aligned}$$

Fig. 3. The image of our running example. Every segment has one label among `Foliage`, `TreeTrunk`, `Window`, `Walls`, `Door`, and `Roof`. The labels are taken from a simple ontology $\mathcal{O}$. The right part shows an excerpt of it.

We explain our method via a running example. Suppose that we start from the labelled image $\mathcal{P}$ of Figure 3. The set of patches of $\mathcal{P}$ and their labels are highlighted by the segments in the figure, e.g. a patch of the image is $p = (P, (\texttt{window}, 1))$. We have manually built a simple ontology $\mathcal{O}$ containing part-whole axioms about houses and vehicles, as well as some concept inclusion axioms. An excerpt of $\mathcal{O}$ is shown on the right side of Figure 3. Despite the simplicity of this example, and the manual construction of $\mathcal{O}$, we believe that this can be highly automatized and scaled to a larger domain since there are several knowledge bases describing objects from a mereological and taxonomical point of view, e.g. Wordnet [10].

Partial Model Initialization. According to the approach described in Section 3, building a semantic image interpretation means to construct a partial model $\mathcal{I}_p$ and the conceptualization function cf. To construct $\mathcal{I}_p$ we have to create the set of individuals $\Delta^{\mathcal{I}_p}$ corresponding to the patches of the picture, assign them the correct concepts, and find relations between them. Finally, we have to check if $\mathcal{I}_p$ is a partial model for $\mathcal{O}$, i.e., if there is a completion of $\mathcal{I}_p$ that is a model for $\mathcal{O}$. This last task can be easily solved by the inference services provided by DL reasoners, such as Racer [11] or Pellet [27]. Reasoners perform the completion of an ABox: they search for a model satisfying the ontology and the statements in the ABox. Moreover, they are able to infer new knowledge from the ABox exploiting the axioms in the ontology. From this consideration it follows that the main steps for the semantic interpretation of $\mathcal{P}$ are:

- for every patch $p \in \mathcal{P}$ create a new individual $\texttt{i}_p$ in the ABox of $\mathcal{O}$;
- typing $\texttt{i}_p$ according to $Labels(p)$;
- starting the reasoner for a possible completion of the ABox.

In the specific, given a patch p we instantiate a statement as $\texttt{Concept}(\texttt{i}_\texttt{p})$ in the ABox of $\mathcal{O}$, where $\texttt{i}_p$ is a new individual and $\texttt{Concept} \in Labels(p)$.

This procedure links together two levels: the concrete level, i.e. the labelled image showing a part of the reality, and an abstract level, i.e. the mathematical entity called partial model. The procedure not only creates the partial domain $\Delta^{\mathcal{I}_p}$ but also the conceptualization function cf. In the running example the partial domain $\Delta^{\mathcal{I}_p}$ is composed by the individuals `foliage1`, `foliage2`, `treeTrunk3`, `treeTrunk4`, `window5`, `window6`, `window7`, `window8`, `walls9`, `door10`, `roof11`. Furthermore, the typing of these individuals brings to the following ABox assertions: `Foliage(foliage1)`, `Foliage(foliage2)`, `TreeTrunk(treeTrunk3)`, `TreeTrunk(treeTrunk4)`, `Window(window5)`, `Window(window6)`, `Window(window7)`, `Window(window8)`, `Walls(walls9)`, `Door(door10)`, `Roof(roof11)`. Now, if we run a reasoner on $\mathcal{O}$ with the ABox it does not raise any inconsistency, this means that there exists a model extending the ABox, thus the latter is a partial logical model of $\mathcal{O}$.

Clustering Parts for Discovering New Complex Objects. The obtained partial model is not so informative, it is necessary to fill it with part-whole relations between individuals. This means to guide the construction of a semantic interpretation of $\mathcal{P}$ towards the most plausible partial model. Such a partial model is obtained according to a general principle, the most plausible model is the one relating together parts of the same object. The idea is to group together the several parts of an object and then inferring a new individual corresponding to that object. We clustered together the several parts of the same object, so different clusters mean different objects. Then, with abductive reasoning, we provide the best explanation for every cluster, that is, the whole object underlying the presence of some parts in the cluster. This approach takes into account geometrical features of the patches and semantic features in a clustering algorithm. Indeed, we need both kind of features because some objects can be close in the Euclidean space but far from a semantic point of view and we do not want to group them together. For example, an house and a tree could be close in the picture, but they are distant in the semantics so they cannot belong to the same cluster. Moreover, two objects can have the same parts but they do not share them. For example, two different houses have as parts some windows, but they do not share them. This is the case where objects can be near in the semantics but distant in the space.

The idea is to define a joint input space for a clustering algorithm. Such a space has to embed low-level with semantic features and its elements are associated to every patch. These elements are vectors representing the joint features of the patch, specifically:

- the (x, y) coordinates of the centroids;
- the semantic distance between the concept expressed by the patch respect to the concepts expressed by other patches.

There are many methods for calculating the semantic distance between concepts, our method is based on the part-whole relations between concepts [16]. Given a patch $p \in \mathcal{P}$ let `L` its label (the concept it expresses), (x_p, y_p) the coordinates of its centroid, $\{\mathtt{L}_i\}_{i=1}^{n} \subseteq \Sigma_C$ the set of concepts expressed by the other patches, $d_{PW}(\mathtt{L}_j, \mathtt{L}_k)$ the semantic distance according part-whole relation between concepts $\mathtt{L}_j$, $\mathtt{L}_k$, the input space function $\mathcal{IS}_{PW}$ associating patches to their features according to part-whole relation is:

$$\begin{aligned} \mathcal{IS}_{PW} &: \mathcal{P} \to \mathbb{R}^{n+2} \\ &: p \mapsto \langle x_p, y_p, d_{PW}(\mathtt{L}, \mathtt{L}_1), ..., d_{PW}(\mathtt{L}, \mathtt{L}_n) \rangle \end{aligned} \tag{3}$$

Thus, our input space is the image of $\mathcal{IS}_{PW}$ over $\mathcal{P}$. In our example, an element of the input space associated to a patch p labelled with `door` has the form:

$$p \mapsto \langle x_p, y_p, d_{PW}(\mathtt{Door}, \mathtt{Walls}), d_{PW}(\mathtt{Door}, \mathtt{Foliage}), d_{PW}(\mathtt{Door}, \mathtt{Roof}), \ldots \rangle$$

With such an input space we aim at clustering together patches both close in the Euclidean space and in the semantics. In this manner we guide the construction of the partial model towards the most plausible one, i.e. the one that groups parts belonging to the same object in the image. After the clustering we have a set of clusters $\mathcal{CL} = \{cl_1, ..., cl_m\}$. In our running example the clustering algorithm (see Section 5 for details) individualized 2 clusters:

$$\begin{aligned} cl_1 &= \{\mathtt{foliage1}, \mathtt{foliage2}, \mathtt{treeTrunk3}, \mathtt{treeTrunk4}\} \\ cl_2 &= \{\mathtt{window5}, \mathtt{window6}, \mathtt{window7}, \mathtt{window8}, \mathtt{walls9}, \mathtt{door10}, \mathtt{roof11}\}. \end{aligned}$$

For the sake of presentation clarity the clusters contain the individuals corresponding to the patches and not the elements of the input space. The first cluster should group only one foliage and a trunk, the reason is these parts are too close in the Euclidean space and the unsupervised learning (as clustering) is not able to distinguish between them, see Section 5 for details.

Inferring New Individuals from Clusters. The construction of the partial model follows from the set of clusters containing parts belonging to the same object. Indeed, we need to create a new individual in the ABox corresponding to this object and typing it. Technically, we have to compute the least common concept containing the types in the cluster. More generally, we have to find the best explanation underlying a certain cluster. The reasoning that gives an explanation to some evidence is called abductive reasoning. We present a method for typing the most likely object given a cluster of its parts and an ontology. The idea is to find, for every cluster, the ontology concept whose existential concept restrictions maximize the concepts expressed by the cluster elements. This procedure is a further step towards the construction of the partial model that mostly adheres to the image.

This idea needs the following formalism to be expressed. Let us consider the axioms of $\mathcal{O}$ with the form $\mathtt{A} \sqsubseteq \bigsqcap_i \exists \mathtt{R}.\mathtt{B}_i$, where $\mathtt{B}_i \subseteq \Sigma_C$ and $\mathtt{R} \in \Sigma_R$. We call $\mathtt{B}_i$ the set of types of the existential restrictions through $\mathtt{R}$. Consequently, let $CF_R : \Sigma_C \to 2^{\Sigma_C}$, where $\mathtt{R} \in \Sigma_R$, the function that assigns to every concept $\mathtt{A} \in \Sigma_C$ the set of types of its existential restriction through $\mathtt{R}$. For example, in our ontology $CF_{\texttt{hasParts}}(\texttt{House}) = \{\texttt{Door}, \texttt{Window}, \texttt{Roof}, \texttt{Chimney}, \texttt{Walls}\}$ and $CF_{\texttt{hasParts}}(\texttt{Tree}) = \{\texttt{Foliage}, \texttt{TreeTrunk}\}$. Our approach is to compare the clusters with our ontology, thus we need to extract the concepts expressed by the parts in the clusters and a similarity measure between set of concepts. Given a cluster cl, the function CE extracts the concepts it expresses: $CE : \mathcal{CL} \to \Sigma_C$. In our running example, $CE(cl_1) = \{\texttt{Foliage}, \texttt{TreeTrunk}\}$. With this formalization it is simple to compare a cluster cl with each concept $\texttt{A}$ by defining a simple kernel set K based on the intersection between sets:

$$K(CE(cl), \texttt{A}) = \frac{\mid CE(cl) \cap CF_{\texttt{hasParts}}(\texttt{A}) \mid}{|CF_{\texttt{hasParts}}(\texttt{A})|}. \tag{4}$$

The abduction step now reduces to:

- perform the kernel set similarity between a given cluster and all the concepts $\texttt{A} \in \Sigma_C$, with $CF_{\texttt{hasParts}}(\texttt{A}) \neq \emptyset$;
- choose the concept that scores best;
- instantiate a new individual, in the ABox of $\mathcal{O}$, with that concept as type.

Thus, given cluster cl, $\texttt{A} \in \Sigma_C$ such that $CF_{\texttt{hasParts}}(\texttt{A}) \neq \emptyset$, we formalize the abductive step as instantiating a new individual $\texttt{newInd} \in \texttt{M}^{\mathcal{I}_p}$ in $\Delta^{\mathcal{I}_p}$, such that:

$$\texttt{M} = \operatorname*{argmax}_{\texttt{A} \in \Sigma_C} K(ce(cl), \texttt{A}). \tag{5}$$

This new individual represents the whole object that best explains the several parts/patches in the cluster. Moreover, the presence of this individual in $\Delta^{\mathcal{I}_p}$ improves the plausibility of the partial model. After its creation we instantiate the `hasParts` relations with the individuals corresponding to its parts. In our running example, the two new individuals after the abductive step are of type `Tree` and `House` for cl_1 and cl_2 respectively.

Remarks. Some considerations are needed. Sometimes, there is not enough semantic information (labels) to discriminate two objects, e.g. can we distinguish a car from a motorbike knowing only the concepts of `Bodywork` and `Wheel`? In this case the kernel could be the same. Objects in the real world are categorized according to a taxonomy (`isA` relation) and a general principle exists: the more general a concept is the less attributes it has. That is, more general concepts have less types of existential restrictions and thus they have a bigger kernel. For example, given the concepts of `Bodywork` and `Wheel`, the kernel with best score will not return the concepts of `Car` or `Motorbike`, but the more general one of `StreetVehicle`.

We have seen that clustering together semantic and low-level features allows to discover objects far in space and semantics, close in space but far in the semantics and vice-versa. But what about objects close in the space and in the semantics? For example, a wheel of a car could be close to the bodywork of a motorbike and the clustering algorithm clusters together the two objects. This is a still open problem, a possible solution will be to exploit further low-level features. We have a partial solution. After the abduction process of creating new individuals in the ABox we start the reasoner in order to: (i) infer knowledge about the new individuals and (ii) to check the consistency of $\mathcal{O}$ with the new assertions in the ABox. This second step allows us to discard wrong clusters. For example, if there is an axiom where cars have only one bodywork and there is a cluster with two of them with some wheels there will have inconsistency. Thus, that cluster will be discarded with the generation of a new one.

5 Experimental Results

We evaluated the task of discovering part-whole relations by defining a gold standard: given the single parts we want to discover the whole object underlying such parts. This evaluation has been achieved by constructing a small dataset of 15 labelled images where every image has been labelled using the tool LabelMe [25]; labels are taken from an ontology $\mathcal{O}$ similar to the one described in Section 4. We concentrated on two image domains: houses with trees and street vehicles, but the method is general and can be easily extended to whatever domain. We obtained our ground truth labelling the single parts composing an object, such as foliages and tree trunks, and the object itself, the tree. Moreover, we also linked the singles parts to the corresponding object according to the part-whole relation. Parts are linked together using only one level of part-whole relation, i.e. we do not have chains of parts connected by the relation.

The next step was to compare the ground truth with the output of our framework: a partial model of $\mathcal{O}$, i.e. a predicted ABox $\mathcal{A}_P$ consistent with the axioms of $\mathcal{O}$. As described in the Section 4, $\mathcal{A}_P$ contains the individuals corresponding to the parts and to the whole objects, this process has been carried out using clustering techniques. Specifically, the experiments were conducted using the Java-ML library [1] with a clustering technique based on Kohonen's Self-Organizing Maps [15]. Such a technique was the one with better performance.

$\mathcal{A}_P$ is a set of assumptions over $\mathcal{O}$, so the goal is to compare such statements with the ground truth. Thus, we converted every labelled image into an ABox $\mathcal{A}_{GT}$ with the corresponding part-whole relations instantiated. In both the ABoxes we used the same identifiers for the individual names of the single parts, while the whole objects have different individual names. This is obvious because our goal is to predict the whole objects, so we cannot use the corresponding name of the ground truth. The idea is to compare the two ABoxes by individualizing groups of parts corresponding to the same object, i.e. in `partOf` relation with it. We are not interested in the name of such an object but only on its parts. Thus, for both the ABoxes we extracted pairs of individuals corresponding to parts of the same object. For $\mathcal{A}_P$ the set of these pairs is called *positive prediction* (P), the pairs coming from $\mathcal{A}_{GT}$ are the ground truth (T) and their intersection are the true positives (TP). Table 1 shows the performance of our framework, for every image in the ground truth, in terms of precision, recall and F-measure. The mean of these metrics are, respectively, 0.89, 0.87 and 0.84.

The results show a high F-measure, and for the 46.7% of the images we generate a fully correct interpretation. Nonetheless, there are problematic cases.

Table 1. Evaluation of the framework in terms of precision, recall and F-measure

Domain	Image	$\|P\|$	$\|T\|$	$\|TP\|$	Precision	Recall	F-measure
Street Vehicles	1	18	18	18	1.00	1.00	1.00
Street Vehicles	2	42	36	26	0.62	0.72	0.67
Street Vehicles	3	14	22	14	1.00	0.64	0.78
Street Vehicles	4	8	8	8	1.00	1.00	1.00
Street Vehicles	12	32	32	32	1.00	1.00	1.00
Street Vehicles	13	4	4	4	1.00	1.00	1.00
Street Vehicles	14	4	12	4	1.00	0.33	0.50
Street Vehicles	15	12	12	12	1.00	1.00	1.00
Houses, Trees	5	242	122	122	0.50	1.00	0.67
Houses, Trees	6	62	62	62	1.00	1.00	1.00
Houses, Trees	7	56	24	24	0.43	1.00	0.60
Houses, Trees	8	54	46	46	0.85	1.00	0.92
Houses, Trees	9	40	110	40	1.00	0.36	0.53
Houses, Trees	10	68	60	60	0.88	1.00	0.94
Houses, Trees	11	12	12	12	1.00	1.00	1.00

This is due to the fact that the clustering algorithm cannot correctly group the parts of an object. In the cases of low precision (e.g. image 7) the algorithm generates less clusters w.r.t. the ground truth; in the cases of low recall (e.g. image 14) the algorithm generates more clusters w.r.t. the ground truth.

6 Conclusions

In this work we addressed the semantic image interpretation as a procedure to extract structured information from images using an ontology. A possible use of such a structure is semantically querying images about their content. The novelty of this work is a fully formalization of the problem in terms of partial logical model of the ontology based on a simple intuition: as an image is a partial view of the world it has to be formalized as a partial model. Moreover, we stated that a partial model should adhere, as much as possible, to the image, so we need a heuristic to guide its construction towards the most plausible partial model. We applied the framework to a specific subtask: the extraction of part-whole relations between objects in an image. The heuristic guiding the construction of the partial model was based on a simple principle: the parts of an object are close in the space. We implemented this idea with a clustering technique that exploits both low-level and semantic features of the image. The method was tested on a built dataset obtaining, in average, good results.

As future work we aim to find a more efficient method for discriminating objects near in the space and in the semantics. In order to better evaluate the soundness of our framework we want to extend the experiments to a larger dataset. Furthermore, we want to generalize our method to patches with more weighted labels, exploring, for example, fuzzy DL approaches. An important open problem is finding heuristics guiding the construction of plausible partial models for other relations. This can be address, for example, using supervised learning techniques o probabilistic graphical models.

References

1. Abeel, T., Van de Peer, Y., Saeys, Y.: Javaml: A machine learning library. J. Mach. Learn. Res. **10**, 931–934 (2009). http://dl.acm.org/citation.cfm?id=1577069.1577103
2. Antanas, L., Frasconi, P., Costa, F., Tuytelaars, T., Raedt, L.D.: A relational kernel-based framework for hierarchical image understanding. In: Gimel'farb, G.L., Hancock, E.R., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR/SPR. LNCS, vol. 7626, pp. 171–180. Springer, Heidelberg (2012)
3. Antanas, L., van Otterlo, M., Mogrovejo, J.O., Tuytelaars, T., Raedt, L.D.: A relational distance-based framework for hierarchical image understanding. In: Carmona, P.L., Sánchez, J.S., Fred, A.L.N. (eds.) ICPRAM (2), pp. 206–218. SciTePress (2012)
4. Atif, J., Hudelot, C., Bloch, I.: Explanatory reasoning for image understanding using formal concept analysis and description logics. IEEE Transactions on Systems, Man, and Cybernetics: Systems **44**(5), 552–570 (2014)
5. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York (2003)
6. Bannour, H., Hudelot, C.: Towards ontologies for image interpretation and annotation. In: Martinez, J.M. (ed.) 9th International Workshop on Content-Based Multimedia Indexing, CBMI 2011, June 13–15, Madrid, Spain, pp. 211–216. IEEE (2011)
7. Dasiopoulou, S., Kompatsiaris, I., Strintzis, M.G.: Applying fuzzy dls in the extraction of image semantics. J. Data Semantics **14**, 105–132 (2009)
8. Diligenti, M., Gori, M., Maggini, M., Rigutini, L.: Bridging logic and kernel machines. Machine Learning **86**(1), 57–88 (2012)
9. Espinosa, S., Kaya, A., Möller, R.: Logical formalization of multimedia interpretation. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Multimedia Information Extraction. LNCS, vol. 6050, pp. 110–133. Springer, Heidelberg (2011). http://dx.doi.org/10.1007/978-3-642-20795-2_5
10. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
11. Haarslev, V., Hidde, K., Möller, R., Wessel, M.: The racerpro knowledge representation and reasoning system. Semantic Web Journal **3**(3), 267–277 (2012)
12. Han, F., Zhu, S.C.: Bottom-up/top-down image parsing by attribute graph grammar. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1778–1785 (October 2005)
13. Hobbs, J.R., Stickel, M.E., Appelt, D.E., Martin, P.: Interpretation as abduction. Artif. Intell. **63**(1–2), 69–142 (1993). http://dx.doi.org/10.1016/0004-3702(93)90015-4
14. Hudelot, C., Atif, J., Bloch, I.: Fuzzy spatial relation ontology for image interpretation. Fuzzy Sets and Systems 159(15), 1929–1951 (2008); from Knowledge Representation to Information Processing and Management Selected papers from the French Fuzzy Days (LFA 2006). http://www.sciencedirect.com/science/article/pii/S0165011408001012
15. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
16. Liu, H., Bao, H., Xu, D.: Concept vector for semantic similarity and relatedness based on wordnet structure. J. Syst. Softw. **85**(2), 370–381 (2012). http://dx.doi.org/10.1016/j.jss.2011.08.029

17. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition **40**(1), 262–282 (2007). http://dx.doi.org/10.1016/j.patcog.2006.04.045
18. Moller, R., Neumann, B., Wessel, M.: Towards computer vision with description logics: some recent progress. In: Proceedings of the Integration of Speech and Image Understanding, pp. 101–115 (1999)
19. Neumann, B., Mller, R.: On scene interpretation with description logics. Image and Vision Computing **26**(1), 82–101 (2008) cognitive Vision-Special Issue. http://www.sciencedirect.com/science/article/pii/S0262885607001394
20. Neumann, B., Weiss, T.: Navigating through logic-based scene models for high-level scene interpretations. In: Crowley, J.L., Piater, J.H., Vincze, M., Paletta, L. (eds.) ICVS 2003. LNCS, vol. 2626, pp. 212–222. Springer, Heidelberg (2003). http://dl.acm.org/citation.cfm?id=1765473.1765497
21. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Sciences **11**(12), 520–527 (2007)
22. Peraldi, I.S.E., Kaya, A., Möller, R.: Formalizing multimedia interpretation based on abduction over description logic aboxes. In: Grau, B.C., Horrocks, I., Motik, B., Sattler, U. (eds.) Description Logics. CEUR Workshop Proceedings, vol. 477. CEUR-WS.org. (2009)
23. Reiter, R., Mackworth, A.K.: A logical framework for depiction and image interpretation. Artificial Intelligence **41**(2), 125–155 (1989)
24. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning **62**(1–2), 107–136 (2006)
25. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. Int. J. Comput. Vision **77**(1–3), 157–173 (2008). http://dx.doi.org/10.1007/s11263-007-0090-8
26. Schroder, C., Neumann, B.: On the logics of image interpretation: model-construction in a formal knowledge-representation framework. In: Proceedings. of the Int. Conf. on Image Processing, vol. 1, pp. 785–788 (September 1996)
27. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. Web Semant. **5**(2), 51–53 (2007). http://dx.doi.org/10.1016/j.websem.2007.03.004
28. Smith, B., von Ehrenfels, C., Verlag, P.: Foundations of Gestalt theory. Philosophia Verlag Munich, Germany (1988)
29. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 129–136 (2011)
30. Staruch, B., Staruch, B.: First order theories for partial models. Studia Logica **80**(1), 105–120 (2005)
31. Straccia, U.: Reasoning within fuzzy description logics. J. Artif. Intell. Res. (JAIR) **14**, 137–166 (2001)
32. Zlatoff, N., Tellez, B., Baskurt, A.: Image understanding and scene models: a generic framework integrating domain knowledge and gestalt theory. In: International Conference on Image Processing, ICIP 2004, vol. 4, pp. 2355–2358 (October 2004)