Multimed Tools Appl (2014) 72:2107–2141
DOI 10.1007/s11042-013-1491-z

Building and using fuzzy multimedia ontologies for semantic image annotation

Hichem Bannour · Céline Hudelot

Published online: 26 May 2013

Abstract This paper proposes a methodology for building fuzzy multimedia ontologies dedicated to image annotation. The built ontology incorporates visual, conceptual, contextual and spatial knowledge about image concepts in order to model image semantics in an effective way. Indeed, our approach uses visual and conceptual information to build a semantic hierarchy that will serve as a backbone of our ontology. Contextual and spatial information about image concepts are then computed and incorporated in the ontology in order to model richer semantic relationships between these concepts. Fuzzy description logics are used as a formalism to represent our ontology and the inherent uncertainty and imprecision of this kind of information. Subsequently, we propose a new approach for image annotation based on hierarchical image classification and a multi-stage reasoning framework for reasoning about the consistency of the produced annotation. In this approach, fuzzy ontological reasoning is used in order to achieve a semantically relevant decision on the belonging of a given image to the set of concepts from the annotation vocabulary. An empirical evaluation of our approach on Pascal VOC'2009 and Pascal VOC'2010 datasets has shown a significant improvement on the average precision results.

Keywords Image annotation · Multimedia ontology · Ontology building · Ontological reasoning · Fuzzy DL · Spatial information · Contextual information

1 Introduction

Automatic image annotation is a challenging problem dealing with the textual description of images. This process usually consists in the building of a computational

H. Bannour (✉) · C. Hudelot
MAS Laboratory, Ecole Centrale Paris, 92 295 Chatenay-Malabry, France
e-mail: hichem.bannour@ecp.fr

C. Hudelot
e-mail: celine.hudelot@ecp.fr

model that enables to associate a text description (often reduced to a set of semantic keywords) to digital images. A wide number of approaches have been proposed to address this concern and to narrow the well-known *semantic gap* problem [35]. Most approaches rely on machine learning techniques to provide a mapping function that allows classifying images in semantic classes using their visual features [5, 9, 27]. However, these approaches face the scalability problem when dealing with broad content image databases [30], i.e. their performances decrease significantly when the concept number is high and depend on the targeted datasets as well [21]. This variability may be explained by the huge intra-concept variability and the wide inter-concept similarities on their visual properties that often lead to conflicted and incoherent annotations. Yet, more and more concept classes are introduced for annotating multimedia content in order to enrich the description of images and to satisfy user expectations in an image retrieval system. Consequently, current techniques are struggling to scale up, and the only use of machine learning seems to be insufficient to solve the image annotation problem. Firstly, because of the lack of a reliable computational model that allows to model the correlation between the low-level features of images and the semantic concepts. Secondly, because it seems that there is a lack of coincidence between the high-level concepts and the low-level features, and that image semantics is not always correlated with the visual appearance. Therefore other alternatives need to be explored in order to improve existing approaches. In particular, some recent work proposed to use explicit semantic structures, such as semantic hierarchies and ontologies, to improve the image annotation [2, 12, 17, 43].

Indeed, ontologies defined as a formal, explicit specification of a shared conceptualization [19] have shown to be very useful to narrow the semantic gap. They allow identifying, in a formal way, the dependency relationships between the different concepts and therefore provide a valuable information source for many problems. Moreover, ontological reasoning can also be used to formulate image annotation and interpretation tasks. For instance, in [12] the authors proposed a framework for the extraction of enhanced image descriptions based on an initial set of graded annotations generated through generic image analysis techniques. Explicit semantics, represented by ontologies, have also been intensely used in the field of image and video indexing and retrieval [2, 26]. In most of these approaches, only the descriptive part of ontologies is used as a common multi-level language to describe image content [34], or more recently as semantic concept networks to refine image annotation [17, 43], or to perform image classification [3, 32].

In this paper, we propose to go deeper in the use of ontologies for image annotation. Our objective is twofold. We first propose an approach to automatically build a fuzzy multimedia ontology dedicated to image annotation. Indeed, given a training database consisting of pairs of image/textual annotation, our approach allows to automatically build an ontology representative of the image semantics by mining these images and their annotations. Thereafter, we propose a generic approach for image annotation combining both machine learning techniques such as hierarchical classification and fuzzy ontology reasoning. The rest of this paper is structured as follows. In Section 2, we review some related work. Section 3 presents an overview of the proposed approach for multimedia ontologies building. Section 4 introduces the proposed formalism for our multimedia ontology and the set of axioms and inferences rules allowing to perform the reasoning tasks. In Section 5, we introduce the proposed method for building multimedia ontologies suitable for reasoning about image annotation and interpretation. Section 6 introduces the proposed multi-stage reasoning framework for image annotation. Section 7 reports the experimental results obtained on the Pascal VOC dataset. A discussion about the proposed approach and the usefulness of our ontology for computer vision tasks is presented in Section 8. The paper is concluded in Section 9.

2 Related work

Despite significant progress shown by statistical approaches for images annotation, the semantic gap problem is still an open issue for image annotation. In this context, several recent approaches have proposed to improve this task by the use of explicit knowledge models. A first category of approaches have proposed to use semantic hierarchies for image annotation and classification [3, 17, 32, 42]. Bannour et al. [3] have identified three types of hierarchies used for image annotation: (1) language-based hierarchies: based on textual information (ex. tags, surrounding context, WordNet, Wikipedia, etc.) [14, 32], (2) visual hierarchies: based on low-level image features [6, 18, 46], and (3) semantic hierarchies: based on both textual and visual features [3, 17, 29]. However, most of these approaches use semantic hierarchies to reduce the complexity of the classification problem or as a framework for hierarchical image classification and they do not use the semantic structure of these hierarchies (i.e. the inherent semantic relationships of concepts within these hierarchies). Consequently, only a limited improvement in the classification results was shown by these approaches.

Other approaches proposed to use multimedia ontologies in order to define a standard for the description of low-level multimedia content [13, 33], or to use it as a semantic repository for storing knowledge about image domain [34], or to allow semantic interpretation and reasoning over the extracted descriptions [12, 22, 24]. Indeed, ontologies allow to model many important semantic relations between concepts which are missing in the semantic hierarchy models, as for instance the contextual and the spatial relationships. These relations have been proved to be of prime importance for image annotation [22, 24, 25, 40]. The reasoning power of ontological models has also been used for semantic image interpretation. In [12, 24, 25], formal models of domain application knowledge are used through fuzzy description logics to help and to guide the semantic image analysis.

However, much remains to be done in order to achieve more expressive ontologies of images semantics. Firstly, almost all existing approaches for building multimedia ontologies start from an existing specification of a domain (defined by an expert or inferred from a generic commonsense ontology). These specifications are not always relevant for modeling image semantics and are often incomplete, subjective and subject to many inconsistencies. Indeed, many assumptions about the concepts, their properties and relationships must be done in order to achieve a given specification, which finally do not hold in the real world. Secondly, most recent approaches for building multimedia ontologies are based either on a conceptual specification, or a visual one. Consequently, these approaches do not accurately model images semantics. Furthermore, many of these approaches are limited to provide a formalism allowing to use ontologies as a repository for storing knowledge about multimedia content. However, since these approaches have not addressed the problem of reasoning about this knowledge, the effectiveness of stored knowledge has to be proved. Finally, ontology modeling in description logics is not an intuitive task. The representation of each single real world object is split into many axioms about concepts and roles, leading to an overall design that is very difficult to apprehend [36]. This makes the design of a well-defined ontology by humans a big challenge, with no guarantee of success (scalability problem of ontology building).

Our approach goes further than the aforementioned ones and allows answering many of the previously stated limitations. Specifically, we propose in this paper a methodology for building multimedia ontologies as knowledge bases that contain explicit and structured knowledge about image context. To ensure that the structure of our ontology is representative of the image semantics, we propose to use a *semantico-visual* specification (which incorporates the visual and conceptual semantics of image concepts) for designing our ontology. In addition, we propose to build our multimedia ontology in an automatic manner and based on mining image databases to gather valuable information about image context. Thereby, we reduce the scalability problem of ontology building and we ensure that the depicted knowledge is faithful to image semantics. Finally, the proposed ontology is built using a highly expressive formalism (*Fuzzy OWL2-DL*), which allows a good interaction with it, i.e. a good querying and reasoning capabilities. Our belief is that such formal ontology will allow performing reasoning tasks in order to achieve an effective decision-making to provide a semantically consistent image annotation.

3 Overview of our approach for building multimedia ontologies dedicated to image annotation

This paper proposes an approach for building a fuzzy multimedia ontology dedicated to image annotation. As illustrated in Fig. 1, our ontology incorporates several types of knowledge about image context in order to achieve a relevant representation of image semantics. Moreover, this knowledge is automatically extracted from a

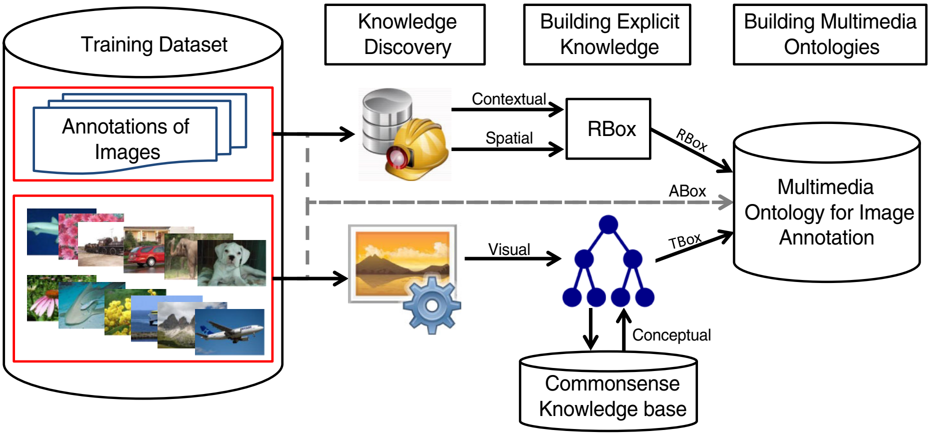

Training Dataset
Annotations of Images
Knowledge Discovery
Building Explicit Knowledge
Building Multimedia Ontologies
Contextual
Spatial
RBox
RBox
ABox
Visual
TBox
Multimedia Ontology for Image Annotation
Conceptual
Commonsense Knowledge base

Fig. 1 From image data to structured knowledge models: architecture of our approach for building multimedia ontologies dedicated to image annotation

training image database using data mining techniques. Therefore, assuming that the considered training dataset is enough representative of current image databases, our approach allows for building multimedia ontologies faithful to the image semantics.

Figure 1 depicts the workflow of our approach. As shown in this figure, the knowledge discovery process is performed through the following steps:

1. Processing the set of images in the training dataset to discover useful knowledge about the image domain (i.e. perceptual semantics), such as the visual similarity between concepts.
2. Mining the image annotations (provided in the metadata) to gather useful information about images context, namely contextual and spatial knowledge about image concepts.
3. Query a commonsense knowledge base to gather precise information about the semantics of image concepts, and in order to link the initial concepts to their hypernyms using the method proposed in [3].

Thereafter, the building of our multimedia ontology is fully automatically performed, i.e. without any human intervention. This is achieved by converting the previously extracted information about image context into explicit knowledge using the formalism described in Section 4.

Problem formalization
Given:

- $\mathcal{DB}$, a training image database consisting of a set of pairs $\langle$*image/textual annotation*$\rangle$, i.e. $\mathcal{DB} = \{[i_1, \mathscr{A}_1], [i_2, \mathscr{A}_2], \cdots, [i_{\mathcal{L}}, \mathscr{A}_{\mathcal{L}}]\}$, where:
  - $\mathfrak{I} = \langle i_1, i_2, \cdots, i_{\mathcal{L}} \rangle$ is the set of all images in $\mathcal{DB}$,
  - $\mathcal{L}$ is the number of images in the database.
  - $\mathcal{C} = \langle c_1, c_2, \cdots, c_{\mathcal{N}} \rangle$ is the annotation vocabulary used for annotating images in $\mathfrak{I}$,
  - $\mathcal{N}$ is the size of the annotation vocabulary.
  - $\mathscr{A}_i$ is a textual annotation consisting of:
    - the set of concepts $\{c_j \in \mathcal{C},\ j = 1..n_{i_i}\}$ associated with a given image $i_i \in \mathcal{DB}$,
    - the spatial location of each concept c_j in the image i_i given by its minimum bounding box defined as $(c_{j_{x\min}}, c_{j_{y\min}}, c_{j_{x\max}}, c_{j_{y\max}})$, where $c_{j_{x\min}}$ and $c_{j_{y\min}}$ are the coordinates of the low left corner of the bounding box (and respectively $c_{j_{x\max}}$ and $c_{j_{y\max}}$ are the coordinates of the upper right corner of the bounding box).
- $\mathcal{CO}$, a generic commonsense ontology containing $\mathcal{N}'$ concepts ($\mathscr{C}$), such that $\mathcal{C} \subseteq \mathscr{C}$. In this paper, we used WordNet as a commonsense ontology.

Our objective is to build a multimedia ontology, consisting of a set of $|\mathcal{C}| + |\mathcal{C}'|$ concepts (*s.t.* $\mathcal{C} \cup \mathcal{C}' \subseteq \mathscr{C}$, and $\mathcal{C}'$ could be probably the empty set), dedicated to this specific annotation problem, i.e. dependent on the initial annotation vocabulary but which could be extended at any time later. This ontology should not only incorporate the subsumption relationships between the different concepts, but also richer semantic relations, such as contextual and spatial relationships. The overall goal is to extend the use of this ontology to previously unseen images (i.e. $\forall\, i_x \notin \mathcal{DB}$)

in order to reason on the consistency of their annotations and to provide them a relevant textual description.

The design of our multimedia ontology as a well defined formal knowledge base is achieved through the following main steps, which are detailed in the remaining of this paper:

- ⋆ Definition of the DL formalism of the proposed ontology, i.e. the expressiveness of the ontology.
- ⋆ Definition of the set of axioms and inferences rules allowing to perform the reasoning tasks on the proposed ontology.
- ⋆ Definition of the main concepts of the ontology.
- ⋆ Definition of the *RBox*, i.e. definition of the key roles (relationships between concepts) and their properties.
- ⋆ Definition of the *TBox*, i.e. definition of the subsumption hierarchy, and consequently the subsumption relationships between the ontology concepts.
- ⋆ Definition of the *ABox*, i.e. the instances of concepts and the relations between them with respect to the roles defined in the *RBox*.

4 Formalism of our multimedia ontology

4.1 Preliminaries

The Web Ontology Language (OWL) is the current standard language for representing ontologies. It allows describing a domain in terms of: concepts (or classes), roles (or properties), individuals and axioms. Concepts ($\boldsymbol{C}$) are a set of objects, individuals ($\boldsymbol{I}$) are instances of concepts in $\boldsymbol{C}$, roles are binary relationships between individuals in $\boldsymbol{I}$, whereas axioms describe how these concepts, individuals, roles, etc. should be interpreted. Three sublanguages of OWL can be used: *OWL-Full* which is the most expressive language but reasoning within it is undecidable, *OWL-Lite* which has the lowest complexity but fewer constructs, and *OWL-DL* which has a good balance/trade-off between expressiveness and reasoning complexity [8].

In our approach, in order to ensure a high expressiveness with a decidable reasoning for our ontology, we used *OWL 2 DL* as a language for designing our ontology. Indeed, *OWL 2 DL* is more expressive than *OWL-DL*, i.e. includes more axioms. Concretely, we have implemented a framework using the *OWL API*[1] [23], which supports *OWL 2* since it last version. The reasoning tasks about concepts, roles and individuals are also performed using our framework, which is based on the *FaCT++* reasoner and extending it with the axioms illustrated in Table 1 to support the *Fuzzy Description Logics* (Fuzzy DL). Initially, *FaCT++* supports the $\mathcal{SROIQ}(D)$ logic (i.e. the DL for *OWL2* ontology). However, our framework supports the fuzzy logic f-$\mathcal{SROIQ}(D)$ thanks to the extension we have made.

Description Logics (DLs) are a family of logics for representing structured knowledge. Fuzzy DLs extend classical DLs by allowing to deal with fuzzy/imprecise concepts [38]. Indeed, in fuzzy logics a statement is no longer true or false, but is changed in a fuzzy statement signifying that it has a degree of truth $\alpha \in [0, 1]$.

[1] http://owlapi.sourceforge.net/index.html

Fuzzy set preliminaries In a formal way, let X be a set of elements. A fuzzy set A over a countable crisp set X is characterized by a membership function $\mu_A : X \rightarrow [0,1]$ (or $A(x) \in [0,1]$), assigning a membership degree $A(x)$ to each element x in X. $A(x)$ gives an estimation of the belonging of x to A. In fuzzy logics, the membership degree $A(x)$ is regarded as the degree of truth of the statement *"x is A"*. Accordingly, a concept C is interpreted in fuzzy DL as a fuzzy set, and thus concepts become imprecise. For instance, the statement $a : C$ (a is an instance of concept C) will have a truth-value in [0,1] given by its membership degree denoted $C^{\mathcal{I}}(a)$. A fuzzy relation R over two countable crisp sets X and Y is a function $R : X \times Y \rightarrow [0,1]$. R is *reflexive* iff for all $x \in X$, $R(x,x) = 1$ holds, while R is *symmetric* iff for all $x, y \in X$, $R(x,y) = R(y,x)$ holds. R is said *functional* iff R is a partial function $R : X \times Y \rightarrow \{0,1\}$ such that for each $x \in X$ there is a unique $y \in X$ where $R(x,y)$ is defined.

4.2 Expressiveness of our ontology

As aforementioned, for the sake of providing a highly expressive multimedia ontology with a decidable reasoning, we used the fuzzy DL f-$\mathcal{SROIQ}(D)$ for designing our ontology. Based on the work of [37, 39], we introduce in the following the specific formalism (constructors and axioms) used for defining our multimedia ontology.

The f-$\mathcal{SROIQ}(D)$ is a fuzzy extension of the $\mathcal{SROIQ}(D)$ DL, which provide both a set of constructors allowing the construction of new concepts and roles. The f-$\mathcal{SROIQ}(D)$ includes $\mathcal{ALC}$ standard constructors (i.e. negation $\neg$, conjunction $\sqcap$, disjunction $\sqcup$, full existential quantification $\exists$, and value restriction $\forall$) extended with transitive roles ($\mathcal{S}$), complex role axioms ($\mathcal{R}$), nominals ($\mathcal{O}$), inverse roles ($\mathcal{I}$), and qualified number restrictions ($\mathcal{Q}$). ($\mathcal{D}$) indicates support for (fuzzy) concrete domains, i.e. datatype properties, data values or data types.

Fuzzy concrete domain A fuzzy concrete domain is a pair $\langle \Delta_D, \Phi_D \rangle$, where Δ_D is an interpretation domain and Φ_D is the set of fuzzy domain predicates d with a predefined arity n and an interpretation $d^D : \Delta_D^n \rightarrow [0,1]$ [41].

In f-$\mathcal{SROIQ}(D)$, concepts (denoted C or D) and roles (R) can be built inductively from atomic concepts (A), atomic roles (R_A), top concept $\top$, bottom concept $\bot$, named individuals (o_i), simple roles S, and universal role U. Simple roles S are inductively defined: (i) R_A is simple if it does not occur on the right side of a Role Inclusion Axioms (RIA), (ii) R^- is simple if R is, (iii) if R occurs on the right side of a RIA, R is simple if, for each $\langle w \sqsubseteq R \rhd \alpha \rangle$, $w = S$ for a simple role S.

Fuzzy concepts Under f-$\mathcal{SROIQ}(D)$, a fuzzy concept is defined by the following assertions:[2]

$$\begin{aligned} C \rightarrow\ & \top \mid \bot \mid A \mid C_1 \sqcap C_2 \mid C_1 \sqcup C_2 \mid \neg C \mid \exists R.C \mid \exists T.d \mid \forall R.C \mid \forall T.d \mid \\ & (\geq m\ S.C) \mid (\geq m\ T.d) \mid (\leq n\ S.C) \mid (\leq n\ T.d) \mid \{o_1, \ldots, o_n\} \\ D \rightarrow\ & d \mid \neg d \end{aligned}$$

[2] n, m are natural numbers, such that $n \geq 0, m > 0$. d is an unary fuzzy domain predicate.

Table 1 Syntax and semantics of the Fuzzy Description Logic $f\text{-}\mathcal{SROIQ}(D)$ used for designing our multimedia ontology

		Syntax	Semantics
C	Constructor		
1	Atomic concept	A	$A^{\mathcal{I}}(a) \in [0, 1]$
2	Top	$\top$	$\top^{\mathcal{I}}(a) = 1$
3	Bottom	$\bot$	$\bot^{\mathcal{I}}(a) = 0$
4	Conjunction	$C \sqcap D$	$(C \sqcap D)^{\mathcal{I}}(a) = C^{\mathcal{I}}(a) \otimes D^{\mathcal{I}}(a)$
5	Disjunction	$C \sqcup D$	$C \sqcup D^{\mathcal{I}}(a) = C^{\mathcal{I}}(a) \oplus D^{\mathcal{I}}(a)$
6	Negation	$\neg C$	$(\neg C)^{\mathcal{I}}(a) = \ominus C^{\mathcal{I}}(a)$
7	Existential restriction	$\exists R.C$	$(\exists R.C)^{\mathcal{I}}(a) = sup_{b \in \Delta^{\mathcal{I}}}\{R^{\mathcal{I}}(a, b) \otimes C^{\mathcal{I}}(b)\}$
8		$\exists T.d$	$(\exists T.d)^{\mathcal{I}}(a) = sup_{v \in \Delta_D}\{T^{\mathcal{I}}(a, v) \otimes d_D(v)\}$
9	Universal restriction	$\forall R.C$	$(\forall R.C)^{\mathcal{I}}(a) = inf_{b \in \Delta^{\mathcal{I}}}\{R^{\mathcal{I}}(a, b) \rightarrow C^{\mathcal{I}}(b)\}$
10		$\forall T.d$	$(\forall T.d)^{\mathcal{I}}(a) = inf_{v \in \Delta_D}\{T^{\mathcal{I}}(a, v) \rightarrow d_D(v)\}$
11	At-least restriction	$\geq\ m\ S.C$	$(\geq\ m\ S.C)^{\mathcal{I}}(a) = sup_{b1,\ldots b_m \in \Delta^{\mathcal{I}}}((\otimes_{i=1}^{m}\{S^{\mathcal{I}}(a, b_i)$ $\otimes C^{\mathcal{I}}(b_i)\}) \otimes (\otimes_{j<k}\{b_j \neq b_k\}))$
12		$\geq\ m\ T.d$	$(\geq\ m\ T.d)^{\mathcal{I}}(a) = sup_{v1,\ldots v_m \in \Delta_D}((\otimes_{i=1}^{m}\{T^{\mathcal{I}}(a, v_i)$ $\otimes d_D(v_i)\}) \otimes (\otimes_{j<k}\{v_j \neq v_k\}))$
13	At-most restriction	$\leq n\ S.C$	$(\leq n\ S.C)^{\mathcal{I}}(a) = inf_{b1,\ldots b_{n+1} \in \Delta^{\mathcal{I}}}((\otimes_{i=1}^{n+1}\{S^{\mathcal{I}}(a, b_i)$ $\otimes C^{\mathcal{I}}(b_i)\} \rightarrow (\oplus_{j<k}\{b_j = b_k\}))$
14		$\leq n\ T.d$	$(\leq n\ T.d)^{\mathcal{I}}(a) = inf_{v1,\ldots v_{n+1} \in \Delta_D}((\otimes_{i=1}^{n+1}\{T^{\mathcal{I}}(a, v_i)$ $\otimes d_D(v_i)\} \rightarrow (\oplus_{j<k}\{v_j = v_k\}))$
15	Local reflexivity	$\exists S.Self$	$(\exists S.Self)^{\mathcal{I}}(a) = S^{\mathcal{I}}(a, a)$
16	Fuzzy nominals	$\bigcup_{i=i}^{m}\{(o_i, \alpha_i)\}$	$\{(o_1, \alpha_1), \ldots, (o_m, \alpha_m)\}^{\mathcal{I}}(a) = sup_{i \mid a \in \{o_i^{\mathcal{I}}\}^{\alpha_i}}$
17	Atomic role	R_A	$R_A^{\mathcal{I}}(a, b) \in [0, 1]$
18	Universal role	U	$U^{\mathcal{I}}(a, b) = 1$
19	Inverse role	R^-	$\forall a, b \in \Delta^{\mathcal{I}}, (R^-)^{\mathcal{I}}(a, b) = R^{\mathcal{I}}(b, a)$
20	Concrete role	T	$T^{\mathcal{I}}(a, v) \in [0, 1]$
A	Axiom		
1	Concept assertion	$\langle a : C \bowtie \alpha \rangle$	$C^{\mathcal{I}}(a^{\mathcal{I}}) \bowtie \alpha$
2	Role assertion	$\langle (a : b) : R \bowtie \alpha \rangle$	$R^{\mathcal{I}}(a^{\mathcal{I}}, b^{\mathcal{I}}) \bowtie \alpha$
3	Concrete role assertion	$\langle (a : b) : T \bowtie \alpha \rangle$	$T^{\mathcal{I}}(a^{\mathcal{I}}, v_D) \bowtie \alpha$
4	Equality assertion	$\langle a = b \rangle$	$a^{\mathcal{I}} = b^{\mathcal{I}}$
5	Inequality assertion	$\langle a \neq b \rangle$	$a^{\mathcal{I}} \neq b^{\mathcal{I}}$
6	Subsumption	$\langle C \sqsubseteq D \rhd \alpha \rangle$	$inf_{a \in \Delta^{\mathcal{I}}}\{C^{\mathcal{I}}(a) \rightarrow D^{\mathcal{I}}(a)\} \rhd \alpha$
7	Concept definition	$\langle C \equiv D \rangle$	$\forall a \in \Delta^{\mathcal{I}}, C^{\mathcal{I}}(a) = D^{\mathcal{I}}(a)$
8	Role inclusion axioms	$\langle R_1 R_2 \cdots R_n$ $\sqsubseteq R \rhd \alpha \rangle$	$sup_{b_1 \ldots b_{n+1} \in \Delta^{\mathcal{I}}} \otimes [R_1^{\mathcal{I}}(b_1, b_2), \ldots,$ $R_n^{\mathcal{I}}(b_n, b_{n+1})] \rightarrow R^{\mathcal{I}}(b_1, b_{n+1}) \rhd \alpha$
9	Disjoint role	$dis(S_1, S_2)$	$\forall a, b \in \Delta^{\mathcal{I}}, S_1^{\mathcal{I}}(a, b) \otimes S_2^{\mathcal{I}}(a, b) = 0$
10	Symmetric role	$sym(R)$	$\forall a, b \in \Delta^{\mathcal{I}}, R^{\mathcal{I}}(a, b) = R^{\mathcal{I}}(b, a)$
11	Reflexive role	$ref(R)$	$\forall a \in \Delta^{\mathcal{I}}, R^{\mathcal{I}}(a, a) = 1$
12	Transitive role	$trans(R)$	$\forall a, b \in \Delta^{\mathcal{I}}, R^{\mathcal{I}}(a, b) \geq sup_{c \in \Delta^{\mathcal{I}}} R^{\mathcal{I}}(a, c)$ $\otimes R^{\mathcal{I}}(c, b)$
13	Irreflexive role	$irr(S)$	$\forall a \in \Delta^{\mathcal{I}}, S^{\mathcal{I}}(a, a) = 0$
14	Asymmetric role	$asy(S)$	$\forall a, b \in \Delta^{\mathcal{I}}$, if $S^{\mathcal{I}}(a, b) > 0$ then $S^{\mathcal{I}}(b, a) = 0$

$a, b \in \Delta^{\mathcal{I}}$ are abstract individuals, $v \in \Delta_D$ is a concrete individual, n, m are natural numbers ($n \geq 0, m > 0$), $\alpha \in [0, 1]$ is the truth degree of a statement, $\rhd \in \{>, \geq\}$, $\bowtie \in \{>, <, \geq, \leq\}$

For more details about the semantics of these assertions cf. Table 1, constructors C1–C16.

Fuzzy $\mathcal{KB}$ A f-$\mathcal{SROIQ}(D)$ knowledge base (denoted $\mathcal{KB}$) is a triple ($\mathcal{T}$,$\mathcal{R}$,$\mathcal{A}$) where $\mathcal{T}$ is a fuzzy Terminological Box (*TBox*), $\mathcal{R}$ is a regular fuzzy Role Box (*RBox*), and $\mathcal{A}$ is a fuzzy Assertional Box (*ABox*) containing statements about individuals. The *TBox* and *RBox* contain general knowledge about the domain application.

Fuzzy ABox The fuzzy *ABox* consists of a finite set of fuzzy concept and fuzzy role assertion axioms. Typically, these assertions include: concept assertion ($\langle a : C \bowtie \alpha\rangle$), role assertion ($\langle (a : b) : R \bowtie \alpha\rangle$), concrete role assertion ($\langle (a : b) : T \bowtie \alpha\rangle$), equality assertion ($\langle a = b\rangle$), and inequality assertion ($\langle a \neq b\rangle$). The semantics of these assertions is defined in Table 1, axioms A1–A5.

Fuzzy TBox The fuzzy *TBox* is a finite set of General Concept Inclusions (GCI) constrained with a truth-value and of the form $\langle C \sqsubseteq D \rhd \alpha\rangle$ between two f-$\mathcal{SROIQ}(D)$ concepts C and D. Concept equivalence $\langle C \equiv D\rangle$ can be captured by two inclusions $C \sqsubseteq D$ and $D \sqsubseteq C$. These assertions and their semantics are defined in Table 1, axioms A6 and A7.

Fuzzy RBox The fuzzy *RBox* consists of a finite set of role axioms which are illustrated in Table 1, axioms A8–A14. These include: role inclusion axioms, disjoint role, symmetric role, reflexive role, transitive role, irreflexive role, and asymmetric role.

Owing to the specific motivations discussed in Section 4.3, we have defined the fuzzy operators used in Table 1 as follows:

1. product t-norm: $a \otimes b = a * b$.
2. product t-conorm: $a \oplus b = a + b - a * b$.
3. Łukasiewicz negation: $\ominus\alpha = 1 - \alpha$.
4. Gödel implication (for GCIs and RIAs): $\alpha \rightarrow \beta = 1$ if $\alpha \leq \beta$, β otherwise.
5. KD implication (for other constructors): $\alpha \rightarrow \beta = \max(1 - \alpha, \beta)$.

Fuzzy interpretation The Semantics of the f-$\mathcal{SROIQ}(D)$ DL is defined in terms of *fuzzy interpretations* [38]. A fuzzy interpretation is a pair $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ where $\Delta^{\mathcal{I}}$ is a non-empty set of objects (called the domain) and $\cdot^{\mathcal{I}}$ is a fuzzy interpretation function, which maps:

- a concept name C onto a function $C^{\mathcal{I}} : \Delta^{\mathcal{I}} \rightarrow [0, 1]$,
- a role name R onto a function $R^{\mathcal{I}} : \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \rightarrow [0, 1]$,
- an individual name a onto an element $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$,
- a concrete individual v onto an element $v_D \in \Delta_D$,
- a concrete role T onto a function $T^{\mathcal{I}} : \Delta^{\mathcal{I}} \times \Delta_D \rightarrow [0, 1]$,
- a concrete feature t onto a partial function $t^{\mathcal{I}} : \Delta^{\mathcal{I}} \times \Delta_D \rightarrow \{0, 1\}$

Satisfiability Finally, a fuzzy interpretation $\mathcal{I}$ satisfies an f-$\mathcal{SROIQ}(D)$ knowledge base $\mathcal{KB} = (\mathcal{T},\mathcal{R},\mathcal{A})$ if it satisfies all axioms of $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{A}$. $\mathcal{I}$ is then called a model of $\mathcal{KB}$, written: $\mathcal{I} \models \mathcal{KB}$.

4.3 Ontology-based reasoning

General automatic reasoning tasks on ontologies include concept consistency, concept subsumption to build inferred concepts taxonomy, instance classification and retrieval, parent and children concept determination, and answering queries over ontology classes and instances [1]. These reasoning tasks are induced by inferring logical consequences from a set of asserted facts or axioms.

Logical consequence A fuzzy axiom τ is a logical consequence of a knowledge base $\mathcal{KB}$, denoted $\mathcal{KB} \models \tau$ *iff* every witnessed model of $\mathcal{KB}$ satisfies τ.

Given a $\mathcal{KB}$ and an axiom τ of the form $\langle C \sqsubseteq D \rangle$, $\langle a : C \rangle$ or $\langle (a, b) : R \rangle$, it is possible to compute the best explanation of a given statement (probably, about an image) as the τ's *best entailment degree* (bed). The *bed* problem can be solved by determining the *greatest lower bound* (glb) [38].

Greatest lower bound The greatest lower bound of τ with respect to a fuzzy $\mathcal{KB}$ is:

$$glb(\mathcal{KB}, \tau) = \sup\{n \mid \mathcal{KB} \models \langle \tau \geq n \rangle\}, \quad where\ \sup \emptyset = 0 \tag{1}$$

Example 1 (Greatest lower bound) For instance, given $\mathcal{KB} = \{\langle (a, b) : R, 0.5 \rangle, \langle b : C, 0.9 \rangle\}$, the greatest lower bound that a is an instance of a concept which is in relation R with concept C is:

$$glb(\mathcal{KB}, a : \exists R.C) = 0.45$$

Best satisfiability degree The *best satisfiability degree* (bsd) of a concept C with respect to a fuzzy $\mathcal{KB}$ is defined as:

$$bsd(\mathcal{KB}, C) = \sup_{\mathcal{I} \models \mathcal{KB}} \sup_{x \in \Delta^{\mathcal{I}}} \left\{ C^{\mathcal{I}}(x) \right\} \tag{2}$$

The *best satisfiability degree* consists in determining the maximal degree of truth that the concept C may have over all individuals $x \in \Delta^{\mathcal{I}}$, among all models $\mathcal{I}$ of the $\mathcal{KB}$.

According to our specific context, and in order to achieve an efficient reasoning (and subsequently an accurate decision) on the best explanation of a given image, it is important to compute a membership degree for this explanation which reflects the likelihood of conjunction of all independent events composing it. The product logic makes possible to dispose of this desirable property for the t-norm. This assumption has motivated our choice for the product t-norm and the product t-conorm as fuzzy operators of our ontology—cf. Section 4.2. For instance, let us consider the following example where we want to compute the membership of an image i to the class *BeachImage*:

Example 2 (Product semantics and Zadeh semantics)

$$\mathcal{KB} = \{\langle i : Image, 1 \rangle, \langle i : \exists depicts.Sea, \alpha_1 \rangle, \langle i : \exists depicts.Sand, \alpha_2 \rangle,$$
$$\langle i : \exists depicts.Sky, \alpha_3 \rangle\}$$

$$BeachImage \equiv Image \sqcap \exists depicts.Sea \sqcap \exists depicts.Sand \sqcap \exists depicts.Sky$$

$$glb\,(\mathcal{KB}, i : BeachImage) = \alpha_1 \otimes \alpha_2 \otimes \alpha_3$$
$$= \begin{cases} \min\{\alpha_1, \alpha_2, \alpha_3\} & \text{under Zadeh semantics} \\ \alpha_1 * \alpha_2 * \alpha_3 & \text{under Product semantics} \end{cases}$$

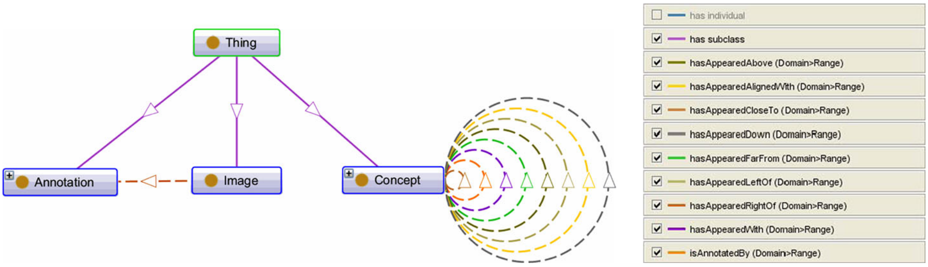
Both explanations and membership degrees are meaningful with respect to a given application. However, according to our target application, the product semantics allows to dispose of a more significant membership value than the one produced by Zadeh semantics. For example, let us suppose that α_1, α_2, and α_3 are produced as a result of an image classification process, or an object detection one. Therefore, it would be more accurate to compute the membership degree of the image i to the class *BeachImage* as the product of the confidence values of these classifiers than as the minimum score of these classifiers. This property is reachable by the use of product semantics.

5 Building of our multimedia ontology

5.1 Main concepts of our ontology

Proposed concepts The proposed multimedia ontology relies mainly on the four following concepts, which can recursively involve similar concepts (Fig. 2a):

- "Thing" represents the top concept ($\top$) of the ontology,
- "Concept" is the generic concept in our ontology to represent a concept from the annotation vocabulary, i.e. any concept $c_j \in \mathcal{C} \cup \mathcal{C}'$ used to describe the content of an image.
- "Image" is the generic concept to represent an image, i.e. each image i_i of the database will be considered as an instance of the concept "Image" with a satisfiability degree of 1 ($\langle i_i : Image, 1\rangle$).
- "Annotation" is a generic concept introduced to represent a given annotation, i.e. a set of concepts as a whole. We will come back on this notion later.


Thing
Annotation
Image
Concept
has individual
has subclass
hasAppearedAbove (Domain>Range)
hasAppearedAlignedWith (Domain>Range)
hasAppearedCloseTo (Domain>Range)
hasAppearedDown (Domain>Range)
hasAppearedFarFrom (Domain>Range)
hasAppearedLeftOf (Domain>Range)
hasAppearedRightOf (Domain>Range)
hasAppearedWith (Domain>Range)
isAnnotatedBy (Domain>Range)

(a) (b)

Fig. 2 Illustration of the used roles for defining concept relationships in our ontology. Figure **a** illustrates the main concepts of our ontology and the used fuzzy roles (in *dashed arrows*) for defining the relationships between concepts. Figure **b** illustrates the roles names

5.2 Definition of the *RBox*

As stated previously, our intent is to design an ontology of spatial and contextual information dedicated to reasoning about the consistency of image annotation. According to this aim, we define in Table 2 the proposed roles and their properties, which constitute the *RBox* of our multimedia ontology. These roles can be categorized as contextual relationships and spatial relationships, and are detailed respectively in Section 5.4.1 and in Section 5.4.2. The choice of these specific roles is motivated by the reasoning scenarios designed to improve the image annotation task. However, these roles can be further enriched depending on referred applications.

5.3 Building the semantic hierarchy and definition of the *TBox*

The subsumption hierarchy (and respectively the subsumption relationships) is a fundamental component of ontologies. It acts as a backbone of the produced ontology, where the subsumption roles allow defining the inheritance of properties from the parent (subsuming) concepts to the child (subsumed) concepts. Thus, any statement that is true (with an α degree) for a parent concept is also necessarily true (with at least an α degree) for all of its subsumed concepts. Furthermore, these subsumption relationships allow defining the *Terminological Box* of ontologies.

In our approach, we propose to automatically build a subsumption hierarchy where leaf nodes are the initial concepts of the considered dataset ($c_j \in \mathcal{C}$), and mid-level nodes are the concepts discovered by a variant of the approach proposed in [3]. Indeed, in order to design a representative ontology of the image semantics, we propose in this paper to automatically build the semantic hierarchy using a

Table 2 Roles and functional roles used for defining concept relationships in our ontology

	Domain	Range	Symetric	Reflexive	Functional	Inverse
Role name						
isAnnotatedBy	Image	Annotation	No	No	No	–
hasAppearedWith	Concept	Concept	Yes	Yes	No	–
hasAppearedAbove	Concept	Concept	No	No	No	hasAppearedBelow
hasAppearedBelow	Concept	Concept	No	No	No	hasAppearedAbove
hasAppearedLeftOf	Concept	Concept	No	No	No	hasAppearedRightOf
hasAppearedRightOf	Concept	Concept	No	No	No	hasAppearedLeftOf
hasAppearedAlignedWith	Concept	Concept	Yes	No	No	–
hasAppearedCloseTo	Concept	Concept	Yes	No	No	–
hasAppearedFarFrom	Concept	Concept	Yes	No	No	–
Functional role name						
hasFrequency	Concept	Float	–	–	Yes	–
hasAppearedAlone	Concept	Float	–	–	Yes	–

Semantico-Visual similarity computed between image concepts. The used *Semantico-Visual* similarity incorporates:

(i) a *visual similarity* which represents the visual distance between concepts, and
(ii) a *conceptual similarity* which defines a relatedness measure between target concepts based on their definitions in WordNet.

Afterwards, the building of the subsumption hierarchy is *bottom-up*, and is based on a set of heuristic rules in order to link together the concepts that are semantically most related w.r.t the previously computed similarity. Consequently, the building of the subsumption hierarchy consists in identifying $|\mathcal{C}'|$ new concepts that link all the concepts of $\mathcal{C}$ in a hierarchical structure that best represents image semantics. For more information about these (visual and conceptual) similarities and the used rules for linking concepts together, the reader is suggested to refer to [3].

Subsequent to the building of the semantic hierarchy, the subsumption relationships between all pairs of concepts ($c_i, c_j \in C \cup C'$) are added to our ontology according to the hierarchy structure. This is achieved automatically using the axiom A6 illustrated in Table 1.

Figure 3 illustrates the built semantic hierarchy on the Pascal VOC'2010 dataset. This semantic hierarchy allowed to define the subsumption relationships between image concepts. We can observe that the produced hierarchy is a *N-ary tree* like-structure, where leaf nodes are the concepts in $\mathcal{C}$. Mid-level concepts are automatically recovered from WordNet based on the previously introduced method. We can also observe that the connected concepts share strong visual and semantic similarity, which justifies the choice of this method in our approach. We therefore concur with the assumption that a suitable semantic hierarchy for representing image semantics should incorporate visual and conceptual (semantic) modalities during the building process [3].

5.4 Definition of the *ABox*

Following the building of the semantic hierarchy that will be used as the backbone of our ontology, information about the context of images is added to our ontology in order to design a more representative knowledge base of image semantics. This information, mainly consisting of contextual and spatial relationships between image

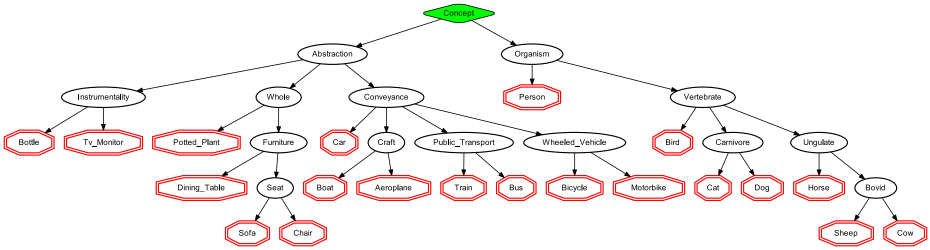


Fig. 3 The semantic hierarchy built on Pascal VOC'2010 dataset. *Double octagon* nodes are original concepts, i.e. concepts of $\mathcal{C}$, and the *diamond* one is the root of the produced hierarchy

concepts will forms the *ABox* of our ontology and will serves for reasoning about image annotation. Furthermore, our intent is to design a fuzzy multimedia ontology in order to model the inherent uncertainty of concept relationships, which should lead to a more efficient decision-making during the image annotation process. Consequently, we introduce in the following how the confidence degrees of each of the proposed fuzzy roles (concept relationships) are computed.

5.4.1 Contextual relationships

Contextual information is of great interest to help understanding the image semantics. A simple form of contextual information is the co-occurrence frequency of a pair of concepts. For example, it is intuitively clear that if two concepts are similar or related, it is likely that their role in the world will be similar, and thus their context of occurrence will be equivalent (i.e. they tend to occur in similar contexts, for some definition of context). For instance, a photo containing "Television" and "Sofa" depicts usually a "Living-room" scene. Nevertheless, contextual similarity is a 'corpus-dependent' measure, i.e. depends on the concepts distribution in the dataset. It is therefore important to normalize the measures based on contextual information.

In our approach, we define three contextual relationships that we estimated important for reasoning about image annotation. These are: $\mathcal{CON} = \{$*"hasFrequency", "hasAppearedWith", "isAnnotatedBy"*$\}$. However, nothing prevents the enrichment of our multimedia ontology with other contextual relationships in order to adapt to other reasoning scenarios. The proposed relations ($\in \mathcal{CON}$) are detailed bellow.

Let us consider an image database $\mathcal{DB}$, where:

- $\mathcal{L}$ is the number of images in the database,
- $\mathcal{N}$ is the size of the annotation vocabulary,
- n_i is the number of images annotated by c_i (occurrence frequency of c_i), and
- n_{ij} the number of images co-annotated by c_i et c_j.

Our objective is to estimate $P(c_i)$ as the probability of occurrence of a given concept c_i (and respectively $P(c_i, c_j)$ as the joint probability of c_i and c_j) in $\mathcal{DB}$. These probabilities can be easily estimated by:

$$\widehat{P(c_i)} = \frac{n_i}{\mathcal{L}} \tag{3}$$

$$\widehat{P(c_i, c_j)} = \frac{n_{ij}}{\mathcal{L}} \tag{4}$$

Based on these probabilities, we define the concept frequency relationship as the concrete feature: $hasFrequency : \Delta^{\mathcal{I}} * \Delta_D \rightarrow \{0, 1\}$, where $\Delta^{\mathcal{I}} = \mathcal{C}$ and $\Delta_D = [0, 1]$ are the interpretation domains. This concrete feature associates to each concept $c_i \in \mathcal{C}$ a fuzzy degree corresponding to its occurrence frequency in $\mathcal{DB}$:

$$\mu_{\text{hasFrequency}(c_i)} = P(c_i) \tag{5}$$

We also define the contextual relationship '*hasAppearedWith*' as the fuzzy role $hasAppearedWith : \Delta^{\mathcal{I}} * \Delta^{\mathcal{I}} \rightarrow [0, 1]$, where $\Delta^{\mathcal{I}} = \mathcal{C}$. The membership degree of this relationship is computed using the Normalized Pointwise Mutual Information

(NPMI). To this purpose, the Pointwise Mutual Information $\rho(c_i, c_j)$ is firstly computed for all pairs of concept $c_i, c_j \in \mathcal{C}$ as follows:

$$\rho(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} = \log \frac{\mathcal{L} * n_{ij}}{n_i * n_j} \tag{6}$$

$\rho(c_i, c_j)$ quantifies the amount of information shared between the two concepts c_i and c_j. Thus, if c_i and c_j are independent concepts, then $P(c_i, c_j) = P(c_i) \cdot P(c_j)$ and therefore $\rho(c_i, c_j) = log\ 1 = 0$. $\rho(c_i, c_j)$ can be negative if c_i et c_j are negatively correlated. Otherwise, $\rho(c_i, c_j)$ is positive and quantifies the degree of dependence between these two concepts. In this work, we only want to estimate the positive correlation between each pair of concepts from the annotation vocabulary and therefore we set the negative values of $\rho(c_i, c_j)$ to 0. Moreover, in order to normalize it into [0,1], the membership degree of the fuzzy role '*hasAppearedWith*' is computed as follows:

$$\mu_{\text{hasAppearedWith}(c_i, c_j)} = \frac{\rho(c_i, c_j)}{-\log[\max(P(c_i),\ P(c_j))]} \tag{7}$$

Finally, we define the fuzzy role '*isAnnotatedBy*' as a relationship between instances of concepts "Image" and "Annotation", i.e. $isAnnotatedBy : \Delta^{\mathcal{I}} * \Delta^{\mathcal{I}} \rightarrow [0, 1]$, where $\Delta^{\mathcal{I}} = \{Image, Annotation\}$. This relationship is intended to represent the probability of finding an image in $\mathcal{DB}$ annotated by a set of concepts $(Annotation_j = \langle c_1, c_2, \cdots, c_\Lambda \rangle)$, or inversely, the likelihood that a given annotation '$Annotation_j$' is associated with an image $i_i \in \mathfrak{I}$. To this end, all the possible annotations in $\mathcal{DB}$ are extracted and are added to our ontology as subconcepts of concept "Annotation". The confidence value of this relationship is computed as follows:

$$\mu_{\text{isAnnotatedBy}(Image_1, Annotation_j)} = \frac{n_{Annotation_j}}{\mathcal{L}} \tag{8}$$

where $Annotation_j = \langle c_1, c_2, \cdots, c_\Lambda \rangle$ is a textual annotation used for annotating a set of images in $\mathcal{DB}$, $n_{Annotation_j}$ is the number of images annotated by $Annotation_j$, and $\mathcal{L} = |\mathfrak{I}|$ is the total number of images in $\mathcal{DB}$.

For instance, Example 3 illustrates some inputs of the added assertions to our *ABox*.

Example 3 (Contextual relationship: '*isAnnotatedBy*')

$$\begin{array}{rcl}
\langle Annotation_1 & \equiv & Aeroplane \sqcap Car \sqcap Person \rangle \\
\langle Annotation_1 & \sqsubseteq & Annotation \geq 1 \rangle\rangle \\
\langle Annotation_2 & \equiv & Dining\_Table \sqcap Chair \sqcap Bottle \sqcap Dog \rangle \\
\langle Annotation_2 & \sqsubseteq & Annotation \geq 1 \rangle \\
\langle a & : & Image \geq 1 \rangle \\
\langle b & : & Annotation_1 \geq 1 \rangle \\
\langle (a : b) & : & isAnnotatedBy \geq 0.023064 \rangle \\
& \cdots &
\end{array}$$

5.4.2 Spatial relationships

Spatial information is a valuable source for the understanding of image semantics. The spatial arrangement of objects provides an important information for the recognition and interpretation tasks, and allows to solve the ambiguity between objects having a similar appearance [7]. For instance, using object detectors if one have detected in an image that “Sky” has appeared bellow “Sea”, it is easy to fix this prediction using spatial information because any well defined knowledge base ($\mathcal{KB}$) would allow to detect and correct this inconsistency.

In our approach, eight spatial relationships are used in order to define the directional positions and distances between image concepts. The directional relationships are defined as follows: $\mathcal{DIR}$ = *{“hasAppearedAbove”, “hasAppearedBelow”, ‘hasAppearedLeftOf”, “hasAppearedRightOf”, “hasAppearedAlignedWith”}*, such as $\forall \mathcal{X} \in \mathcal{DIR}, \mathcal{X} : \Delta^{\mathcal{I}} * \Delta^{\mathcal{I}} \rightarrow [0, 1]$, with $\Delta^{\mathcal{I}} = \mathcal{C}$.

The relationships in $\mathcal{DIR}$ are derived from the following primitives: ’*left*’, ‘*right*’, ‘*above*’, ‘*below*’ and ‘*aligned*’, which are computed according to the angle between the segment joining two points ’*a*’ and ’*b*’ (where ‘*a*’ and ‘*b*’ are the centroids of two given objects in a given image) and the *x-axis* of the image—cf. Fig. 4. This angle, denoted $\theta(a, b)$, takes values in $[-\pi, \pi]$ which constitutes the domain of definition of these primitives. They are then computed using $cos^2\theta$ and $sin^2\theta$, and are functions from $[-\pi, \pi]$ into $\{0, 1\}$. Thus, any of the previous primitives can be computed by an angle α with the *x-axis* as illustrated in Fig. 5.

Regarding the primitive ‘*aligned*’, it takes 1 when $\theta \in [-\pi/6, \pi/6] \cup [5\pi/6, -5\pi/6]$ and 0 otherwise. A comprehensive survey about spatial relationships for image processing can be found in [7].

The confidence value of a given directional relationship is finally computed as follows:

$$\mu_{\mathcal{X}(c_i,c_j)} = \frac{\sharp \text{ of instances where } \mathcal{X}(c_i, c_j)}{n_{ij}} \qquad (9)$$

where $c_i, c_j \in \mathcal{C}$, and $\mathcal{X}$ is a directional relationship, i.e. $\mathcal{X} \in \mathcal{DIR}$.

In addition, we define in our approach the distance relationships as $\mathcal{DIS}$ = {*“hasAppearedCloseTo”, “hasAppearedFarFrom”*}, such as $\forall \chi \in \mathcal{DIS}, \chi : \Delta^{\mathcal{I}} * \Delta^{\mathcal{I}} \rightarrow [0, 1]$, with $\Delta^{\mathcal{I}} = \mathcal{C}$. These distance relationships are computed according to the Euclidean distance on the considered objects. To this purpose, let us consider in a given image two objects O and P defined by their centroids (x_1, y_1) and (x_2, y_2), and

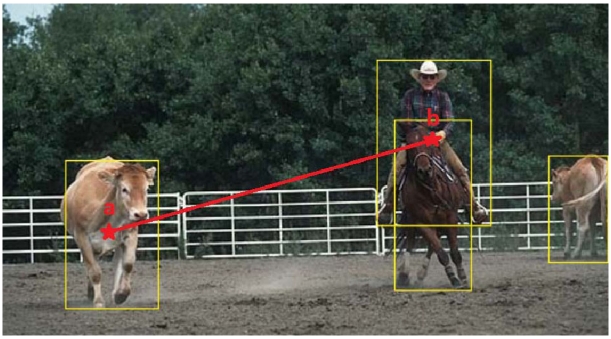


Fig. 4 Spatial primitives are computed according to the angle between the segment joining two points ‘*a*’ and ‘*b*’ and the *x-axis* of the image. ‘*a*’ and ‘*b*’ are the centroids of two given objects (here “Cow” and “Person”) in a given image

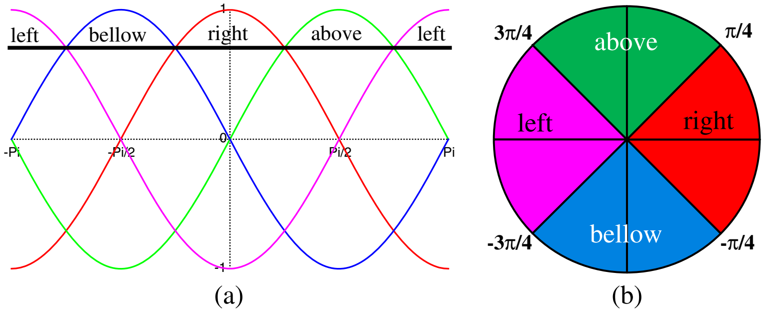


Fig. 5 Directional relationships are computed according to an angle α with the *x-axis*

their bounding box $(O_{x\min}, O_{x\max}, O_{y\min}, O_{y\max})$ and $(P_{x\min}, P_{x\max}, P_{y\min}, P_{y\max})$. We define then the following primitives:

$$distance(O, P) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{10}$$

$$size(O) = \sqrt{(O_{x\max} - O_{x\min})^2 + (O_{y\max} - O_{y\min})^2} \tag{11}$$

$$close(O, P) = \begin{cases} 1 & \text{if } distance(O, P) < 2(size(O) + size(P)) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$farfrom(O, P) = \begin{cases} 1 & \text{if } distance(O, P) \geq 2(size(O) + size(P)) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Using the previous primitives, distance relationships can easily be computed by the following equation:

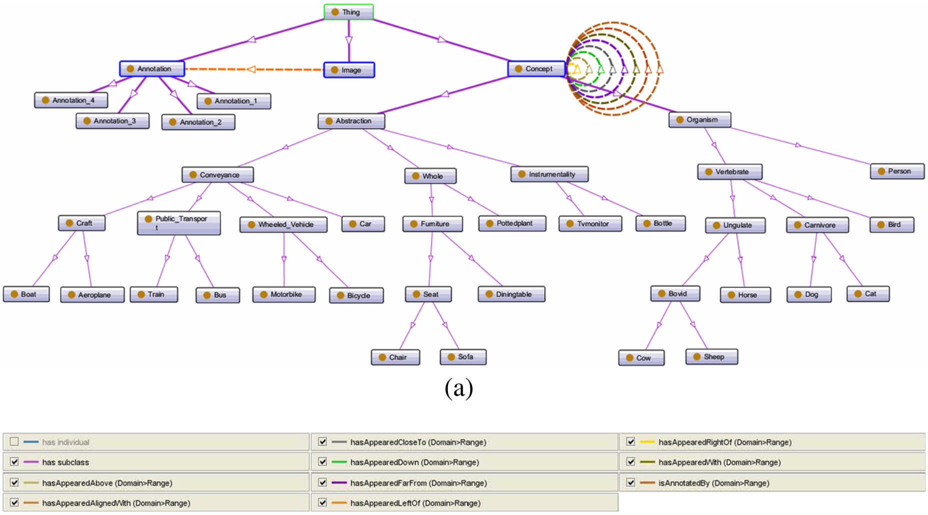
$$\mu_{\chi(c_i,c_j)} = \frac{\sharp \text{ of instances where } \chi(c_i, c_j)}{n_{ij}} \tag{14}$$

where $c_i, c_j \in \mathcal{C}$, and χ is a distance relationship, i.e. $\chi \in \mathcal{DIS}$.

Example 4 (Spatial relationships)

$$\begin{aligned}
\langle a &: Bottle \geq 1\rangle \\
\langle b &: Dining\_Table \geq 1\rangle \\
\langle (a:b) &: hasAppearedAbove \geq 0.76\rangle \\
\langle (a:b) &: hasAppearedBelow \geq 0.02\rangle \\
\langle (a:b) &: hasAppearedAlignedWith \geq 0.62\rangle \\
\langle (a:b) &: hasAppearedCloseTo \geq 0.97\rangle \\
&\ldots
\end{aligned}$$

In order to illustrate our approach for building multimedia ontologies, we show in Fig. 6 an extract of the built ontology on Pascal VOC dataset. This figure depicts the main concepts of the built ontology and the used roles for defining concepts



(a)

☐ has individual	☑ hasAppearedCloseTo (Domain>Range)	☑ hasAppearedRightOf (Domain>Range)
☑ has subclass	☑ hasAppearedDown (Domain>Range)	☑ hasAppearedWith (Domain>Range)
☑ hasAppearedAbove (Domain>Range)	☑ hasAppearedFarFrom (Domain>Range)	☑ isAnnotatedBy (Domain>Range)
☑ hasAppearedAlignedWith (Domain>Range)	☑ hasAppearedLeftOf (Domain>Range)	

(b)

Fig. 6 An extract of the built multimedia ontology on Pascal VOC dataset is illustrated in figure **a**. *Dashed arrows* represent the fuzzy roles used for defining the contextual and spatial relationships between concepts. Figure **b** illustrates the roles names

relationships. Full arrows represent the subsumption relationships between the ontology concepts. Dashed arrows represent the fuzzy roles used for defining the contextual and spatial relationships between concepts. For the clarity of the illustration we restricted the *Annotation*$_j$ concept number to 4 and we did not displayed the instances (individuals).

6 Proposed method for image annotation: Multi-stage reasoning framework for image annotation

Automatic image annotation is still a challenging problem despite more than a decade of research. Indeed, current approaches are struggling to scale up because of the lack of a computational model allowing to model such a complex system, the uncertainty introduced by the statistical learning algorithms, the dependency on the accuracy of the ground truth of the training dataset and the well-known semantic gap problem. Given a training dataset, automatic image annotation often consists in building a computational model that enables to predict a set of concepts from the annotation vocabulary to previously unseen images.

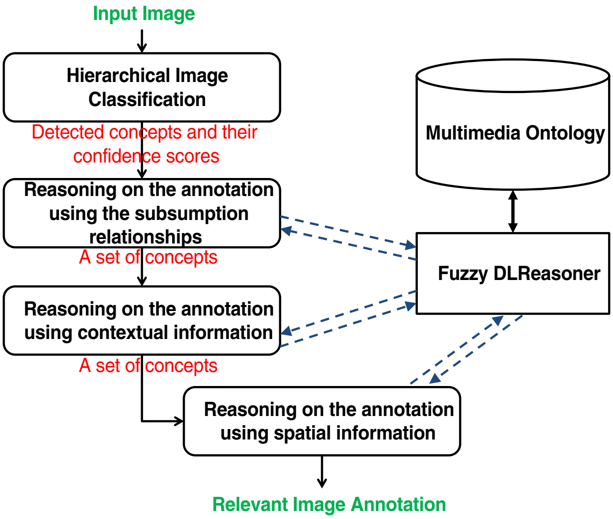
Image classification is a widely used technique for image annotation. It consists in performing several binary SVM classifiers on an input image to find to which classes it belongs to. The annotation of an image depends therefore on the classifier outputs, i.e. an image is annotated by a concept $c_i \in \mathcal{C}$ if the output of the classifier associated to c_i is positive. Usually, such a process involves considerable uncertainty because of the errors introduced by the machine learning algorithms. However, this uncertainty can be reduced using reasoning over the produced image annotation. For instance, it is most often easy to compute a confidence score (membership value) for the classification of an image to a given class. Such information is valuable and can be of great importance to improve image classification accuracy. For instance, one can improve image annotation in a post-classification process based on these confidence scores and an explicit knowledge source, such as an ontology which models images context. In that way, this uncertainty is used itself as a knowledge source in order to achieve a better decision-making on the image annotation. Furthermore, the use of an explicit knowledge model can help model, reduce, or even remove this uncertainty by supplying a formal framework to reason about the consistency of extracted information from images.

Our approach is motivated by the above assumption. Indeed, we propose in the following a multi-stage reasoning framework for image annotation based on the earlier built multimedia ontology. The proposed framework allows reasoning on the provided annotations by the image classification algorithm in order to achieve a semantically relevant image annotation. A global overview of the proposed approach is illustrated in Fig. 7.

Specifically, we consider the following problem. Given a formal multimedia ontology designed as a fuzzy knowledge base $\mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$, where $\mathcal{T}$ is a fuzzy Terminological Box (*TBox*), $\mathcal{R}$ is a regular fuzzy Role Box (*RBox*), and $\mathcal{A}$ is a fuzzy Assertional Box (*ABox*). This fuzzy knowledge base is assumed to contain the following explicit knowledge about ontology concepts: i) subsumption relationships, ii) contextual relationships, and iii) spatial relationships. This multimedia ontology is then used within our framework for annotating previously unseen images. As illustrated in Fig. 7, this is achieved by the following steps:

- A hierarchical classification is performed on the input image, and the confidence score for each concept $c_j \in \mathcal{C} \cup \mathcal{C}'$ is recovered.
- These concepts and their confidence scores are thereafter transformed into fuzzy description logics assertions, and their consistency is checked using the

Fig. 7 Proposed method: a knowledge-based multi-stage reasoning framework for image annotation


Input Image
Hierarchical Image Classification
Detected concepts and their confidence scores
Reasoning on the annotation using the subsumption relationships
A set of concepts
Reasoning on the annotation using contextual information
A set of concepts
Reasoning on the annotation using spatial information
Relevant Image Annotation
Multimedia Ontology
Fuzzy DLReasoner

subsumption relationships and our fuzzy DL reasoner. Inconsistent concepts are removed from the candidate annotation[3] of the input image.

- Thereafter, the consistency of the set of concepts from the candidate annotation is checked with respect to the contextual relationships and our fuzzy DL reasoner. Inconsistent concepts are again removed from the candidate annotation of the input image.
- Finally, the consistency of the candidate annotation is checked with respect to the spatial information, and the final (candidate) annotation is associated with the input image. This final annotation is supposed to be semantically consistent.

6.1 Hierarchical image classification

Based on the subsumption hierarchy, we propose in the following to train several classifiers that represent the same concept at different levels of abstraction. These classifiers are consistent with each other since they are linked by the subsumption relationship, and then represent the same information with different levels of details. Therefore, it is possible to reason on the outputs of these classifiers in order to achieve a relevant decision on the belonging of an image to a given class.

Concretely, given a semantic (subsumption) hierarchy, a classifier for each concept node of the hierarchy is trained by performing a One-Versus-All (OVA) Support Vector Machines [11]. Specifically, for training the classifier of a target concept node we took as positive samples all images associated with its children leaf nodes. Negative samples are all the other images of the training database. Therefore, the semantic hierarchy is only used to recover the set of positive and negative sample images for training the classifiers of each concept node at the different layers of the hierarchy. Consequently, the decision function of each classifier is independent from its subsumed (child) and subsuming (parent) concept nodes.

Let x_i^v be any visual representation of an image $i_i \in \mathfrak{I}$ (a visual feature vector), we train for each concept class ($c_j \in \mathcal{C} \cup \mathcal{C}'$) in the hierarchy a classifier that can associate c_j with its visual features. This is achieved by the use of $|\mathcal{C}| + |\mathcal{C}'|$ binary SVM OVA, with a decision function:

$$\mathcal{G}(x_i^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x_i^v) + b \tag{15}$$

where $\mathbf{K}(x_k^v, x_i^v)$ is the value of a kernel function for the training sample x_k^v and the test sample x_i^v, $y_k \in \{1, -1\}$ is the class label of x_k^v, α_k is the learned weight of the training sample x_k^v, and b is a learned threshold parameter.

Radial Basis Function (RBF) kernel is used for the training of our SVM:

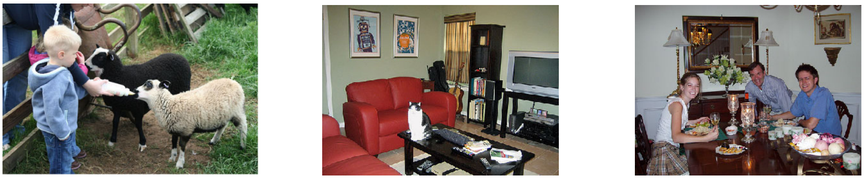
$$\mathbf{K}\left(x_k^v, x_i^v\right) = \exp\left(\frac{\|x_k^v - x_i^v\|^2}{\sigma^2}\right) \tag{16}$$

[3] A candidate annotation $\mathcal{P}$ consists of a set of candidate concepts $\{c_j \in \mathcal{C} \cup \mathcal{C}', j = 1..n_{i_i}\}$ and their confidence values $\{\alpha_j, j = 1..n_{i_i}\}$, predicted as describing the image content.

6.2 Reasoning on image annotation using the subsumption hierarchy

Based on the classifiers outputs and the subsumption relationships, we propose in the following to check the consistency of candidate concepts. So, let us consider a previously unseen image $i'_i \in \mathfrak{I}'$. Performing a hierarchical image classification on i'_i produces an output $\mathcal{P}$ which consists of a set of candidate concepts $\{c_j \in \mathcal{C} \cup \mathcal{C}', \ j = 1..n_{i'_i}\}$ and their confidence values $\{\alpha_j, \ j = 1..n_{i'_i}\}$, i.e. $\mathcal{P} = \langle(c_0, \alpha_0), (c_1, \alpha_1), \cdots (c_m, \alpha_m)\rangle$ as illustrated in Fig. 8. Subsequently, these concepts and their confidence scores are transformed into fuzzy description logics assertions. In order to do so, we first normalize into [0, 1] the outputs $\{\alpha_j, \ j = 1..n_{i'_i}\}$ of the SVM classifiers by assigning zero to negative values and performing min-max normalization on the positive values. Thereafter, the consistency of each concept $c_j \in \mathcal{C}$ is checked using the subsumption relationships and our fuzzy DL reasoner. Inconsistent concepts are removed from the candidate annotation.

Specifically, our objective is to check the consistency of a candidate concept $c_j \in \mathcal{C}$ to a given image i'_i using the subsumption relationships, and thus the set of



<table>
<tr><th colspan="3">Groundtruth:</th></tr>
<tr><td>Sheep, Person</td><td>Cat, Tv_monitor, Sofa</td><td>Chair, Dining_Table: <i>Marked as Difficult</i>, Person</td></tr>
<tr><th colspan="3">Classifier Outputs for Concepts $c_j \in \mathcal{C}$:</th></tr>
<tr><td>Aeroplane: -1.192, Bicycle: -0.012, Bird: -0.639, Boat: -0.474, Bottle: -0.347, Bus: -0.367, Car: -0.525, Cat: -0.244, Chair: -0.310, Cow: 0.310, Dining_table: 0.162, Dog: -0.0211, Horse: 0.391, Motorbike: 0.262, Person: 0.805, Potted_plant: -0.012, Sheep: 0.519, Sofa: -0.465, Train: -0.259, Tv_monitor: -0.701</td><td>Aeroplane: -0.491, Bicycle: 0.196, Bird: -0.723, Boat: 0.055, Bottle: -0.296, Bus: -0.464, Car: -0.108, Cat: 0.758, Chair: 0.428, Cow: -0.900, Dining_table: 0.391, Dog: -1.031, Horse: -0.118, Motorbike: -0.098, Person: 0.069, Potted_plant: 0.148, Sheep: -0.925, Sofa: 0.858, Train: 0.098, Tv_monitor: 0.421</td><td>Aeroplane: -1.086, Bicycle: 0.106, Bird: -0.752, Boat: -0.792, Bottle: 0.807, Bus: -0.330, Car: -0.185, Cat: -0.207, Chair: 1.024, Cow: -0.458, Dining_table: 0.854, Dog: 0.271, Horse: -0.109, Motorbike: 0.147, Person: 1.240, Potted_Plant: 0.584, Sheep: -0.670, Sofa: -0.046, Train: -0.530, Tv_monitor: 0.158</td></tr>
<tr><th colspan="3">Classifier Outputs for Concepts $c_j \in \mathcal{C}'$:</th></tr>
<tr><td>Abstraction: 0.109, Bovid: 0.499, Carnivore: -0.012, Conveyance: -0.377, Craft: -1.040, Furniture: -0.135661, Instrumentality: -0.659, Organism: 0.636, Public_transport: -0.377, Seat: -0.243, Ungulate: 0.391, Vertebrate: 0.056, Wheeled_vehicle: -0.088, Whole: -0.106</td><td>Abstraction: 1.098, Bovid: -0.976, Carnivore: 0.875, Conveyance: 0.033, Craft: -0.671, Furniture: 1.229, Instrumentality: 0.785, Organism: 0.488, Public_transport: -0.108, Seat: 0.361, Ungulate: -0.682, Vertebrate: 0.508, Wheeled_vehicle: -0.294, Whole: 1.065</td><td>Abstraction: 1.072, Bovid: -0.368, Carnivore: -0.049, Conveyance: -1.077, Craft: -1.446, Furniture: 1.145, Instrumentality: 0.775, Organism: 0.647, Public_transport: -0.185, Seat: 0.513, Ungulate: -0.202, Vertebrate: -0.138, Wheeled_Vehicle: 0.020, Whole: 1.179</td></tr>
<tr><th colspan="3">Reasoning on the annotations using the subsumption hierarchy:</th></tr>
<tr><td>Cow: 0.310, Horse: 0.391, Person: 0.805, Sheep: 0.519
~~Motorbike~~, ~~Dining_table~~</td><td>Cat: 0.616, Chair: 0.348, Dining_table: 0.318, Person: 0.056, Potted_plant: 0.120, Sofa: 0.698, Tv_monitor: 0.342
~~Bicycle~~, ~~Boat~~, ~~Train~~</td><td>Bottle: 0.650, Chair: 0.825, Dining_table: 0.688, Person: 1.00, Potted_Plant: 0.470, Tv_monitor: 0.127
~~Bicycle~~, ~~Dog~~, ~~Motorbike~~</td></tr>
<tr><th colspan="3">Reasoning on the annotations using image context:</th></tr>
<tr><td>Person: 0.805, Sheep: 0.519
~~Horse~~, ~~Cow~~</td><td>Cat: 0.616, Chair: 0.348, Dining_table: 0.318, Sofa: 0.698, Tv_monitor: 0.342
~~Person~~, ~~Potted_plant~~</td><td>Bottle: 0.650, Chair: 0.825, Dining_table: 0.688, Person: 1.00, Potted_Plant: 0.470
~~Tv_monitor~~</td></tr>
<tr><th colspan="3">Reasoning on the annotations using spatial information:</th></tr>
<tr><td>Person: 0.805, Sheep: 0.519</td><td>Cat: 0.616, Chair: 0.348, Dining_table: 0.318, Sofa: 0.698, Tv_monitor: 0.342</td><td>Bottle: 0.650, Chair: 0.825, Dining_table: 0.688, Person: 1.00, Potted_Plant: 0.470</td></tr>
</table>

Fig. 8 Illustrative examples of the proposed method for annotating images

its hypernyms $\{c_k \in \mathcal{C}' \mid c_j : C > 0, c_k : D > 0, C \sqsubseteq D > 0\}$. Therefore, the reasoning process can be formulated using conjunctive queries as follows:

$$valid(c_j) \leftarrow \mathcal{P}(c_j) > 0 \wedge c_j : C > 0 \wedge c_k : D > 0 \wedge C \sqsubseteq D > 0 \wedge valid(c_k)$$
$$valid(\top) = 1$$

where $\top$ is the root of the ontology, and $\mathcal{P}(c_j)$ represents the confidence score of the concept c_j given by α_j.

In DL, given an abstract individual 'a' (an instance of a given candidate concept), the consistency checking of concept inclusions is performed as follows. For $C \sqsubseteq D$, we compute the greatest lower bound $glb(\mathcal{KB}, C \sqsubseteq D)$ using Axiom A6 in Table 1, i.e. as the minimal value of x such that $\mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \cup \{\langle a : C, \alpha_1 \rangle\} \cup \{\langle a : D, \alpha_2 \rangle\} \rangle$ is satisfiable under the constraints expressing that $\alpha_1 \rightarrow \alpha_2 \leq x$, with α_1 and $\alpha_2 \in [0, 1]$. This process is then iterated until the root of the ontology is reached. Thus, we come up with the following hierarchy: $C_1 \sqsubseteq C_2 \geq x_1, C_2 \sqsubseteq C_3 \geq x_2, \cdots, C_n \sqsubseteq \top \geq 1$. Thereafter, a confidence score for the considered candidate concept is computed as follows:

$$bed(\mathcal{KB}, a : ValidCC) = x_1 \otimes x_2 \otimes \cdots \otimes 1 = x_1 * x_2 * \cdots * 1 \tag{17}$$

where *ValidCC* stands for a *Valid Candidate Concept*, which is a concept defined to regroup all the consistent candidate concepts.

Finally, all candidate concepts with a confidence score equal to zero are removed from the annotation of the image i'_i.

In order to illustrate our approach, let us consider the first example in Fig. 8 where evaluations were performed on Pascal VOC'2010 dataset. The image classification algorithm has detected "Motorbike" as a candidate concept (among others) for the considered image. However, according to the subsumption hierarchy (cf. Fig. 3) a "Motorbike" $\sqsubseteq$ "Wheeled_vehicle" $\sqsubseteq$ "Conveyance", etc., and therefore the classifiers should also have detected these concepts to stay coherent. The consistency checking of the concept "Motorbike" is performed according to the previously described procedure [–cf. Example 5], and thus this concept is removed from the list of candidates since $bed(\mathcal{KB}, Motorbike : ValidCC) = 0$.

Example 5 (Consistency checking of concept "Motorbike")

$$\begin{aligned} \mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} &\cup \{\langle a : Motorbike \geq 0.262 \rangle\} \\ &\cup \{\langle a : Weeled\_vehicule \geq 0 \rangle\} \\ &\cup \{\langle a : Conveyance \geq 0 \rangle\} \\ &\cup \{\langle a : Abstraction \geq 0.109 \rangle\} \\ &\cup \{\langle a : Concept \geq 1 \rangle\} \rangle \end{aligned}$$

$$bed(\mathcal{KB}, Motorbike : ValidCC) = 0.262 \otimes 0 \otimes 0 \otimes 0.109 \otimes 1 = 0$$

6.3 Reasoning on image annotation using image context

As aforementioned, contextual information can provide valuable information for the understanding of image context or to reason about the consistency of image annotation. For instance, it is evident that an image which contains the set of concepts {“Aeroplane”, “Person”, “Car”} represents a scene of an *airport tarmac*, and not the one of a *flying plane*. And conversely, it is obvious that an image that contains “Dining_table” and “Sofa” should not contain “Boat” or “Bus”. Thus, contextual information, if processed, can be helpful to check the consistency of image annotations.

Using our multimedia ontology, it is easy to recover contextual information about images. Consequently, we propose in the following to use this information to recover from our ontology all consistent annotations with respect to contextual information, and to compute the best explanation of a considered image. Specifically, the fuzzy role “isAnnotatedBy” allows predicting a confidence score (based on contextual information) for a given set of candidate concepts. Given a *Candidate Annotation* $CA_j = \langle c_1, c_2, \cdots, c_m \rangle$ and a target image $i'_i \in \mathfrak{I}'$, a confidence score is computed to estimate the correlation likelihood between CA_j and i'_i. This confidence score increases according to the likeliness of the candidate annotation CA_j, or it is equal to 0 when the annotation is not valid.

Concretely, given an image i'_i and $\mathcal{P}' : \langle (c_0, \alpha_0), (c_1, \alpha_1), \cdots (c_m, \alpha_m) \rangle, m = |\mathcal{P}'|$, a set of valid candidate concepts with respect to the subsumption relationships, we build first the set of candidate annotation (CA_j, $j \in 1..|combinaisons|$) by taking all the possible combination of the concepts in $\mathcal{P}'$. A confidence score for each valid candidate annotation (*ValidCA*) is then computed. For instance, let us assume that we dispose of one candidate annotation consisting of 3 concepts. Its confidence score is computed as follows:

Example 6 (Reasoning using image context)

$\mathcal{P}' : \langle (c_1, \alpha_1), (c_2, \alpha_2), (c_3, \alpha_3) \rangle$, (classifier outputs)
$\langle c_1 : C_1 \geq \alpha_1 \rangle$, $\langle c_2 : C_2 \geq \alpha_2 \rangle$, $\langle c_3 : C_3 \geq \alpha_3 \rangle$
$\langle CA \equiv C_1 \sqcap C_2 \sqcap C_3 \rangle$
$\langle b : CA \geq \alpha_b \rangle$, $s.t.\ \alpha_b = \alpha_1 \otimes \alpha_2 \otimes \alpha_3$
$\mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \cup \{\langle a : Image \geq \alpha_a \rangle\} \cup \{\langle b : CA \geq \alpha_b \rangle\} \rangle$
$\langle (a, b) : isAnnotatedBy \geq \alpha_r \rangle$, is already stored in the $\mathcal{KB}$ during the ontology building process, where $\alpha_r = \mu_{\text{isAnnotatedBy(a,b)}}$ (cf. (8)).

Therefore, according to (1), the correlation likelihood between a candidate annotation CA and a given image i'_i can be computed as follows:

$$glb\,(\mathcal{KB}, a : \exists\, isAnnotatedBy.CA) = \alpha_b \otimes \alpha_r = (\alpha_1 \otimes \alpha_2 \otimes \alpha_3) \otimes\ \mu_{\text{isAnnotatedBy}(a,b)} \tag{18}$$

then,

$$ValidCA \equiv \exists\, isAnnotatedBy.CA \tag{19}$$

Finally, the best explanation (bex) of i'_i is retrieved as the *ValidCA* having the maximum correlation likelihood among all the others. This explanation is computed as follows:

$$bex(\mathcal{KB}, ValidCA) = \{\langle a, r \rangle | r = bed(\mathcal{KB}, a : ValidCA)\} \tag{20}$$

For instance, let us consider the first example in Fig. 8. We show below some cases of DL reasoning using the contextual information:

Example 7 (DL Reasoning using image context)

$$\begin{aligned}
&\mathcal{P}' : \langle (c_1 : Horse, 0.391), (c_2 : Person, 0.805), (c_3 : Sheep, 0.519), (c_4 : Cow, 0.310)\rangle \\
&\langle CA_0 \equiv Horse \sqcap Person \sqcap Sheep\rangle \\
&\langle CA_1 \equiv Person \sqcap Sheep\rangle \\
&\langle CA_2 \equiv Cow \sqcap Person\rangle \\
&\langle b_0 : CA_0 \geq 0.163\rangle \\
&\langle b_1 : CA_1 \geq 0.417\rangle \\
&\langle b_2 : CA_2 \geq 0.249\rangle \\
&\mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \cup \{\langle a : Image \geq 1\rangle\} \cup \{\langle b_0 : CA_0 \geq 0.163\rangle\} \cup \{\langle b_1 : CA_1 \geq 0.417\rangle\} \\
&\qquad \cup \{\langle b_2 : CA_2 \geq 0.249\rangle\}\rangle \\
&glb\,(\mathcal{KB}, a : \exists\, isAnnotatedBy.CA_0) = (0.391 \otimes 0.805 \otimes 0.519) \otimes 0.003548 = 0.00057 \\
&glb\,(\mathcal{KB}, a : \exists\, isAnnotatedBy.CA_1) = (0.805 \otimes 0.519) \otimes 0.027413 = \mathbf{0.01145} \\
&glb\,(\mathcal{KB}, a : \exists\, isAnnotatedBy.CA_2) = (0.391 \otimes 0.805) \otimes 0.025455 = 0.00635 \\
&bex(\mathcal{KB}, ValidCA) = 0.01145
\end{aligned}$$

Consequently, with respect to the contextual information, the best explanation for the left image in Fig. 8 is: $CA_1 \equiv Person \sqcap Sheep$.

Please note that, since most images of the Pascal VOC dataset contain only one or two concepts [10], and thus the distribution of multi-labeled images is not uniform, we computed (8) for this dataset as:

$$\mu_{\text{isAnnotatedBy}(Photo, Annotation_i)} = \frac{n_{Annotation_i}}{\mathcal{L}} * \exp(\Lambda) \tag{21}$$

where $\Lambda = |Annotation_i|$.

6.4 Reasoning using spatial information

Contextual knowledge can help the recognition of objects within a scene by providing predictions about objects that are most likely to appear in a specific setting, i.e. *topological information*, along with the locations that are most likely to contain objects in the scene, i.e. *spatial information*. Specifically, the spatial arrangement of objects provides important information for the recognition and interpretation tasks, and allows to solve ambiguity between objects having a similar appearance. As part of this work, we have proposed an approach based on image classification for annotating images. Consequently, we do not dispose of the spatial position of detected concepts, and therefore the reasoning capabilities using spatial information are limited in the current approach. However, we propose in the following a simple but effective usage scenario that relies on the spatial arrangement of the currently detected concepts in order to provide a semantically consistent image annotation. In Section 8, we propose some usage scenarios that illustrate the usefulness of spatial information and the reasoning over this kind of knowledge in order to improve image annotation.

Given an image $i'_i \in \mathfrak{I}'$ and $\mathcal{P}'' : \langle(c_0, \alpha_0), (c_1, \alpha_1), \cdots (c_m, \alpha_m)\rangle, m = |\mathcal{P}''|$, a set of a valid candidate concepts with respect to the subsumption relationships and contextual information. We propose first to query the ontology in order to retrieve all possible spatial arrangement of all pairs of concepts $(c_j, c_k) \in \mathcal{P}''$, and to recover the confidence score of each of these spatial arrangements. A score can then be computed as the maximum likelihood of all spatial arrangements of these concepts to find the best explanation of i'_i. Algorithm 1 details the different steps of this method.

Algorithm 1 Reasoning using spatial Information

```
Input: A valid candidate annotation: ValidCA
Result: Semantically consistent image annotation
begin
    Find:
    - $C, D \leftarrow \operatorname{argmax}_{x,y \in ValidCA} x.hasAppearedwith(y)$
    - Spatial arrangement $\leftarrow \operatorname{argmax}_{\chi \in \mathcal{DIR}} C.\chi(D)$
    - $E \leftarrow \operatorname{argmax}_{x \in ValidCA} x.hasAppearedwith(C \sqcup D)$
    - Max spatial arrangement of $E$ and $C$ s.t Spatial arrangement of $E$ and $D$ is
    satisfiable
    - Reiterate the process with the remaining concepts in ValidCA
end
```

Reasoning on spatial information should also allow to provide a good image interpretation. For instance, computing the maximum spatial arrangement likelihood allows to retrieve the likeliness of spatial arrangement of each detected concept in a given image. This will allow for example, to provide a textual description of a given image in the following way:

Figure 8, first example:	"*This picture depicts a person standing on the left of a sheep. They are close to each other.*"
Figure 8, second example:	"*This picture depicts a cat sitting on a table in a living room. There is a table, a sofa and a television in the living room.*"

It is easy to implement such a system for image interpretation once we dispose of information about detected concepts and their spatial location [20]. We will address the implementation of such a system in our future work.

7 Experiments

In this paper, evaluations are performed on Pascal VOC'2009 dataset [15] and Pascal VOC'2010 dataset [16]. These datasets contain about 11,000 images and 20 concepts. Each image is annotated with one or more concepts from the annotation vocabulary. In the following, we introduce the used method for visual representation of images, then we present the obtained results on the used datasets and we compare our proposal to recent work.

7.1 Visual representation of images

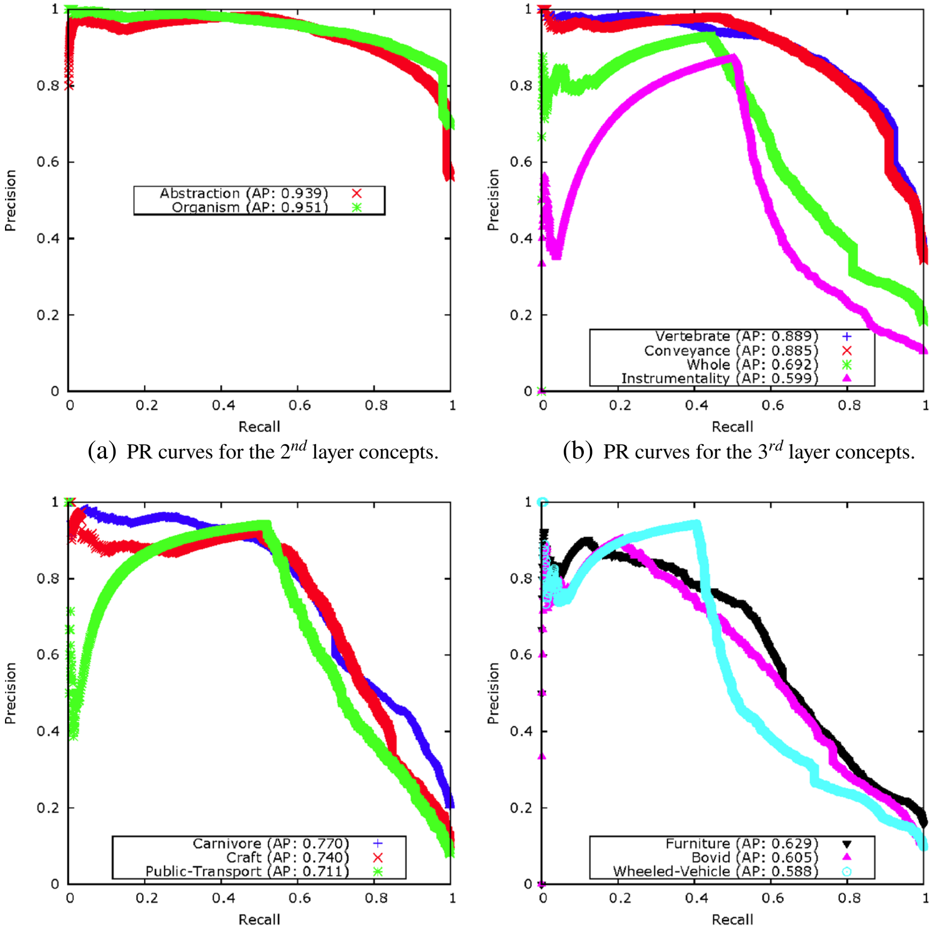
The Bag-of-Features (BoF) representation, also known as Bag-of-Visual-Word (BoVW), is used in this paper to describe image features. The BoF model has shown excellent performances and became one of the most widely used model for image classification and object recognition [28]. In our approach, image features are described as follows: Lowe's DoG Detector [31] is used for detecting a set of salient image regions. A signature of these regions is then computed using SIFT descriptor [31]. Afterwards, given the collection of detected region from the training set of all categories, we generate a codebook of size $K = 1{,}000$ by performing the k-means algorithm. Thus, each detected region in an image is mapped to the most similar visual word in the codebook through a KD-Tree. Each image is then represented by a histogram of K visual words, where each bin in the histogram corresponds to the occurrence number of a visual word in that image.

7.2 Evaluation of image annotation

As aforementioned, experiments are performed on Pascal VOC'2009 and VOC'2010 datasets. Since we do not dispose of the test set used in these challenges, we used 50 % of the image dataset for training the classifiers and the other images are used for evaluating our approach.

In order to emphasize the importance of hierarchical image classification and ontological reasoning using the subsumption relationships, we illustrate in Fig. 9 the obtained average precision and Precision/Recall (PR) curves for all the concepts of each level of the hierarchy. As depicted in this figure, the concepts in the higher levels of the hierarchy have strong average precision, and we can also observe that the classifier accuracy decreases as we go deeper in the hierarchy. These results can be explained as follows. Firstly, the classes in the higher levels of the hierarchy are widely different in their visual appearance, i.e. it is easy to find a boundary that separates these classes. They are also more balanced, i.e. these classes dispose of more positive samples for training their classifiers than the ones in lower levels of the hierarchy. We can therefore conclude that the subsumption relationships should allow improving the image annotation results as they provide a formal framework for reasoning about concepts consistency. Moreover, as the classification accuracy increases as we move to the upper levels of the hierarchy, the overall classification accuracy should increase also.

In Fig. 10, we compare our framework for image annotation to the following methods: a flat classification method, a hierarchical classification one and a baseline method. The baseline method is built by taking the average submission results to Pascal VOC'2010 challenge. The flat classification is performed by using $|C|$ SVM One-Versus-All (OVA), where the inputs are the BoF representation of images and the outputs are the desired SVM responses for each image (1 or -1). We used cross-validation to overcome the unbalanced data problem, taking at each fold as many positive as negative images. Hierarchical classification is performed by training a set of $(|C| + |C'|)$ hierarchical classifiers (OVA) consistent with the structure of the hierarchy illustrated in Fig. 3—for more details about hierarchical classification see Section 6.1. Results are evaluated in terms of Average Precision (AP) scores.



(a) PR curves for the 2^{nd} layer concepts.

(b) PR curves for the 3^{rd} layer concepts.

(c) PR curves for the fourth layer concepts.

Fig. 9 Hierarchical classification: Precision/Recall (PR) curves for the concepts of each level of the hierarchy

As illustrated in Fig. 10, our method for image annotation performs better results than the other ones on Pascal VOC'2010 dataset, with an average precision of 66.49 % and a gain of +8.6 % comparing to the baseline method, a gain of +14.8 % comparing to the hierarchical classification method and a gain of +32.6 % comparing to the flat classification method. These results confirm the effectiveness of the proposed approach, and the importance of contextual and spatial information for improving image annotation. These improvements could be further significant when using a dataset containing more multi-labeled images. Indeed, in Pascal VOC dataset the proportion of images labeled with more than two concepts is small compared with the total number of images [10].

In Fig. 11, we compare our framework for image annotation to the following methods: Bottom-Up Score Fusion (*BUSF*) [4], Top-Down Classifiers Voting (*TDCV*) [4] and Hierarchy of SVM (*H-SVM*) [32]. As it can be seen in this figure, our multi-stage reasoning framework for image annotation outperforms on all classes comparing

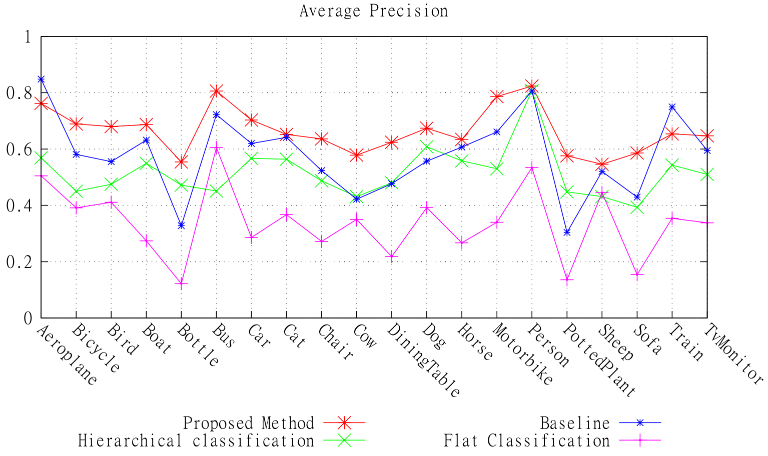


Fig. 10 Comparison of our method for image annotation with: a flat classification method, a hierarchical classification one, and the baseline method. Comparison is performed on VOC'2010 dataset

to the other ones. Please note that this comparison was performed using the same experimental setup, i.e. the same training/validation sets from the VOC'2010 dataset and the same visual representation of images. Therefore, it is clear that the proposed multimedia ontology and the proposed framework for reasoning about the consistency of image annotation allow achieving a significant improvement in the image annotation accuracy. These results also put into evidence the effectiveness of using explicit knowledge models, such as ontologies, for achieving semantically relevant image annotation.

In Table 3, we compare our multi-stage reasoning framework for image annotation to the methods of [47] and [45] on Pascal VOC'2009 dataset. In [47], the authors proposed a method for image classification using local visual descriptors

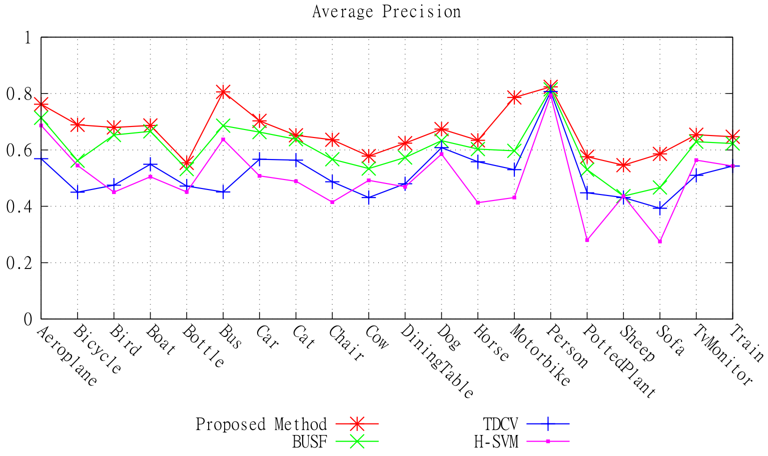


Fig. 11 Comparison of our framework for image annotation to previous work on Pascal VOC'2010 dataset. Our approach outperforms on all classes comparing to the other ones

Table 3 Comparison of our method for image annotation with the ones of [47] and [45] on Pascal VOC'2009 dataset

	Proposed method (AP)	[47] (AP)	[45] (AP)
Aeroplane	82.2	87.1	87.7
Bicycle	74.1	67.4	67.8
Bird	69.2	65.8	68.1
Boat	64.5	72.3	71.1
Bottle	52.1	40.9	39.1
Bus	80.4	78.3	78.5
Car	70.1	69.7	70.6
Cat	61.7	69.7	70.7
Chair	63.8	58.5	57.4
Cow	62.7	50.1	51.7
Dining_table	68.9	55.1	53.3
Dog	63.2	56.3	59.2
Horse	62.7	71.8	71.6
Motorbike	76.1	70.8	70.6
Person	83.2	84.1	84.0
Potted_plant	57.1	31.4	30.9
Sheep	64.4	51.5	51.7
Sofa	58.1	55.1	55.9
Train	72.8	84.7	85.9
Tv_monitor	66.7	65.2	66.7
AP on all concepts	67.7	64.29	64.6

and their spatial coordinates. Their method consists in performing first a nonlinear feature transformation on local appearance descriptor, termed as super-vector, which exploits the residual vector information obtained from the vector quantization (VQ). These descriptors are then aggregated to form image-level feature vector. The image-level feature vector is finally fed into a classifier to perform image classification. In [45], an efficient sparse coding algorithm with a mixture model is proposed and which is assumed to work with much larger dictionaries that often offer higher classification performances. The mixture model softly partitions the descriptor space into local sub-manifolds, where sparse coding with a much smaller dictionary can fast fit the data. As illustrated in Table 3, our approach performs better than the other ones and achieves a gain of 3.41 % compared to the method of [47] and a gain of 3.1 % compared to the method of [45]. This result is promising especially because we did use only the half of the training set for training our classifiers and the other images for evaluating our approach, since we did not dispose of the testing set. We also wish to recall that we have included in our evaluation the images and the concepts marked as difficult, which are ignored in the challenge because they are considered as difficult to recognize. For instance, in the third example of Fig. 8, we can easily observe a "Dining_table" in the illustrated image. However, "Dining_table" is marked as difficult in the ground-truth of this image in the VOC'2009 challenge, and thus it will not count for computing the average precision of this concept. In our evaluation, we included these concepts, i.e. if they are not detected they will count as false negative. Furthermore, the scope of our paper was to study the potential of adding contextual and spatial information into the image annotation process through the use of ontology and ontological reasoning. Thus, we have focused our contribution on these points and we did not seek to implement a very efficient image descriptor since this is not the aim of our paper. Accordingly, the obtained results can be further improved as for example by incorporating other image features.

Finally, we want to highlight that some images in the VOC dataset are badly annotated. For instance, in the third example of Fig. 8 we can distinguish a bottle partially hidden by a vase and a potted flower in the background of the image. However, these concepts (i.e. "Bottle" and "Potted_plant") are missing in the ground-truth of this image. Thus, despite that our method succeeded to recognize these concepts, they counted as false positive detections in the evaluation of our method since they are missing in the ground-truth. For the second example of Fig. 8, our method has detected the concept "Dining_table" which is absent from the ground-truth. However, the image depicts indeed a "coffee table" and therefore our prediction is semantically relevant, especially since the annotation vocabulary does not provide concepts such as "Table" or "Coffee_Table". In Fig. 12, we illustrate another image which is badly annotated in the dataset. Indeed, the ground-truth of this image contains only the concept "Person". However, the image depicts much more concepts: a bottle, chairs, tables, and screens. Our method has detected

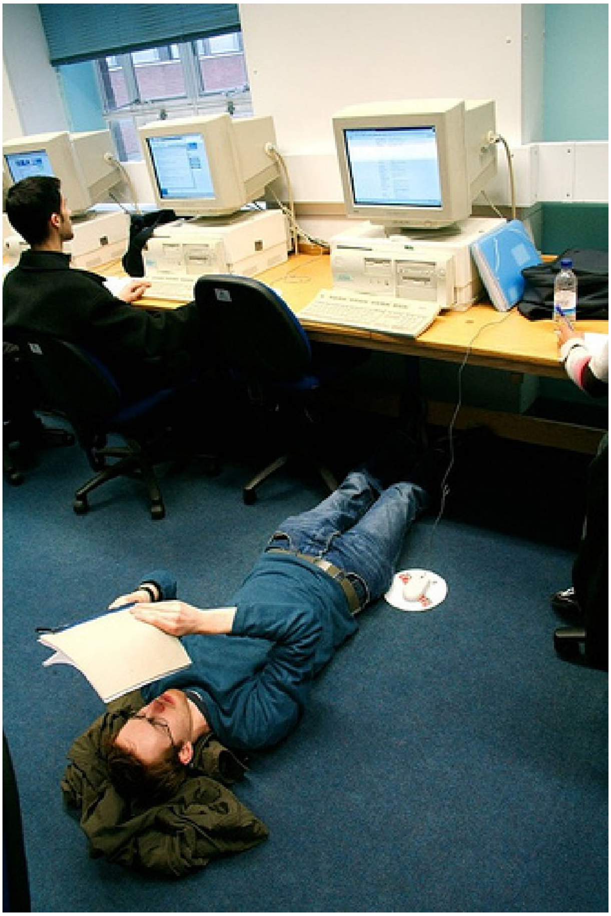


Fig. 12 An example of a badly annotated image in the VOC'2010 dataset. Ground-truth: Person. Annotation provided by our method: Bottle: 0.982, Chair: 0.281, Dining_table: 0.493, Person: 1.00, Tv_monitor: 0.333

these concepts, but according to the ground-truth these detections counted as false positives.

8 Discussion

The proposed methodology for building multimedia ontologies is original, and is useful for the modeling and the understanding of image semantics, i.e. identify and formalize the semantic relationships between image concepts. Indeed, the representation of our concepts and their semantic relationships are automatically extracted from image datasets, which provides an efficient modeling of image semantics and allows for extending our ontology at any time by mining new image datasets. Efficient modeling of image semantics means here: less sensitive to the subjectivity of human perception and less sensitive to the semantic gap.

Regarding the usefulness of our multimedia ontology for computer vision tasks, we propose in the following some usage scenarios. Let us consider an expressive amount of multimedia content, it is possible to extend our approach in order to model (or to learn), in a simple way, complex concepts by the mining of this multimedia content. For instance, let us suppose that we dispose of a well annotated image database and which is representative of the scenes from real life. It is obvious that when we find a '*Computer monitor*' in a given image, it is very likely to find a '*Mouse*' and a '*Keyboard*', and thus, these concepts will share a high co-occurrence confidence score. One can therefore use our proposed approach to define complex concepts, which are not previously included in the annotation vocabulary, based on the fuzzy role '*hasAppearedWith*' and the co-occurrence confidence score. Specifically, if the context of appearance of a set of concepts is sufficiently high (greater than a predefined threshold), therefore using their definition in WordNet we can find the common concept that connects them, and consequently define automatically this (complex) concept. To illustrate this proposal, here are some examples of defined concepts by the above described method:

Example 8 (Scenario 1: Defining complex concepts)

$$\langle Sitting\_room \equiv Sofa \sqcap Table \sqcap Television\rangle$$
$$\langle Beach \equiv Sea \sqcap Sand \sqcap Sky \sqcap \exists hasAppearedAbove(Sea, Sand) \sqcap$$
$$\exists hasAppearedBellow(Sea, Sky)\rangle$$
$$\langle Computer \equiv Screen \sqcap Keyboard \sqcap Mouse \sqcap \exists hasAppearedAbove(Screen,$$
$$Keyboard) \sqcap \exists hasAppearedRightOf(Mouse, Keyboard)\rangle$$

Another usage scenario consists in a knowledge-driven approach for image annotation using object detection. Indeed, one popular technique for identifying and localizing objects in an image is by the use of sliding-window object detection. It consists in defining a fixed-size rectangular window and applying a classifier to the sub-image defined by the window. The classifier extracts image features from within the window and returns the probability that the window bounds a particular object. The process is repeated on successively scaled copies of the image so that objects can be detected at any size.

So, let us suppose that one dispose of a multimedia database annotated with an average of 3,000 concepts, as for instance the SUN database [44]. Thus, we will dispose of 3,000 object detectors that will be performed on all images of the database and at different scales, which is computationally very expensive. The complexity of this task can be decreased significantly by the use of our multimedia ontology and the scenario defined in the following.

Example 9 (Scenario 2: A knowledge-driven approach for object detection.) Given a previously unseen image:

1. Apply progressively the detectors of the most frequent concepts (w.r.t '*hasFrequency*' concrete feature) in $\mathcal{KB}$, until a first concept $c_i \in \mathcal{C}$ is detected.
2. Query the ontology ($\mathcal{KB}$) for the most likely concept ($c_j \in \mathcal{C}$) to appear with c_i and its spatial location.
3. Apply the detector for c_j by delimiting the retrieving space according to the predicted spatial location. If it fails go to 2, else go to 4.
4. Query the ontology for candidate textual annotations with respect to the already detected concepts and their locations.
5. According to the decreasing confidence scores of these annotations, apply the detectors for the concepts of the selected annotation. If all concepts of the considered annotation are detected go to 6, else go to 4 (to select another annotation consistent w.r.t the already detected concepts).
6. Stop the processing and return the object detection result (i.e., the set of detected concepts and their spatial location) for the input image.

This usage scenario allows reducing significantly the complexity of the object detection process. In order to perform object detection, it requires performing much less detectors than the classical approach and targeting the detection zone according to the already detected concepts. Thus, it is clear that the proposed ontology is useful to effectively manage image processing tasks, and to efficiently perform image annotation. These usage scenarios will be addressed in our future work.

9 Conclusion

In this paper, we proposed a new approach to automatically build a fuzzy multimedia ontology dedicated to image annotation and interpretation. In our approach, visual and conceptual information are used to build a semantic hierarchy faithful to image semantics, and which will serves as a backbone of our ontology. The ontology is thereafter enriched with contextual and spatial information. Fuzzy description logics are used as a formalism to represent our ontology and to deal with the uncertainty and the imprecision of concept relationships. Some usage scenarios are then proposed to show the usefulness of the proposed ontology.

We subsequently proposed a new method for image annotation based on hierarchical image classification and a multi-stage reasoning framework for reasoning about the consistency of the produced annotation. An empirical evaluation of our approach on Pascal VOC'2009 and Pascal VOC'2010 datasets has shown a significant improvement on the average precision results.

References

1. Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) (2003) The description logic handbook: theory, implementation, and applications
2. Bannour H, Hudelot C (2011) Towards ontologies for image interpretation and annotation. In: Content-based multimedia indexing (CBMI'11)
3. Bannour H, Hudelot C (2012) Building semantic hierarchies faithful to image semantics. In: International conference on advances in multimedia modeling (MMM'12), pp 4–15
4. Bannour H, Hudelot C (2012) Hierarchical image annotation using semantic hierarchies. In: Proceedings of the 21st ACM international conference on information and knowledge management (CIKM'12), pp 2431–2434
5. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135
6. Bart E, Porteous I, Perona P, Welling M (2008) Unsupervised learning of visual taxonomies. In: Computer vision and pattern recognition (CVPR)
7. Bloch I (2005) Fuzzy spatial relationships for image processing and interpretation: a review. Image Vis Comput 23(2):89–110
8. Bobillo F, Straccia U (2011) Reasoning with the finitely many-valued lukasiewicz fuzzy description logic sroiq. Inform Sci 181(4):758–778
9. Carneiro G, Chan AB, Moreno PJ, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 29(3):394–410
10. Choi MJ, Lim J, Torralba A, Willsky A (2010) Exploiting hierarchical context on a large database of object categories. In: Computer vision and pattern recognition (CVPR), pp 129–136
11. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
12. Dasiopoulou S, Kompatsiaris I, Strintzis M (2009) Applying fuzzy DLs in the extraction of image semantics. In: Spaccapietra S, Delcambre L (eds) Journal on data semantics XIV. Lecture notes in computer science, vol 5880. Springer Berlin, Heidelberg, pp 105–132
13. Dasiopoulou S, Tzouvaras V, Kompatsiaris I, Strintzis MG (2010) Enquiring mpeg-7 based multimedia ontologies. Multimed Tools Appl 46:331–370
14. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Computer vision and pattern recognition (CVPR)
15. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2009) The PASCAL visual object classes challenge 2009 (VOC2009) results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html
16. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The PASCAL visual object classes challenge 2010 (VOC2010) results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html
17. Fan J, Gao Y, Luo H (2008) Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. IEEE Trans Image Process 17(3):407–426
18. Griffin G, Perona P (2008) Learning and using taxonomies for fast visual categorization. In: Computer vision and pattern recognition (CVPR)
19. Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing. Int J Hum-Comput Stud 43(5):907–928
20. Gupta A, Mannem P (2012) From image annotation to image description. Neural Inf Process 7667:196–204
21. Hauptmann A, Yan R, Lin WH (2007) How many high-level concepts will fill the semantic gap in news video retrieval? In: International conference on image and video retrieval (CIVR)
22. Hollink L, Nguyen G, Schreiber G, Wielemaker J, Wielinga B, Worring M (2004) Adding spatial semantics to image annotations. In: International workshop on knowledge markup and semantic annotation
23. Horridge M, Bechhofer S (2011) The owl api: a java api for owl ontologies. Semant Web 2(1): 11–21
24. Hudelot C, Atif J, Bloch I (2008) Fuzzy spatial relation ontology for image interpretation. Fuzzy Set Syst 159:1929–1951
25. Hudelot C, Atif J, Bloch I (2010) Integrating bipolar fuzzy mathematical morphology in description logics for spatial reasoning. In: European conference on artificial intelligence (ECAI), pp 497–502

26. Kompatsiaris Y, Hobson P (2008) Semantic multimedia and ontologies: theory and applications. Springer
27. Lavrenko V, Manmatha R, Jeon J (2003) A model for learning the semantics of pictures. In: Neural information processing systems. MIT, Cambridge
28. Li FF, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'05), vol 2. Washington, DC, USA, pp 524–531
29. Li LJ, Wang C, Lim Y, Blei DM, Li FF (2010) Building and using a semantivisual image hierarchy. In: Computer vision and pattern recognition (CVPR)
30. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recogn 40(1):262–282
31. Lowe DG (1999) Object recognition from local scale-invariant features. In: International conference on computer vision (ICCV)
32. Marszalek M, Schmid C (2007) Semantic hierarchies for visual object recognition. In: Computer vision and pattern recognition (CVPR)
33. Simou N, Tzouvaras V, Avrithis Y, Stamou G, Kollias S (2005) A visual descriptor ontology for multimedia reasoning. In: WIAMIS
34. Simou N, Athanasiadis T, Stoilos G, Kollias SD (2008) Image indexing and retrieval using expressive fuzzy description logics. Signal Image Video Process 2(4):321–335
35. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
36. Spaccapietra S, Cullot N, Parent C, Vangenot C (2004) On spatial ontologies. In: Brazilian symposium on geoinformatics
37. Stoilos G, Stamou GB (2007) Extending fuzzy description logics for the semantic web. In: Workshop on OWL: experiences and directions (OWLED)
38. Straccia U (2001) Reasoning within fuzzy description logics. J Artif Intell Res 14:137–166
39. Straccia U (2006) A fuzzy description logic for the semantic web. In: Sanchez E (ed) Fuzzy logic and the semantic web. Capturing intelligence, vol 1. Elsevier, pp 73–90
40. Straccia U (2010) An ontology mediated multimedia information retrieval system. In: Multiple-valued logic (ISMVL), pp 319–324
41. Straccia U (2012) Description logics with fuzzy concrete domains. In: Computing research repository (CoRR). arXiv:abs/1207.1410
42. Tousch AM, Herbin S, Audibert JY (2012) Semantic hierarchies for image annotation: a survey. Pattern Recogn 45(1):333–345
43. Wu L, Hua XS, Yu N, Ma WY, Li S (2012) Flickr distance: a relationship measure for visual concepts. IEEE Trans Pattern Anal Mach Intell 34(5):863 –875
44. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR). IEEE, pp 3485–3492
45. Yang J, Yu K, Huang T (2010) Efficient highly over-complete sparse coding using a mixture model. In: Proceedings of the 11th European conference on computer vision: part V, ECCV'10, pp 113–126
46. Yao B, Yang X, Lin L, Lee MW, Zhu SC (2010) I2t: Image parsing to text description. Proc IEEE 98(8):1485–1508
47. Zhou X, Yu K, Zhang T, Huang T (2010) Image classification using super-vector coding of local image descriptors. In: European conference on computer vision (ECCV)

Hichem Bannour received the B.S. and the M.S. degrees in computer science from University of Monastir, Tunisia and the PhD degree in computer science from the MAS (Applied Mathematics and Systems) laboratory of Ecole Centrale Paris, France, in early 2013. He is currently a postdoctoral fellow at the CEA-List Vision & Content Engineering laboratory, Atomic Energy Commission, France. His research interests include image semantics modeling, image annotation, multimedia information retrieval, image and data mining, statistical machine learning, and ontological engineering.

Céline Hudelot obtained her PhD in electrical and computer engineering from INRIA and the University of Nice Sophia Antipolis in 2005. She is an assistant professor (Maître de Conférences) at the MAS Laboratory (Applied Mathematics and Systems Research Laboratory) of ECP since 2006, in charge of the research axis on formal methods for semantic multimedia understanding in the MAS Laboratory. Her research interests include knowledge and ontological engineering for semantic image analysis, 2D and 3D image processing, information fusion, formal logics, graph-based representation and reasoning, spatial reasoning and machine learning.