Building and Using Knowledge Models for Semantic Image Annotation

Hichem Bannour

ECOLE CENTRALE DES ARTS
ET MANUFACTURES
"ECOLE CENTRALE PARIS"

PHD THESIS

in candidacy for the degree of

Doctor of Ecole Centrale Paris

Specialty: Computer Science

Defended by

Hichem Bannour

Building and Using Knowledge Models for Semantic Image Annotation

prepared at Ecole Centrale Paris, MAS Laboratry

defended on March 8^{th}, 2013

Jury:

Chairman:	Dr. Marcin Detyniecki	CNRS Paris, France
Reviewers:	Pr. Jean-Marc Ogier	University of La Rochelle, France
	Dr. Philippe Mulhem	CNRS Grenoble, France
Examiners:	Dr. Adrian Popescu	CEA Saclay, France
	Dr. Jamal Atif	University of Paris-Sud 11, France
Advisors:	Dr. Céline Hudelot	Ecole Centrale Paris, France
	Pr. Marc Aiguier	Ecole Centrale Paris, France

Building and Using Knowledge Models for Semantic Image Annotation

Hichem Bannour

A Dissertation
Presented at Ecole Centrale Paris
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Mathematics Applied to Systems
Advisors: Dr. Céline Hudelot and Pr. Marc Aiguier

March 2013

To my parents,
To my sisters and my niece *Yasmine*,
who have contributed to my work like nobody else.

Acknowledgements

Firstly, I would like to thank my advisor Dr. Céline Hudelot for offering me the opportunity to achieve my PhD within the MAS laboratory of Ecole Centrale Paris. Thank you Céline for your trust, the freedom you gave me to explore the topics that I was interested in, your valuable advices and your unconditional support. At your side I learned many things both scientific and personal. I also want to thank Pr. Marc Aiguier who co-directed my PhD work and who was always available for giving me priceless advices on my scientific career. I am also thankful to Marc and Céline for offering me the possibility to teach at Ecole Centrale Paris.

Afterward, I would like to thank all the committee members for accepting to review my modest work. Specifically, I'm grateful to my reviewers Pr. Jean-Marc Ogier and Dr. Philippe Mulhem, the chairman Dr. Marcin Detyniecki and the examiners Dr. Adrian Popescu and Dr. Jamal Atif for their relevant comments and questions which contributed to enhance this dissertation.

Then, I would like to acknowledge all the members of the MAS laboratory, with special thanks to Frederic Abergel, Nikos Paragios, Pascale Legall, Pascal Laurent, Gilles Faye, Iasonas Kokkinos, Marie-Aude Aufaure, Anirban Chakraborti and the others for their kindness, their support and with whom it was always a pleasure to discuss. I am also thankful to Annie Glomeron and Sylvie Dervin who were always concerned to provide me the necessary conditions for the completion of my thesis.

I could not forget to thank Pr. Bechir Ayeb who introduced me in the field of research and who has been often my inspiration source. Without his constant support, I could never have completed this work. I would like also to acknowledge Pr. Rahul Singh for the valuable discussions we had and for the interesting position he offered me to evolve in his lab in San Francisco. I hope that we could collaborate very soon.

I am also thankful to my colleagues and friends who have shared with me unforgettable moments within and outside the lab. Namely, Marc-Antoine Arnaud, Nesrine Ben Mustapha, Hassen & Hana Ben Zineb, Dung Bui, Clément Courbet, Emmanuelle Gallet, Nicolas James, Bilal Kanso, Sofiane Karchoudi, Adrian Maglo, Casio Melo, Adith Perez, Florent Pruvost, Rania Soussi, Olivier Teboul, Konstantin Todorov, Amel Znaidia, and the others. It was my pleasure to share with you lunches, table football and skiing. I am also thankful to Arunvady Xayasenh, Ka Ho Yim, Cédric Zaccardi, Géraldine Carbonel and Catherine Lhopital who have evolved with me within the UJ2CP (association of junior researchers of Ecole Centrale Paris).

I would close these acknowledgements by those who are dearest to me: my family. Specifically, I am grateful to my father Abdelaziz and my mother Zakia for providing

me all the needed support to succeed in this PhD work. I am also thankful to my sister Asma Bennour and my brother in law Taoufik Hnia who always provided me logistic support and shared with me joyful moments in Paris. I also thank my sister and private doctor Arij Bennour who was always present to push me forward. Finally, I thank my sister Khaoula, my sweetest niece Yasmine and my brother in law Amine Omri who always brought joy in my life.

Abstract

This dissertation aims at building and using knowledge-driven models in order to improve the accuracy of automatic image annotation. Currently, many image annotation approaches are based on the automatic association between low-level or mid-level visual features and semantic concepts using machine learning techniques. Nevertheless, the only use of machine learning seems to be insufficient to bridge the well-known semantic gap problem, and therefore to achieve efficient systems for automatic image annotation. Structured knowledge models, such as semantic hierarchies and ontologies, appear to be a good way to improve such approaches. These semantic structures allow modeling many valuable semantic relations between concepts, as for instance subsumption, contextual and spatial relationships. Indeed, these relationships have been proved to be of prime importance for the understanding of image semantics. Moreover, such structured knowledge models about high-level concepts enable to reduce the complexity of the large-scale image annotation problem.

In this thesis, we propose a new methodology for building and using structured knowledge models for automatic image annotation. Specifically, our first proposals deal with the automatic building of explicit and structured knowledge models, such as semantic hierarchies and multimedia ontologies, dedicated to image annotation. Thereby, we propose a new approach for building semantic hierarchies faithful to image semantics. Our approach is based on a new image-semantic similarity measure between concepts and on a set of rules that allow connecting the concepts with higher relatedness till the building of the final hierarchy. Afterwards, we propose to go further in the modeling of image semantics through the building of explicit knowledge models that incorporate richer semantic relationships between image concepts. Therefore, we propose a new approach for automatically building multimedia ontologies consisting of subsumption relationships between image concepts, and also other semantic relationships such as contextual and spatial relations. Fuzzy description logics are used as a formalism to represent our ontology and to deal with the uncertainty and the imprecision of concept relationships.

In order to assess the effectiveness of the built structured knowledge models, we propose subsequently to use them in a framework for image annotation. We propose therefore an approach, based on the structure of semantic hierarchies, to effectively perform hierarchical image classification. Furthermore, we propose a generic approach for image annotation combining machine learning techniques, such as hierarchical image classification, and fuzzy ontological-reasoning in order to achieve a semantically relevant image annotation. Empirical evaluations of our approaches have shown significant improvement in the image annotation accuracy.

Keywords:

Automatic Image Annotation, Hierarchical Image Classification, Multimedia Ontologies, Semantic Hierarchies, Knowledge-Driven Models, Ontological Reasoning, Fuzzy-Description Logics.

Résumé

Cette thèse vise à construire et à utiliser des modèles à base de connaissances pour l'annotation automatique d'images. En effet, la plupart des approches d'annotation d'images sont basées sur la formulation d'une fonction de correspondance entre les caractéristiques de bas niveau, ou de niveau intermédiaire, et les concepts sémantiques en utilisant des techniques d'apprentissage automatique. Cependant, la seule utilisation des algorithmes d'apprentissage semble être insuffisante pour combler le problème bien connu du fossé sémantique, et donc pour produire des systèmes efficaces pour l'annotation automatique d'images. L'utilisation des connaissances structurées, comme les hiérarchies sémantiques et les ontologies, semble être un bon moyen pour améliorer ces approches. Ces structures de connaissances permettent de modéliser de nombreuses relations sémantiques entre les concepts, comme par exemple, les relations de subsomption, les relations contextuelles et les relations spatiales. Ces relations se sont avérées être d'une importance primordiale pour la compréhension de la sémantique d'images. En outre, l'utilisation de ces modèles à base de connaissances structurées permet de réduire la complexité du problème de l'annotation d'images à grande échelle, i.e. sur des bases avec un grand nombre d'images et un nombre de concepts assez conséquent.

Dans cette thèse, nous proposons une nouvelle méthodologie pour la construction et l'utilisation de ces modèles à base de connaissances dédiés à l'annotation d'images. Plus précisément, nous proposons dans un premier lieu des approches pour la construction automatique de modèles de connaissances explicites et structurés, à savoir des hiérarchies sémantiques et des ontologies multimédia adaptées pour l'annotation d'images. Ainsi, nous proposons une approche pour la construction automatique de hiérarchies sémantiques. Notre approche est basée sur une nouvelle mesure "sémantico-visuelle" entre les concepts et un ensemble de règles qui permettent de relier les concepts les plus apparentés jusqu'à la création de la hiérarchie finale. Ensuite, nous proposons de construire des modèles de connaissances plus riches en terme de sémantique et qui modélisent donc d'autres types de relations entre les concepts de l'image. Par conséquent, nous proposons une nouvelle approche pour la construction automatique d'une ontologie multimédia qui modélise non seulement les relations de subsomption, mais aussi les relations spatiales et contextuelles entre les concepts de l'image. L'ontologie proposée est adaptée pour raisonner sur la cohérence de l'annotation d'images.

Afin d'évaluer l'efficacité des modèles de connaissances construits, nous proposons de les utiliser par la suite dans un cadre d'annotation automatique d'images. Nous proposons donc une approche, basée sur la structure des hiérarchies sémantiques, pour la classification hiérarchique d'images. Puis, nous proposons une approche générique, combinant des techniques d'apprentissage automatique et le raisonnement ontologique flou, permettant de produire des annotations d'images sémantiquement pertinentes.

Des évaluations empiriques de nos approches ont montré une amélioration significative de la précision des annotations d'images.

Mots clés :
Annotation automatique d'images, classification hiérarchique d'images, ontologies multimédia, hiérarchies sémantiques, modèles à base de connaissances, raisonnement ontologique, logiques de description floues.

Contents

List of Figures

List of Tables

List of Definitions and Examples

Introduction

Chapter 1

Introduction

Nowadays, image databases are becoming very large and of potential use in many areas, including public domain applications and also domain-specific applications. For instance, the public domain applications regroup internet services such as social networks[1], photo sharing[2,3], image and video monitoring, and more generally, information retrieval related services and image search engines. Specific-domain applications include medicine, aerial surveillance, audiovisual archiving and many others.

As the amount of these digital images increases, the problem of finding a desired image (or a subset of them) becomes critical. For example, in a distributed medical database (probably worldwide), given the image of a patient, it can be beneficial to find other images of the same modality, of the same anatomic region, and/or of the same disease that was already diagnosed [Müller et al., 2004]. Doubtless, this will have a great impact on the clinical decision-making process. Consequently, there is an important need for efficient techniques for storage, indexing and retrieval of multimedia information. In particular, image retrieval is still a big challenge despite 20 years of research.

The image retrieval field is carrying on the set of techniques/systems for browsing, searching and retrieving images from a large collection of digital images. Usually, such systems operate in two steps: i) image indexing: which could be defined as the process of extracting, modeling and storing the content of the image, the image data relationships, or other patterns not explicitly stored, and ii) image search: which consists in executing a matching model to evaluate the relevance of previously indexed images with the user query. Figure 1.1 illustrates the workflow of image retrieval systems. It is common sense that image search in these system is based on the same features (or modalities) than the ones used for image indexing.

This thesis deals with the problem of semantic image annotation. Specifically, we focus in this dissertation on how to model in an effective way the image content. Our approach is based on the building and the use of explicit and structured knowledge models in order to improve image annotation.

[1] More than 2300 images were loaded on Facebook every second during 2011 [Pixable Blog, 2011].
[2] Flickr (http://www.flickr.com/) exceeded 5 billion hosted pictures in late 2010.
[3] Picasa (https://picasaweb.google.com/) exceeded 7 billion hosted pictures in late 2010.

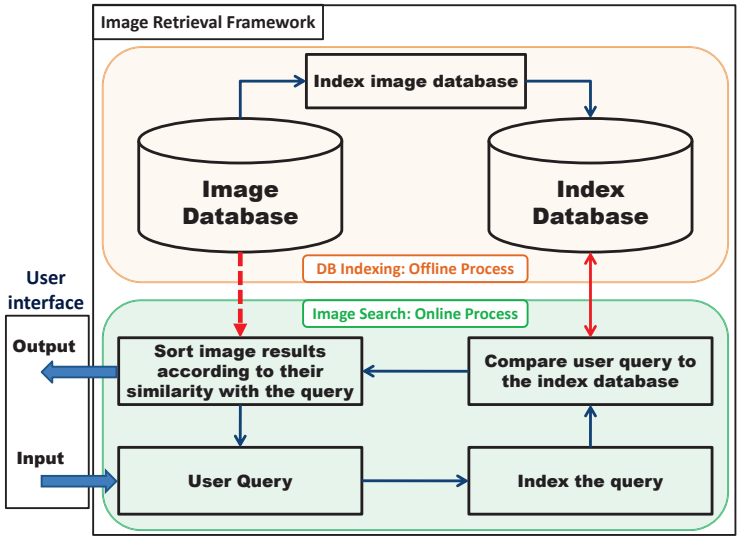


Figure 1.1: Workflow of image retrieval systems.

1.1 Context and Problem Statement

Initially, images were manually annotated by text descriptors (or tags) which are then used by an image retrieval system. This process is called '*iconography*'. '*Iconography*' means the description and interpretation of works in visual arts[4]. With respect to image retrieval field, 'iconography' means the process of human annotation of images. The iconography may be of prime importance for image retrieval systems since it provides valuable information about image content. It allows to dispose of the *visual* content description of an image, and very often of its *subjective*, *spatial*, *temporal* and *social dimensions* which are priceless for understanding image semantics. Online photo sharing systems, such as Flickr, Picasa and Getty Images[5], provide a valuable source of human-annotated photos. However, the iconography requires a considerable level of human labour, and it cannot be considered for large image databases [Liu et al., 2007].

Current approaches for automatic image indexing and retrieval can be classified into *text-based approaches* or *content-based approaches*, according to the used content (or modality) to index images - cf. Figure 1.2. In the text-based approaches, images are indexed by a set of text descriptors which are extracted from the surrounding context[6]. Nevertheless, although this paradigm is adopted by many current image

[4]Iconography: is the branch of art history which studies the identification, description, and the interpretation of the content of images: the subjects depicted, the particular compositions and details used to do so, and other elements that are distinct from artistic style. [Wikipedia-Iconography, 2012]

[5]`http://www.gettyimages.com/` Compared to Flickr and Picasa, Getty Images provides well detailed and less subjective image annotations.

[6]Surrounding context: We refer to surrounding context as the textual information surrounding a given image, such as user textual annotations, HTML tags, metadata or contextual text surrounding the image.

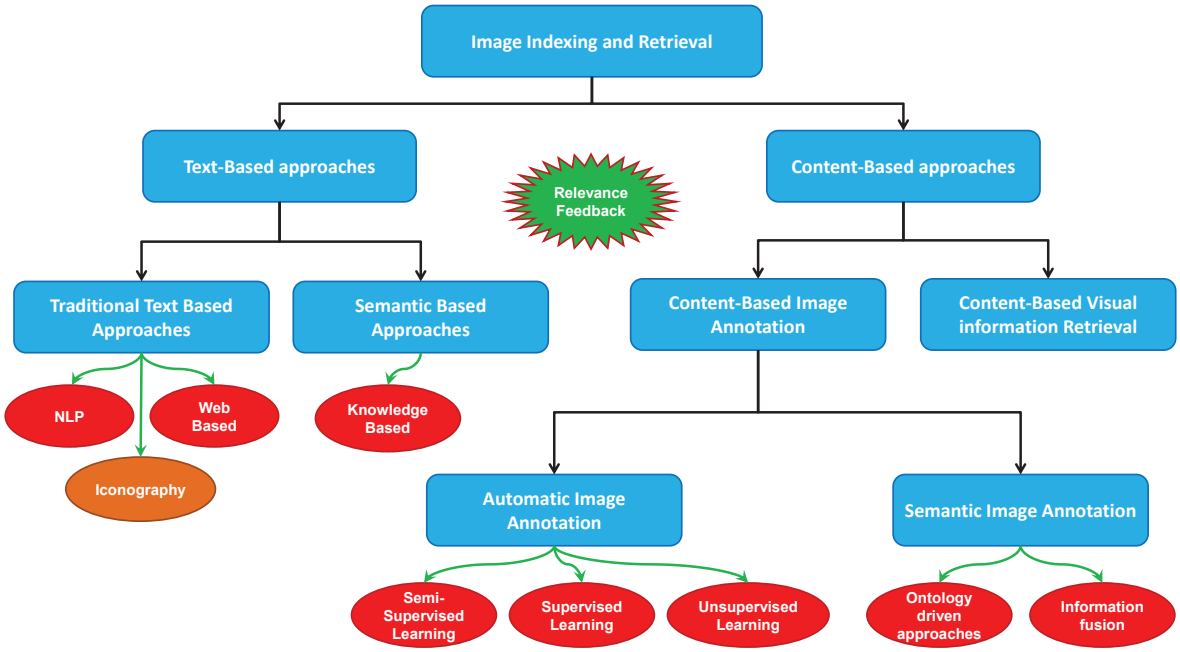


Figure 1.2: Proposed taxonomy of image retrieval approaches.

search engines (e.g. previous version of *Google Image Search*[7], *Bing Images* [8], etc.), it suffers from the following problems:

- As illustrated in Figure 1.3, the surrounding context is not always relevant with the image content, or sometimes only a small part of it describes its content. Consequently, the surrounding context is not always relevant for indexing images. Moreover, these methods do not consider the image content during the indexing process, and therefore there is no guarantee that the provided annotation is relevant with respect to image content.

- Text-based approaches are subject to the subjectivity of image semantics, also known as the problem of subjectivity of human perception[9] [Rui et al., 1998, Sethi and Coman, 2001]. This problem occurs when the one who provided the image description has a different background, and/or he want to express a different semantics in the image content, compared to the one who is searching the image.

- Currently, many voluminous image databases are generated daily without any surrounding context. For instance, the uploaded pictures of many Facebook users. Text-based approaches are unable to process these images.

[7] http://images.google.fr/

[8] www.bing.com/images/

[9] "*Different persons, or the same person under different circumstances, may perceive the same visual content differently. This is called human perception subjectivity*" [Rui et al., 1998]

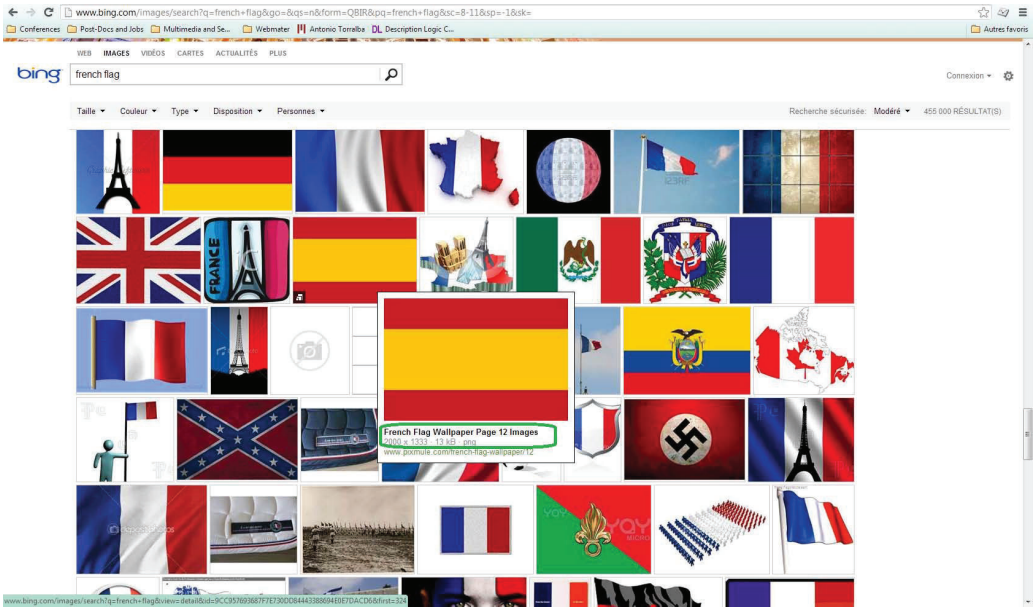


Figure 1.3: The surrounding context is not always relevant with respect to image content. Used query: *French Flag*, used search engine: *Bing Images*. As illustrated in this example, the search engine returns for instance the image of the Spanish flag because its surrounding context contains "French Flag Wallpaper Page 12 $\cdots$".

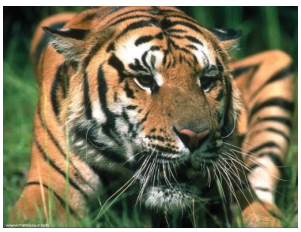
To overcome these problems, *Content-Based Image Retrieval* (CBIR) was introduced in the early 1980s. In this type of approaches, images are indexed and retrieved using their content. In particular, many approaches have been proposed to index images by their visual content, modeled by a set of low-level or mid-level visual features [Smeulders et al., 2000]. Nevertheless, although several sophisticated algorithms have been proposed to extract and to store efficiently these visual features, it was shown that these approaches are unable to model the semantic content of images. This is known as the **semantic gap** problem, which is defined as *"the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation"* [Smeulders et al., 2000].

In Figure 1.4, we illustrate an example of the semantic gap problem. Indeed, the images (a) and (b) have similar color histograms (i.e. a similar visual appearance) but depict different concepts, whereas, the images (a) and (d) have different color histograms but depict same concepts, i.e. {"House", "Forest", "Cabin"}.

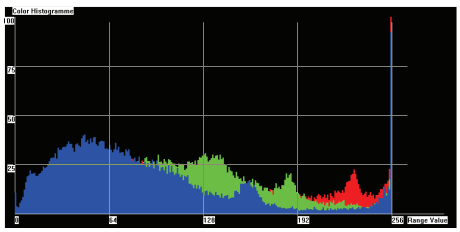
Hence, the main issue in image retrieval field is how to relate the visual content of images (low-level or mid-level visual features) to its semantic content (or high-level concepts). Consequently, the development of new approaches allowing to narrow[10] the semantic gap has been a core research topic since already ten years. In particular, **automatic image annotation** or *content-based image annotation* has been a very active research domain since a decade. Indeed, automatic image annotation can be

[10]There is a general consensus in which the semantic gap would not be bridged for the near future, and it is safe to use 'narrow' (instead of 'bridge') the semantic gap.

(a) House, forest, cabin.

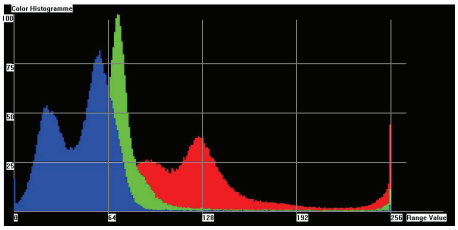


(b) Tiger, head, grass.



(c) Color Histogram for 1.4(a) and 1.4(b)

(d) House, forest, cabin.



(e) Color Histogram for 1.4(d)

Figure 1.4: Semantic gap problem. Images (a) and (b) have similar color histograms but different meanings. Images (a) and (d) have different color histograms but the same meaning.

defined as the process of associating to a previously unseen image a textual description (often reduced to a -set of- semantic keywords or high-level concepts) which depicts its content. Once again, a wide number of approaches have been proposed for automatic image annotation. These approaches have focused on the problems of recovering effective image descriptors on one hand, and on the other hand on developing efficient machine learning algorithms in order to provide robust methods that allow mapping visual features into semantic concepts [Duygulu et al., 2002, Barnard et al., 2003, Lavrenko et al., 2003, Djeraba, 2003, Carneiro et al., 2007]. These approaches are also known as image classification or image categorization approaches.

However, these approaches seem to be also insufficient to bridge the semantic gap, specifically when dealing with broad content image databases [Liu et al., 2007, Deng et al., 2010]. Specifically, automatic image annotation approaches are sensitive to the dimension of the physical representation space of image features and to the size of the semantic space used for describing image concepts. Moreover, these approaches face the scalability problem when the concept number is high and depend on the targeted datasets as well [Hauptmann et al., 2007]. Furthermore, automatic image annotation approaches are subject to many types of uncertainty introduced by machine learning algorithms:

- Uncertainties in input data, i.e. images are subject to noise, outliers, and errors, but also the representation (projection in a given space) of these data introduces some form of uncertainty. This refers to the sensory gap [Smeulders et al., 2000]. The noisy data, in theory, can sometimes have a positive effect on the generalization behavior of the machine learning algorithm, since it is forced to develop some form of invariance and to abstract from the

noise [Hammer and Villmann, 2007]. But in practice, noisy data usually have a decreasing impact on the method efficiency.

- Uncertainties in model parameters [Solomatine and Shrestha, 2009]. Indeed, machine learning algorithms are sensitive to parameter setup. The tuning of these parameters is not an easy task, and is not always relevant for all the considered classes.

- Uncertainties due to the lack of a perfect model structure. Even a representation of data in a high-dimensional space does not guarantee that the data is separable (i.e. find a boundary that separates the output classes).

As a consequence, decision functions should provide a confidence value which allows to judge the certainty or the belief of the output.

Moreover, these approaches do not adapt to the user background, nor to the specific semantics of the information sought by him in an image retrieval system, i.e. the meaning sought by a given user in the image content with respect to his background and within (or not) a specific domain application. Besides, in these approaches, the image semantics is often limited to its perceptual manifestation. Nevertheless, as shown in Figure 1.5, the image semantics is not (always) explicitly stored in its visual content, but it is rather an emergent property of the interaction of the user and the image content [Santini et al., 2001]. Very often, it also depends on prior knowledge on the domain and on the context of use of the image [Hudelot, 2005, Alm, 2006]. Consequently, **we consider that semantic image annotation is a knowledge-based process.**

In Figure 1.5, we have illustrated these ideas with some examples where the meaning is not explicitly encoded in the visual appearance of images. For instance, only people with deep knowledge in astronomy may know that the image 1.5(g) depicts the Crab Nebula which is a supernova remnant and pulsar wind nebula in the constellation of Taurus. Also, understanding that the image 1.5(a) depicts love requires a cognitive analysis based on knowledge about human behaviors. Also, the understanding of the image 1.5(e) requires a comprehensive image analysis and a deep cognitive process to figure out that it depicts the global warming. Indeed, one can observe in this image that the ice floe is small and it is crowded by penguins. Figure 1.5(f), which illustrates a fractal Broccoli, is usually interpreted by common people as depicting a Broccoli, but many mathematicians and physicists will only see in it a fractal equation of the form $D = \lim_{S \to 0} \frac{\ln(N)}{\ln(\frac{1}{S})}$, where N is the number of closed balls of diameter S needed to cover the object. So, image interpretation is subjective to the user background knowledge and to the user objective, and therefore automatic image annotation approaches should incorporate knowledge models in order to adapt to the requested semantics by the user in an image retrieval system. Finally, Figures 1.5(c) and 1.5(d) represent two different concepts but sharing the same name "Jaguar". This refers to the polysemy problem which is very common in information retrieval field. This problem requires to analyze the user query in order to understand which of both concepts the user is looking for, and consequently, **there is a need for semantic structures that allow to make explicit the semantics of concepts**.

(a) Love

(b) Family

(c) Jaguar (animal)

(d) Jaguar (car)

(e) Antarctic penguins on ice floes.

(f) Fractal Broccoli.

(g) The crab nebula.

Figure 1.5: Images and their semantic. These examples show that image semantics is not always straightforward and subtle. The picture (a) depicts Love (a man kissing a woman under a red umbrella), (b) depicts a family picture: a father carrying his daughter on his shoulders who is holding a soft toy in her hands, pictures (c) and (d) depict Jaguars, (e) depicts a global warming scene, (f) depicts a fractal broccoli, and (g) depicts the Crab Nebula which is the result of a supernova that occurred in 1054 AD. This image was taken by the Hubble telescope, and in the center is a pulsar, or a neutron star.

To sum up, automatic image annotation is a very promising and important issue to improve image search and retrieval. Nevertheless, most of the current approaches are still insufficient to satiate the user need due to the following problems:

- **Uncertainty** introduced by machine learning algorithms.
- **Scalability problems** in terms of the image number, the dimension of the representation space of image visual features, and the concept number (dimension of the semantic space).
- **Semantic gap problem**.
- **Subjectivity of image semantics** (these approaches do not adapt to the user background).
- **Sensitivity to the accuracy of the ground truth of the learning dataset**.
- **Polysemy problem**.

Therefore, automatic image annotation is still a challenging problem and current approaches need to be improved. In particular, some recent work have

proposed to model and to use explicit knowledge about image context in order to answer to the above mentioned problems, as for instance [Hollink et al., 2004, Maillot et al., 2004, Popescu et al., 2007, Fan et al., 2008a, Hudelot et al., 2008, Tousch et al., 2008, Wu et al., 2008, Simou et al., 2008, Dasiopoulou et al., 2009, Fan et al., 2009, Straccia, 2010, Li et al., 2010]. We refer to these approaches as **semantic image annotation**, which can be defined as the set of methods mainly based on the use of explicit semantic structures, such as semantic hierarchies and ontologies, in order to narrow the semantic gap. However, semantic image annotation approaches are currently facing the following shortcomings:

- The availability of knowledge structures modeling the image semantics, or failing that, their building or their learning. Indeed, currently there are a number of available knowledge resources which have been used for image annotation, like WordNet [Fellbaum, 1998] or Wikipedia[11], and also multimedia knowledge resources such as LSCOM [Naphade et al., 2006], ImageNet [Deng et al., 2009], and so on. However, these knowledge sources are not completely suitable for image annotation, since they are built using a conceptual specification which not necessarily reflects image semantics, i.e. they do not consider image content.
- The lack of an unified model for knowledge representation and reasoning.
- Only few approaches have considered this knowledge in a formal framework and exploited reasoning on it.
- At last, unlike other domains, multimedia knowledge extraction is still in its infancy, i.e. much efforts are needed to assess the strength and the limits of these approaches on real data applications.

1.2 Objectives

In this dissertation, our objective is to make advances in the field of semantic image annotation by the proposition of new knowledge-based frameworks dedicated to image annotation. Indeed, the achieved contributions in this thesis were motivated by the following assumptions:

- Image semantics is not fully included in the image itself, and therefore the use of explicit structured-knowledge models is necessary for performing accurate image annotations.
- Existing knowledge models are not fully adequate for image annotation. These are either: i) very generic knowledge resources (commonsense ontologies such as WordNet and Wikipedia) which do not reflect image semantics, or ii) built automatically using available data, but often only a single piece of information on images is considered (for instance, visual information in [Sivic et al., 2008] or conceptual information in [Snoek et al., 2007]). Therefore, a new paradigm

[11] http://www.wikipedia.org/

for building knowledge models should be adopted and which should incorporate the different image modalities (e.g. visual, conceptual, contextual, spatial, etc.) in order to be relevant with respect to image semantics.

- The building of these knowledge models should be performed automatically, using data mining techniques, in order to reduce the scalability problem of knowledge building and to achieve a consistent representation of image semantics.

Consequently, the proposed contributions involve capturing and modeling suitable knowledge models about the image semantics, and thereafter building effective tools to exploit this knowledge in order to improve the image annotation process. The overall goal remains to narrow the semantic gap using the available visual features of images and relevant domain knowledge in order to support the varied search categories, ultimately to satiate the user expectations. Specifically, we address the following issues:

1. **The building of explicit knowledge models**, in an automatic and effective manner, for the purpose of image annotation.
2. **The use of these knowledge models** with the aim of improving the image annotation accuracy, through:
   (a) An hierarchical image classification framework.
   (b) A fuzzy ontological reasoning framework.

Therefore, we believe that the outcomes of our work will allow to address the following problems:

- Narrowing the semantic gap, i.e. providing effective tools that allow mapping the visual features of images into semantics concepts.
- Reducing the uncertainty introduced by machine learning algorithms by the use of fuzzy ontological reasoning with the aim of assessing the consistency of image annotations.
- Reducing the following problems: i) polysemy, ii) subjectivity of image semantics, and ii) sensitivity to the accuracy of the ground truth, by providing explicit semantic structures allowing the reasoning on concepts semantics.
- Reducing the scalability problem of automatic image annotation by the proposition of a hierarchical image classification method, which scales well with large databases.

1.3 Contributions

As aforementioned, the aims of this thesis are twofold. As a first step, we have focused our contributions on providing effective tools for the automatic building of structured and explicit knowledge models dedicated to image annotation. This is intended to propose structured knowledge models that are relevant with respect to image semantics. Subsequently, we have concentrated our contributions on the effective use of these knowledge models in order to improve image annotation. Our final purpose is to narrow the semantic gap, i.e. decrease the gap between the visual content of images and their meaning (image semantics). The achieved contributions in this dissertation are detailed in the following subsections.

1.3.1 Building Explicit and Structured Knowledge Models for Image Annotation

Our first contribution in this dissertation deals with the automatic building of semantic hierarchies for the purpose of image annotation. Indeed, as it is detailed in Chapter 3 - Section 3.2.4, our standpoint is that a semantic hierarchy dedicated to image annotation should take account of the *visual* information, the *conceptual* information and the *contextual* one. Consequently, we propose a new image-semantic measure which allows estimating the semantic similarity between image concepts, and which incorporates the aforementioned information in order to provide a meaningful measure of image semantics. Subsequently, a new methodology, based on the previously proposed measure and on a set of effective rules, is proposed in order to automatically build semantic hierarchies suitable for image annotation. This contribution is detailed in Chapter 3, and has served for the following publications [Bannour and Hudelot, 2012a, Bannour and Hudelot, 2012b].

Thereafter, we propose to go further in the modeling of image-concepts relationships, and consequently the modeling of image semantics. We propose therefore a new approach for building an ontology of spatial and contextual information suitable for reasoning about the consistency of image annotation. Our approach uses visual and conceptual information in order to build a semantic hierarchy that will serve as a backbone of our ontology. Contextual and spatial information about image concepts are afterwards 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. This work is the subject of Chapter 4, and is published in [Bannour and Hudelot, 2013a].

Figure 1.6, illustrates the general workflow of the proposed approaches for building structured knowledge models dedicated to image annotation.

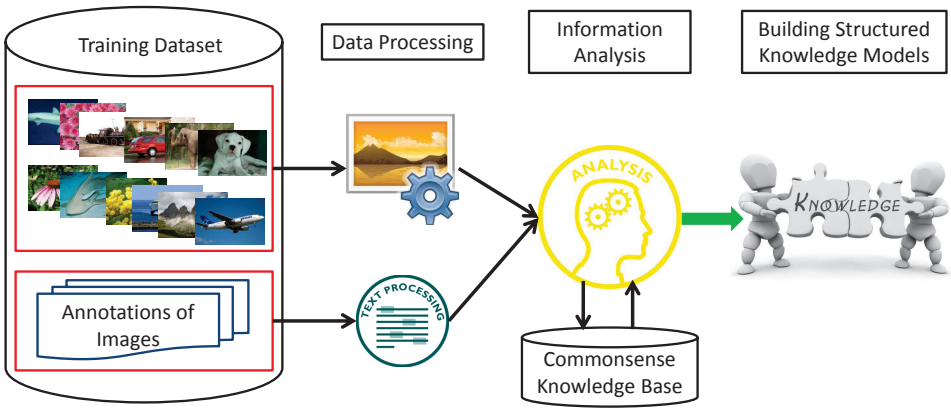


Figure 1.6: From image data to structured knowledge Models. Workflow of our approaches for building structured knowledge models dedicated to image annotation.

1.3.2 Using Structured Knowledge Models for Image Annotation

In order to make an efficient use of the built structured knowledge models, we propose in this dissertation to investigate the contribution of: i) semantic hierarchies, and ii) DL ontologies and ontological reasoning, for image annotation. The overall workflow of our methods is illustrated in Figure 1.7. Given a set of previously unseen (unlabeled) images and a structured knowledge model dedicated to image annotation, our approach allows to predict the set of concepts from the annotation vocabulary that are relevant with respect to the image content/semantics.

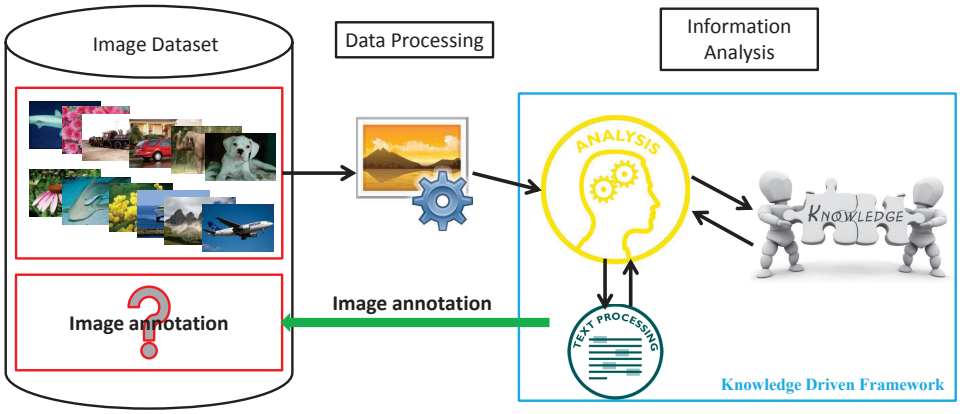


Figure 1.7: Using structured knowledge models for image annotation.

We propose therefore a new methodology, based on the structure of semantic hierarchies, to efficiently train hierarchical classifiers. Our method allows decomposing the problem into several independent tasks and therefore it scales well with large databases. We also propose two methods for computing a hierarchical decision function that serves to annotate new image samples. The former is performed using a top-down classifiers voting, while the second is based on a bottom-

up score fusion. This methodology is detailed in Chapter 5, and is published in [Bannour and Hudelot, 2012c, Bannour and Hudelot, 2013b].

Finally, we propose a novel 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 relevant decision on the belonging of a given image to the set of concept classes. The ontological reasoning is performed using the previously built multimedia ontology (in Chapter 4), and involves reasoning about contextual and spatial information of image concepts. This framework is introduced in Chapter 6, and is published in [Bannour and Hudelot, 2013a].

1.4 Organization of this Dissertation

This dissertation is organized as follows.

In **Chapter 2**, we present a thorough survey on relevant research topics to the image retrieval domain. The covered topics include: image retrieval, automatic image annotation, hierarchical image classification and knowledge-driven approaches for image annotation. We also propose a thorough discussion of the related work in our contribution chapters, so as to put our proposals into perspective.

The main contributions of this dissertation start with **Chapter 3**. Indeed, we propose in **Chapter 3** a new approach for building semantic hierarchies suitable for image annotation. Our approach takes advantages of the different image modalities to compute an image-semantic measure between the different concepts of the annotation vocabulary. This measure is thereafter used to connect the concepts with a higher semantic relatedness till the building of the final hierarchy.

Chapter 4 proposes to go further in the use of structured knowledge models for image annotation. Specifically, we address the following issues: how to extract and store knowledge about image context in an effective way, how to query the ontology to retrieve valuable information about image context, and how to perform reasoning using multimedia ontologies in order to reduce the uncertainty about image annotations. We therefore propose to automatically build multimedia ontologies by mining image databases to gather valuable information about image context. Thus, we show that ontologies are not necessarily limited to be defined by humans, as it was the case for most of existing approaches in the multimedia domain. Thereby, we reduce the scalability problem of ontology building.

In **Chapter 5**, we explore the contribution of semantic hierarchies for image annotation. A natural way to use semantic hierarchies is to operate within a framework for hierarchical image classification. Consequently, we propose a new approach based on the semantic hierarchy structure to efficiently train the hierarchical classifiers, and for computing an accurate hierarchical decision function which is thereafter used to annotate previously unseen images.

In **Chapter 6**, we explore the use of formal ontologies and fuzzy ontological reasoning for improving image annotation. Our intent is to reduce the uncertainty of image annotation introduced by the machine learning algorithms. Therefore, we pro-

pose 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.

Finally, this dissertation is concluded in **Chapter 7** with a feedback on our contributions and a discussion on the directions that can be borrowed by the presented research topics.

Chapter 2

State of the Art on Image Annotation: from Classical to Knowledge-Driven Approaches

Contents

2.1 Introduction

As introduced in the previous chapter, image indexing and retrieval has been a very active research domain since two decades, hence many approaches have been proposed to solve this problem. These different approaches can be classified in several different ways and from different points of views, as for example, the application domain, the indexing technique, the used content (modality) for image description, and so on. In Figure 1.2, we have proposed a taxonomy to classify image retrieval approaches, based on the used modality for describing image content. In particular, as introduced in Section 1.1, we consider two main families of approaches:

i) *Text-based approaches*: which do not consider the image content during the indexing process, but rely on the only surrounding textual information (HTML tags, metadata, contextual text surrounding the image, etc.), and

ii) *Content-based approaches*: which are basically based on the use of image content (visual and semantic content) in order to index and retrieve images.

The scope of this dissertation is content-based image retrieval, and in particular content based-image annotation. Thus, as illustrated in Figure 2.1, we focus in this chapter on this family of approaches, and specifically on automatic image annotation approaches and semantic image annotation approaches.

This chapter is structured as follows. Section 2.2 presents a brief introduction to the image annotation problem. In Section 2.3, we introduce what is meant by image semantics, and we put in evidence the different challenges in order to achieve effective image annotation. In Section 2.4, we propose a review of classical image annotation approaches and we discuss their limitations. Section 2.5 proposes, as a first contribution, a state of the art on the use of implicit/explicit knowledge-based models to improve the annotation process. Section 2.6 proposes a general discussion about existing approaches for automatic image annotation. This chapter is concluded in Section 2.7.

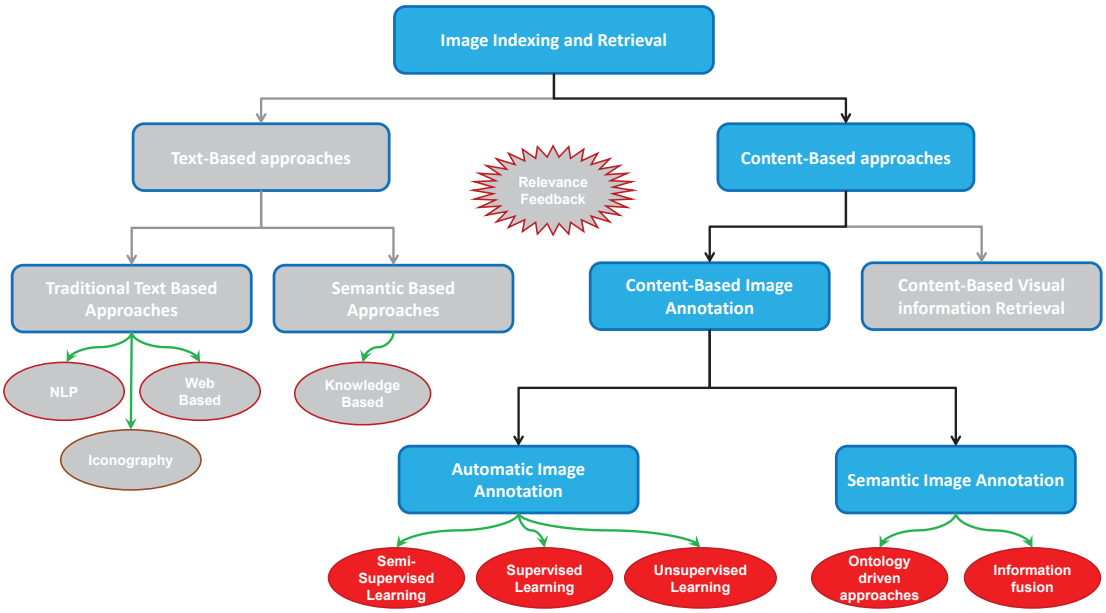


Figure 2.1: Proposed taxonomy of image indexing and retrieval approaches. The scope of our work is in blue boxes.

2.2 The Image Annotation Problem

Automatic image annotation can be defined as the process of automatically assigning a text description (often reduced to a set of semantic keywords) to a digital image through a computational model (or a computer system). Automatic image annotation is usually used in image retrieval systems in order to index and retrieve images of interest from a large database. Very often, this task is regarded as an image classification problem (usually a multiclass classification problem), and is involved by the following steps:

1. Gather a training image dataset consisting of a set of images with their textual annotations. These textual annotations generally consist of a set of high-level concepts (semantic concepts) depicting image content, and are usually called the ground truth. The set of all concepts composed the annotation vocabulary.

2. Build a computational model enabling to find a correspondence model between the low-level or mid-level representations of images and the concepts of the annotation vocabulary.

3. Test the system and adjusting the parameters of the computational model.

Problem Formalization. In a formal way, automatic image annotation can be defined as follows:

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 at least a set of concepts $\{c_j \in \mathcal{C}, j = 1..n_{i_i}\}$ associated with a given image $i_i \in \mathcal{DB}$. According to the used dataset, these textual annotations may also contain other pieces of information, such as information about depicted concepts (bounding box, truncated or not, etc.), image source, user tags and so on.

The objective of automatic image annotation is to build a computational model that enables to associate a set of concepts $\{c_j \in \mathcal{C}, 1 \leq j \leq \mathcal{N}\}$ to any given image $i_i \in \mathfrak{I}$. The overall goal is to extend this computational model to previously unseen images (i.e. $\forall\ i_x \notin \mathcal{DB}$) in order to provide them a textual description. Figure 2.2 illustrates the aim of automatic image annotation by an example.

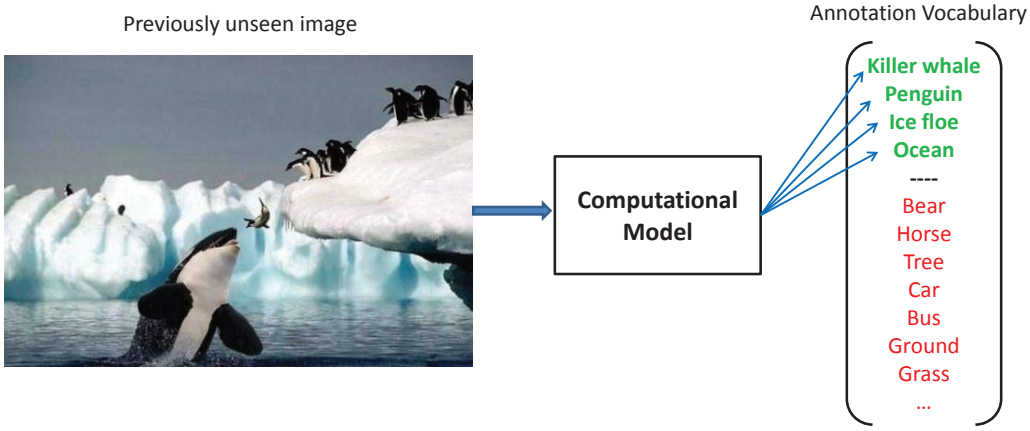


Figure 2.2: Example of automatic image annotation.

Terminology. For a sake of clarity, we define in this section the used terms and their meaning. Thereby, we refer to:

- *Concept*: as a label (or a word) with a precise semantics (a given meaning) used for annotating images. For instance, in Figure 2.2 "Killer Whale" and "Penguin" are concepts. Furthermore, according to our definition a label may correspond to several concepts, as for instance in Figure 1.5, where concept "Jaguar" refers to an animal (c), and to a car (d).

- *Annotation vocabulary*: as the set of concepts used for annotating the images of a given database.
- *Image features*: as a set of low-level or mid-level descriptors used for representing the visual content of an image.
- *Visual similarity*, *conceptual similarity*, *contextual similarity* as follows:

Definition 1 (Visual similarity)**.** *The visual similarity is a measure reflecting the similarity between concepts with respect to their visual appearances.*

Definition 2 (Conceptual similarity)**.** *The conceptual similarity is a measure reflecting the semantic relatedness between two concepts from a linguistic and/or a taxonomic point of view. With respect to image retrieval field, this measure is usually computed using textual information sources (such as the surrounding context, user tags, etc.), or using commonsense ontologies (such as WordNet [Fellbaum, 1998], Wikipedia*[1]*, etc.).*

Definition 3 (Contextual similarity)**.** *The contextual similarity is a measure that enables to group the concepts that share a same context. With respect to image annotation, we define this context as the context of appearance (co-occurrence) of concepts in the image database.*

2.3 Image Semantics

Many recent image annotation approaches have been proposed to narrow the semantic gap problem defined by [Smeulders et al., 2000] as the lack of coincidence between the low-level image descriptions using visual features and the richness of human semantics, i.e. "the interpretation that the image data have for a user in a given situation". According to this definition, images get their semantic meaning as an act of image interpretation or understanding, and it is consequently difficult for an image retrieval system to discern the meaning sought by a user when he is searching for a particular image. This is not a new idea and as recalled by [Boucher and Le, 2005], the fundamental work of [Treisman and Gelade, 1980, Marr, 1982, Biederman, 1987] have previously carried this insight.

Nevertheless, even if the notion of image semantics seems to be important for image retrieval related tasks, it is still vague and can vary according to the different approaches. Indeed, to our knowledge, only few work have tried to provide a precise definition of the notion of image semantics. Therefore, we propose in this section a tentative to define, through some examples, this notion of image semantics according to our objectives. Then, we relate our definition to the existing ones that have been sparsely defended in the literature.

[1] http://www.wikipedia.org/

2.3.1 Our Standpoint on Image Semantics

Semantics is the study of meaning. In linguistics, it is also the study of the meaning or the interpretation of a word, sentence, or other language form. With respect to the image retrieval field, we can define it as the meaning sought by the user in the image content. Figure 2.3 illustrates some images and their associated semantic meaning.

Definition 4 (Image Semantics)**.** *With respect to the image retrieval field, image semantics can be defined as the meaning sought by the user in the image content.*

As illustrated in Figure 2.3, the intended meaning may be obvious, intuitive and explicit which make it easy to access. However, this meaning is also very often implicit, induced and subtle making it difficult to discern, specifically through a multimedia information retrieval (MIR) system. For instance, in the images 2.3(a) and 2.3(b), how one can define a spectacular jump? How an image retrieval system can model or access this information? As well in the images 2.3(f) and 2.3(g), what is the difference between a race track and a road? How machines can infer this information directly from image pixels? And finally, in Figures 2.3(c), 2.3(d) and 2.3(e), the global warming depicted in these images is subtle. Indeed, we can see in the image 2.3(d) a polar bear who has trouble standing on a tiny ice, and in the image 2.3(c) a piece of ice overcrowded by penguins. Based on their background knowledge, humans know that the global warming is responsible of the arctic ice melting. Therefore, humans can infer from these images that the survival of the Antarctic animals, such as penguins and polar bears, are threatened by the ice melting caused by the global warming.

Consequently, our first view on **image semantics** is that it **is not fully, nor explicitly stored in the image pixels**, and it is usually hard for a machine to access the image semantics using only image features. Indeed, as stated in [Santini et al., 2001], image semantics is rather an *emergent property of the interaction of the user and the image content*. We can therefore conclude that the image interpretation process requires often a reasoning mechanism over the detected objects in the image, which is usually based on cognition and on the past experiences.

Accordingly, in order to match user expectations in an image retrieval system, it is important to provide **tools that allow building and using knowledge models representing the image context** (and consequently the image semantics). These will allow reproducing the reasoning process in order to deduce a more consistent image annotation/interpretation. Specifically, we believe that these knowledge models should go further than the simple description of specific objects that may appear in images, and rather model the image context through the description of concepts and the semantic relationships between them. For instance, a good description of the image 2.3(d) is: a "polar bear" is *standing on* a "tiny ice". This description could probably allow inferring that this image depicts a global warming, since it is meaningless that a big bear is standing on a tiny ice unless he was obliged to be. Hence, if it was obliged to do so, it is possibly because large ice floes have melt. The reason of ice melting could therefore be intuitively deduced.

(a) A spectacular jump of a killer whale in an aquatic parc.

(b) A spectacular jump of a killer whale to eat penguins.

(c) Global warming threatens the survival of polar animals (Antarctic penguins on ice floes).

(d) Global warming threatens the survival of polar bears.

(e) Our survival is threatened by global warming.

(f) Cars mercedes-benz race tracks.

(g) Corvette racing wins at american Le Mans - inside track.

Figure 2.3: Images and their semantic. Images (a) and (b) depict a spectacular jump of a killer whale; images (c), (d) and (e) depict the global warming; and images (f) and (g) depict cars on a racetrack.

Let us look at a definition for the image semantics from the perspective of the theory of meaning.

Definition 5 (Image Semantics 2)**.** *Referring to the theory of meaning, image semantics can be defined as the meaning sought by the user in the image content* ***within a particular context****.*

According to the above definition, we can notice that the **context of searching** is very important for understanding the image semantics thought and sought by a user. For instance, depending on the observer's background knowledge, the image in Figure 2.4(a) can be interpreted as: i) a statue, ii) a statue of a Carthaginian general, iii) a statue of Hannibal, iv) a statue of Hannibal counting the signet rings of Roman nobles killed during the battle, statue by Sébastien Slodtz, 1704, Louvre. Also, the image in Figure 2.4(b) can be interpreted as: i) airplane, ii) military aircraft, iii) an aircraft of the Second World War, iv) the Swordfish, which is a two-place torpedo bomber used during the Second World War. All these interpretations are relevant with respect to the observer's background, his past experiences, and the semantics level sought by him in these images.

(a) Hannibal counting the signet rings of Roman nobles killed during the battle, statue by Sébastien Slodtz, 1704, Louvre.

(b) The Swordfish was a two-place torpedo bomber used during the Second World War.

Figure 2.4: An image may have different interpretations depending on the observer's background knowledge.

These examples enable us to put into perspective our two next views on image semantics: i) **image semantics is a multi-level paradigm**, and ii) **image semantics is context-sensitive**, i.e. it *depends on the user's objective* and *on his background knowledge*. Consequently, one of the major challenges of image annotation and image retrieval systems is to be able to extract these different levels of image semantics and to adapt themselves to the user's objective in order to be efficient and useful. For example, a user would like to go beyond a query like "I seek an image that visually looks like this one" or "an image that contains a car", and would prefer to be able to ask queries such as "find me an image that contains a spectacular jump of a killer whale" or "a figure which depicts the global warming" or "find me a picture depicting cars on a racetrack" - cf. Figure 2.3.

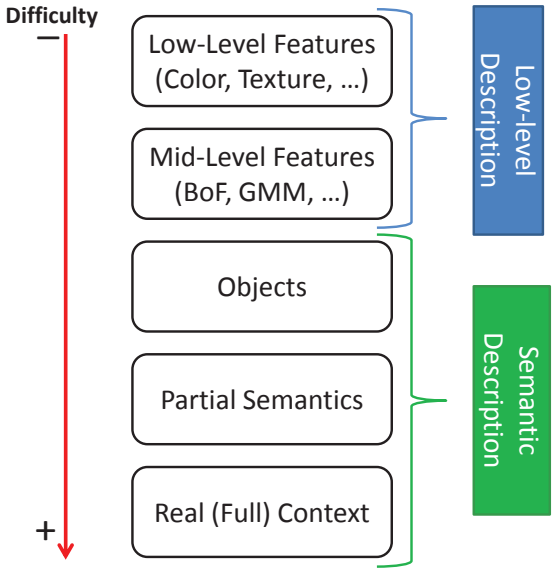


Figure 2.5: Levels of abstraction for multimedia content, and respectively image content.

Definition 6 (Image Semantics 3)**.** *Image semantics is a multi-level paradigm, i.e. there are several levels of semantics (or interpretation) for a given image and the major challenge of image retrieval systems is then to be able to extract such semantics from images and to adapt to the user background in order to be efficient and useful.*

In Figure 2.5, we illustrate the multi-level paradigm of image semantics. In particular, we distinguish three semantic levels: objects, partial semantics and full semantics. For instance, if we look at image (d) of Figure 2.4, the semantics at the object level could be {"Bear", "Iceberg"}, the semantics at the partial level could be "Polar Bear standing on a small iceberg" and the semantics at the full level could be "global warming threatens the survival of the polar bears". Therefore, we can notice that the difficulty of processing and extracting the semantics from images increases significantly according to the sought level of abstraction. Currently, most approaches for image retrieval deals with the first level of semantic content. These approaches target to provide efficient methods to learn semantics classes from visual image features. They allow therefore providing a description for images in the form of a set of keywords (e.g. Car, Building, Sky, Dog, etc.). Other approaches are trying to tackle the second level of semantics, i.e. the partial semantics level. These approaches aim at recognizing the concepts (objects) depicted in an image and subsequently try to identify the semantic relationships between them (e.g. a person eating an apple, penguins on ice floes, etc.). The last category of semantic description, i.e. the real context one, consists in identifying the full semantic context of a given image (e.g. see the description of images 2.3(c) and 2.4(a)). To our knowledge no attempts have focused on this kind of semantic description because of its difficulty, and the lack of a computational model allowing to model/infer such knowledge.

As a consequence, our last view on image semantics is that **contextual knowledge**, i.e. the semantic relationships between the different concepts, **are very important information to achieve image annotation at the partial or the full semantics level**.

In the next section, we show that these views on image semantics have been partially defended in the literature.

2.3.2 Related Work on Image Semantics

Following the idea of [Marr, 1982], some work have proposed multiple-level paradigms for image representation. [Jaimes and fu Chang, 2000] proposed a pyramidal indexing structure composed of 10 levels visual structure. The authors have considered that these levels can be divided into perceptual levels (i.e. the syntactic description of the image) and into semantic levels (i.e. the semantic description of the image). They differentiated 6 different semantic or conceptual levels consisting of: generic objects, generic scene, specific objects, specific scene, abstract objects and abstract scene. These 6 different conceptual levels can be related to the three different conceptual image description levels proposed by [Shatford, 1986, Shatford, 1994], i.e. generic, specific and abstract content which can be themselves considered as a generalization to generic image of the 'pre-iconographic', 'iconographic' and 'iconologic' levels of expression proposed by the art historian [Panofsky, 1972]. The authors have also stated that *a priori* knowledge is very useful to interpret images at high semantic levels, i.e. the higher the level is, the more knowledge is involved in the interpretation of images.

In [Enser and Sandom, 2003], the authors proposed to extend a model initially proposed by [Jörgensen, 1996, Jörgensen, 1998] and proposed to separate the image content into perceptual, generic-interpretive, specific-interpretive and abstract attributes. Interpretive attributes correspond to the semantic level and "are those which require both interpretation of perceptual cues and application of a general level of knowledge or inference from that knowledge to name the attribute" [Jörgensen, 1996].

In [Santini et al., 2001], the authors proposed a discussion on image semantics and its role in the design of image databases. In particular, the authors defended the interesting idea that the semantics is not an intrinsic property of the images but an emergent property which depends on the interaction between the user and the database. In particular, the authors emphasize the idea that the semantics depends on the contextual knowledge and on the cultural background of the user. Similar ideas are defended in [Hudelot et al., 2005] in the context of image understanding.

[Enser, 1993, Armitage and Enser, 1997, Eakins, 2002] have studied the nature of the image database query task. For instance, [Eakins, 2002] mentioned three kinds of queries in CBIR: 1) retrieval by primitive or perceptual features (e.g. find pictures sharing visual features with this query image), 2) retrieval by naming objects (e.g. find pictures of "Car"), and 3) retrieval by abstract attributes which involves reasoning about the objects and their relationships by taking into account different kinds of knowledge (contextual knowledge, cultural knowledge, personal background

knowledge and domain knowledge). The authors have considered that the two last kinds of query are situated at the semantic level.

In [Hare et al., 2006], the authors tried to characterize the semantic gap. In particular, they made the distinction between the gap between image descriptors and image labels and the gap between object labels and the full semantics. For the authors, full semantics implies to consider not only high-level concepts but also the relationships between these different concepts.

Finally, the key work of [Smeulders et al., 2000] remains the most complete currently existing work setting out the image semantics problem. The authors have succeeded to highlight the main challenges of image retrieval systems, even if many of the mentioned issues have been partially identified before. Specifically, among their statements, the authors mentioned that "on the coverage side, labeling is seldom complete, context sensitive, and, in any case, there is a significant fraction of requests whose semantics can not be captured by labeling alone". However, according to the current advances in image retrieval field, automatic image annotation remains the only alternative to solve the problem of semantic gap and image retrieval related tasks. Indeed, we believe that, in anyway, automatic image annotation is the preliminary step for any process of image description and interpretation. Nevertheless, as aforementioned by the previous work, the image semantics is a subjective property acquired by the end user during the image interpretation process. Therefore, we believe that the use of knowledge models about image semantics and the use of inference mechanisms are priceless in order to automatically perform semantic image annotation and interpretation.

2.4 An Overview of Classical Image Annotation Approaches

The rapid growth of multimedia content comes with the need to effectively manage this content by providing mechanisms for image indexing and retrieval that can meet user expectations. Towards this goal, semantic image analysis and interpretation has been one of the most interesting challenges during this last decade, and several attempts have addressed the, previously introduced, *semantic gap* problem. In particular, a typical method for narrowing the semantic gap is to perform automatic image annotation. Automatic image annotation was introduced in the early 2000s, and first efforts focused on statistical learning approaches as they provide powerful and effective tools to establish associations between the visual features of images and the semantic concepts [Barnard et al., 2003, Lavrenko et al., 2003, Monay and Gatica-Perez, 2003, Carneiro et al., 2007]. A recent review on automatic image annotation techniques is proposed in [Zhang et al., 2012].

Early efforts aiming to narrow the semantic gap have focused on providing mechanisms/methods for mapping low-level features (such as color, texture, shape and salient points) directly to some specific semantic concepts such as indoor/outdoor [Szummer and Picard, 1998], nature, animal, food [Smith and Chang, 1997], pedestrian [Papageorgiou and Poggio, 1999] and so on. Therefore, many dedicated detec-

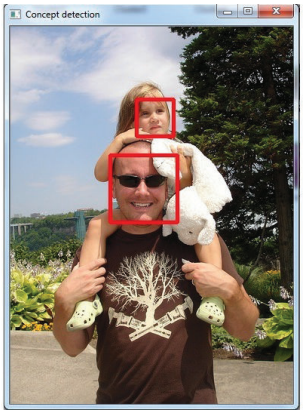
tors based on simple decision rules have been proposed. We refer to these approaches as *specific-concepts* dedicated approaches, since they are dedicated to a specific set of concepts with a specific low-level representation of images. These approaches have quickly become cumbersome and impractical following the normal request of a larger annotation vocabulary. Indeed, it would be impossible to build a detector for each potential concept, as they are too many. To overcome the limitations of the methods dedicated to specific concepts, other approaches have proposed to build *generic* detectors/classifiers. More precisely, given a set of positive and negative training samples, generic classifiers are trained separately using a single approach. We refer to these approaches as image classification approaches or concept-detectors based approaches if they also aim at localizing the concept in the image. However, since these approaches do not consider concept-specific knowledge/properties, they may be less accurate than specific-concepts based methods, but allow in the other hand to deal with a large-scale annotation vocabulary [Snoek et al., 2006]. In the remaining of this section, we review some of the existing approaches for image annotation, emphasizing each time the difference between each category of approaches and their limitations.

2.4.1 Concept-Detectors Based Approaches

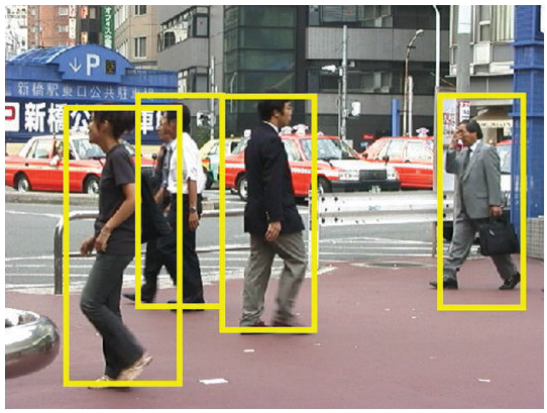
As aforementioned, the concept-detectors based approaches (also called object detection approaches) aim at building a (set of) specific concept detector(s), i.e. a dedicated detector to each specific concept/object of interest. The main difference between this category of approaches and image classification based approaches, is that the former one requests to slide a window across the image (possibly at multiple scales), and to classify each such local window as containing the target concept or a background [Papageorgiou and Poggio, 2000, Viola and Jones, 2001, Mohan et al., 2001, Dalal and Triggs, 2005, Torralba et al., 2007, Wei and Tao, 2010]. Consequently, concept-detectors based approaches allow *identifying* and *localizing* the recognized concepts in processed images, i.e. predicting in a given image the bounding box and the label of each detected concept from the annotation vocabulary - cf. Figure 2.6. While the image-classification-based approaches allow only predicting in a given image the presence/absence of a set of concepts from the annotation vocabulary ($c_j \in \mathcal{C}$).

More precisely, the concept-detectors based approaches, also called sliding-window object detection methods, consist in defining a fixed-size rectangular window and applying a classifier, such as boosting [Viola and Jones, 2001] or support vector machine [Papageorgiou and Poggio, 2000], 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.

This technique has shown impressive results for some specific concepts such as face detection [Osuna et al., 1997, Viola and Jones, 2001] or pedestrian detection [Dalal and Triggs, 2005], but these approaches have failed to generalize to other concepts. Doubtless, the most famous object/concept detector is the one proposed by [Viola and Jones, 2001] for face detection, as it provides very competitive object detection rates in real-time. They proposed to use a cascade of weak classifiers in order



(a) Face detection using Viola and Jones object detection framework [Viola and Jones, 2001]



(b) Pedestrian detection using Histograms of Oriented Gradients (HOG) [Dalal and Triggs, 2005]

Figure 2.6: Concept-detectors based approaches, also called sliding-window object detection methods.

to create a strong classifier with a good detection rate. As illustrated in Figure 2.7, a sub-window (also called a patch) of a given image is classified as containing a face only if it passes tests in all nodes in the cascade. Usually, non face patches are quickly rejected by the early nodes. [Viola and Jones, 2001] proposed to use: i) the integral image representation to compute rapidly the image features (called rectangle features), and ii) the AdaBoost algorithm [Schapire et al., 1998] to select rectangle features and combine them into a cascade of classifiers.

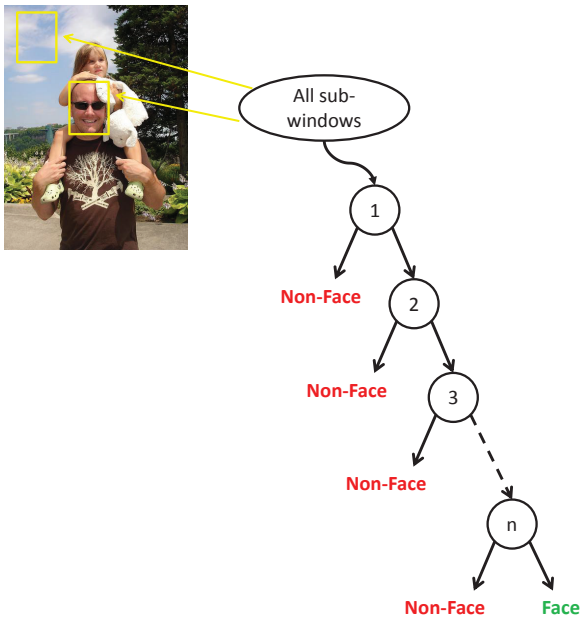


Figure 2.7: Scheme of the Viola and Jones face detector [Viola and Jones, 2001].

[Dalal and Triggs, 2005] proposed a method for pedestrian detection which achieves high accuracy. They proposed to use a single filter on Histogram of Oriented Gradients (HOG) features to represent an object category. This detector uses a sliding window approach, where a filter is applied at all positions and scales of an image. The classifier allows to detect whether or not there is an instance of the target concept at a given position and scale. Their method has shown very good performance on the INRIA human dataset. [Mikolajczyk et al., 2004] proposed a method for human detection based on a probabilistic assembly of robust part detectors. These parts were represented by co-occurrences of local features which capture the spatial layout of the part appearance. Separate detectors were then trained for each part using AdaBoost algorithm. The target location was determined by maximizing the joint likelihood of body parts occurrences combined according to the geometric relations.

Other approaches have tackled the problem of concept detection in a generic way, i.e. without targeting specific concepts. [Torralba et al., 2004, Torralba et al., 2007] proposed an approach allowing to share features among different object categories, thus providing a faster solution for multi-classes object detection. [Lampert et al., 2008] proposed to consider the object detection problem as finding the optimal bounding box that gives the highest detection score in the image. They proposed an efficient variant of the branch-and-bound method for retrieving the optimal bounding box in the image. In [Felzenszwalb et al., 2010], an object detection system based on mixtures of multiscale deformable part models is proposed. This system is able to represent highly variable object classes and achieves state-of-the-art results in the PASCAL object detection challenges.

Object detection is one of the fundamental tasks of visual recognition and computer vision, and consequently the literature is very rich concerning this topic. The aim of this section is not to provide an exhaustive review of this kind of approaches, i.e. concept-detectors based approaches. Therefore, for more information the reader is suggested to refer to [Galleguillos and Belongie, 2010].

Discussion

Concept-detectors based approaches have shown good accuracy for detecting specific objects/concepts, such as faces, pedestrians, cars, etc. These approaches select the parameters of the model so as to minimize the detection error on a set of training images. Such approaches aim to directly optimize the decision boundary between positive and negative samples [Felzenszwalb et al., 2010]. However, these approaches seem unlikely to scale up to the detection of a large number of concepts, or to many different views of objects, since each classifier computes many image features independently [Torralba et al., 2007]. These features typically involve convolutions with part templates [Fergus et al., 2003] or with a set of basis filters [Viola and Jones, 2001, Dalal and Triggs, 2005]. Therefore, computing these features is slow and requires a big set of training samples to find the useful features.

Moreover, sliding window based approaches are computationally-expensive. For instance, given an image of size $n * n$ and a window of size $r * r$, a histogram-based sliding window approach needs to scan n^2 windows, to scan r^2 pixels per window to

construct the histogram, and to scan B bins of the histogram to evaluate the objective function [Wei and Tao, 2010]. Thus, the overall complexity of this kind of approach is $\mathcal{O}(n^2(r^2 + B))$, which is unaffordable when either n, r or B is large. Many attempts have been proposed to reduce the complexity of sliding window based approaches, but despite this, the complexity is still very high with respect to current needs for detecting concepts in images and videos.

2.4.2 Image Classification Based Approaches

Image classification based approaches allow predicting the presence/absence of a set of concepts (from the annotation vocabulary $c_j \in \mathcal{C}$) in a given image. These approaches usually follow the scheme illustrated in Figure 2.8. These, either extract:

- global features (computed on the whole image or by the use of dense sampling) [Tong and Chang, 2001, Monay and Gatica-Perez, 2004, Carneiro et al., 2007, Lazebnik et al., 2006], or
- require a prior segmentation of the image as regions/blobs[2] [Barnard and Forsyth, 2001, Carson et al., 2002, Duygulu et al., 2002, Blei and Jordan, 2003, Romdhane et al., 2010], or as blocks [Feng et al., 2004, Carbonetto et al., 2004].

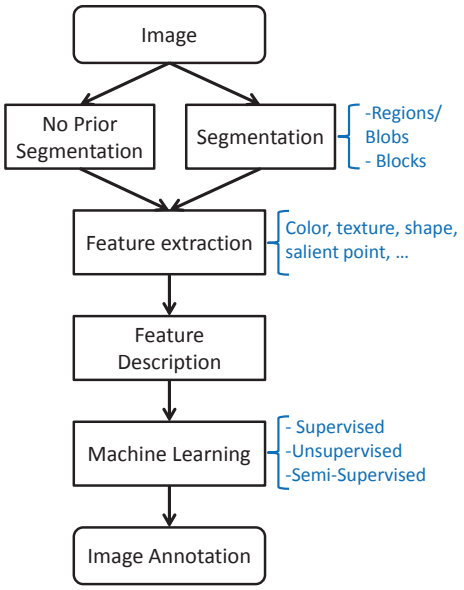


Figure 2.8: Basic scheme for image annotation as a classification problem.

Thereafter, these approaches carry out the extraction of features. A feature is defined to capture a certain visual property of an image, either on the whole image or on a region of it. Most commonly used features for image classification are the ones

[2]Blobs are regions of an image that are somehow coherent.

reflecting: color, texture, shape, and salient points in images. Feature description is subsequently performed to assign a signature to the extracted features. Finally, a machine learning algorithm is trained to recognize/detect the concepts from the annotation vocabulary using the visual features of images.

Early efforts in image classification based approaches have focused on the extraction of reliable specific semantics, for instance, indoor vs. outdoor [Szummer and Picard, 1998], city vs. landscape [Vailaya et al., 1998] and human vs. horses [Forsyth and Fleck, 1997], buildings vs. non-building [Li and Shapiro, 2002], etc. These approaches have addressed the problem of extracting semantics using **supervised learning techniques**. Consequently, Bayesian classifiers [Vailaya et al., 2001], Support Vector Machines [Griffin and Perona, 2008], multiple instance learning [Zhang et al., 2002], statistical models [Hastie et al., 2001], k-NN [Tang et al., 2011], and artificial neural networks [Romdhane et al., 2010] are often used to learn high-level concepts (i.e. semantic content) from low-level features (i.e. visual content). This is achieved by collecting a set of positive and negative training images for the concept of interest, and therefore a binary classifier is trained to detect this concept. The process is repeated for each concept of the annotation vocabulary. These classifiers are then applied to all database images which are, accordingly, annotated with respect to the presence or absence of the concepts from the annotation vocabulary.

Other approaches have attempted to address the problem more generally by using **unsupervised learning** [Barnard and Forsyth, 2001, Duygulu et al., 2002, Blei and Jordan, 2003, Lavrenko et al., 2003]. Usually, these approaches employ probabilistic models to explain the co-occurrence between image features and semantic labels. The basic idea of these approaches is to introduce a set of latent variables that encode hidden states of data, and where each state induces a joint distribution on the space of semantic labels and image visual descriptors (local features computed over image neighborhood) [Torralba et al., 2007]. During the training process, each image is assigned a set of labels from the annotation vocabulary, these images are thereafter segmented into a set of regions (using a block-based decomposition [Feng et al., 2004, Carbonetto et al., 2004] or using image segmentation methods [Barnard and Forsyth, 2001, Duygulu et al., 2002, Blei and Jordan, 2003, Lavrenko et al., 2003]), and an unsupervised learning algorithm is performed on the whole database to estimate the joint density of semantic labels and visual features. Therefore, the annotation of previously unseen images is performed as follows. Firstly, visual feature vectors are extracted from these images. Then, these feature vectors are used to instantiate the joint probability model. Thereafter, the state variables are marginalized, and finally the set of labels that maximize the joint density of concepts and visual appearance are assigned to the input image [Carneiro et al., 2007].

In the following, we propose a brief overview of the different categories of image classification according to the used machine learning paradigm (i.e. supervised, unsupervised or semi-supervised). We refer to the notations of Section 2.2.

2.4.2.1 Supervised Learning

The formulation of automatic image annotation as a supervised learning problem has been proposed in [Carneiro and Vasconcelos, 2005, Carneiro et al., 2007]. It consists firstly in performing a training phase, where $\mathcal{N} =\mid \mathcal{C} \mid$ distinct classifiers are trained, each to detect the presence/absence of a given concept $c_i \in \mathcal{C}$. Let x_k^v be any visual representation of an image i_k (a visual features vector), Y_i a random variable such that $Y_i = 1$ if i_k is annotated with the concept c_i, otherwise $Y_i = 0$, X a random vector of visual features. The training of a classifier $\langle classifier_\lambda \rangle$ is performed as follows:

i) all images in the training dataset annotated with c_λ are collected as positive samples, remaining images from the training dataset are considered as negative samples, and

ii) some density estimation algorithm is trained to estimate $P_{X|Y_i}(x_k^v|j)$, i.e. the conditional density that the feature vector x_k^v of the input image i_k could be associated/or-not to the the semantic class c_i (where $j \in \{0, 1\}$).

According to the statistical decision theory [Duda et al., 2001], the decision function to annotate an image i_k with a concept c_i could be easily computed as:

$$P_{X|Y_i}(x_k^v|1)P_{Y_i}(1) \geq P_{X|Y_i}(x_k^v|0)P_{Y_i}(0) \tag{2.1}$$

Equation 2.1 produces, for a given input image i_k, an output $\mathcal{P}$ which consists of a set of candidate concepts $c_i \in \mathcal{C}$ and their confidence values $\alpha_i = P_{Y_i|X}(1|x_k^v)$, i.e. $\mathcal{P} : \langle (c_0, \alpha_0), (c_1, \alpha_1), \cdots (c_m, \alpha_m) \rangle$, $0 \leq m \leq \mathcal{N}$.

For instance, a simple way to perform image annotation using supervised learning techniques is to train a set of binary classifiers, one for each concept from the annotation vocabulary [Tong and Chang, 2001, Goh et al., 2001]. Other approaches have also applied this method in order to annotate images in restricted domains, to distinguish cities from landscapes [Vailaya et al., 1998], to detect indoor scenes from the outdoor ones [Szummer and Picard, 1998], and so on. [Smith et al., 2003] presented a two-step Discriminative Model Fusion (DMF) approach to discover the implicit/indirect relationships between concepts by constructing model vectors based on detection scores of individual classifiers. Support vector machines are then trained to refine the detection results of the individual classifiers.

[Li and Wang, 2003] used a statistical modeling approach to convert images into keywords. They considered the images of any given concept as instances of a stochastic process that characterizes this concept. Thus, they computed the extent of association between an image and the textual description of a concept as the likelihood of the occurrence of the image based on the characterizing stochastic process. [Boutell et al., 2004] applied multi-label learning techniques for scene classification. They decomposed the multiclass learning problem into several independent binary classification tasks. They also provided many labeling criteria to predict a set of labels, for a given test image, based on the outputs of the different binary classifiers.

Other approaches have proposed to use Multiple-Instance Learning in order to perform image annotation [Chen and Wang, 2004, Carneiro et al., 2007]. In Multiple-Instance Learning (MIL) approaches, each image is treated as a bag of instances, and the goal is to maximize the diverse density or the soft margins between the negative and the positive samples of a given concept. For example, [Carneiro et al., 2007] proposed a supervised multiclass labeling (SML) method, in which images are represented as bags of localized feature vectors, and a mixture hierarchies is built to learn the correspondence between images and their labels. The mixture hierarchy was composed of one mixture density estimated for each image and the mixture density estimated on all images annotated with a same given concept. [Chen and Wang, 2004] proposed a MIL framework named Diverse-Density-SVM (DD-SVM). DD-SVM learns first a collection of instance prototypes according to a Diverse Density (DD) function. The training is performed in a feature space constructed from a mapping defined by the local maximizers and minimizers of the DD function. Through the feature mapping, DD-SVM essentially converts MIL to a standard supervised learning problem.

2.4.2.2 Unsupervised Learning

Unsupervised learning approaches for image annotation [Barnard and Forsyth, 2001, Duygulu et al., 2002, Jeon et al., 2003, Blei and Jordan, 2003, Lavrenko et al., 2003] aimed at finding hidden structure in unlabeled images. These structures are represented by a set of states which define each a joint distribution for semantic labels and image features [Carneiro et al., 2007]. Three categories of approaches exist in the literature, which differ according to the definition of the states of the hidden variable:

- associating a state to each image in the database [Lavrenko et al., 2003],
- associating a state to each image cluster [Barnard and Forsyth, 2001, Duygulu et al., 2002],
- or associating a state to a topic[3] [Blei and Jordan, 2003, Monay and Gatica-Perez, 2007].

The form of the overall model is as follows :

$$P_{X|W}(x_k^v|c_i) = \sum_{l=1}^{S} P_{X|L}(x_k^v|l) P_{W|L}(c_i|l) \tag{2.2}$$

where x_k^v is a visual representation of the image i_k (a visual features vector), $c_i \in \mathcal{C}$ is a concept from the annotation vocabulary, X a random vector of visual features, W is a random variable that takes value in $\{1, \ldots, \mathcal{N}\}$ such that $W = i$ if and only if X is a sample of c_i, S is the number of possible states of L.

Equation 2.2 illustrates a mixture model, and therefore learning is usually based on the Expectation-Maximization (EM) algorithm [Dempster et al., 1977].

[3] A topic is a latent variable which represents a distribution over words. These approaches are known as 'topic models'.

For more information, the readers are suggested to refer to [Duda et al., 2001, Carneiro et al., 2007].

[Mori et al., 1999] proposed a co-occurrence model based on the co-occurrence of words and image regions created using a regular grid. [Duygulu et al., 2002] proposed a Machine Translation model (MT) to translate image blobs into concepts. First, the image is segmented into regions using a segmentation algorithm. Features are then computed on each image region, and blobs are generated by clustering the image features for these regions across all images. Therefore, each image is generated by using a collection of these blobs. Finally, the proposed translation model applies one of the classical statistical machine translation models to translate from the set of keywords of an image to the set of blobs forming the image. [Jeon et al., 2003] proposed a Cross-Media Relevance Model (CMRM), which considers that concepts and blobs are conditionally independent in a given an image. [Blei and Jordan, 2003] extended the Latent Dirichlet Allocation (LDA) model to propose a Correlation LDA model which relates words and images. In their approach, they used a Dirichlet distribution to generate a mixture of latent factors, which is then used to generate words and regions. Expectation-Maximization is again used to estimate this model. [Lavrenko et al., 2003] proposed a model called Continuous-space Relevance Model (CRM), which is closely related to the models proposed by [Blei and Jordan, 2003, Jeon et al., 2003], but it brings some improvements. CRM makes no assumptions about the topological structure, it directly models continuous features, i.e. does not rely on clustering, and consequently does not suffer from the granularity issues.

Some other approaches proposed to use Latent Dirichlet Allocation (LDA) and *graphical models* to learn a joint distribution model for a set of keywords and image regions [Barnard and Forsyth, 2001, Barnard et al., 2003, Feng and Lapata, 2010]. These approaches assume that a variable (representing the hidden state) exists in the data generative process, and which links the concepts and the visual feature vectors through conditional relationships. Indeed, several variations of Latent Dirichlet Allocation (LDA) based mixture models are based on this assumption. Therefore, in these approaches, images are represented as a collection of region-based image features, and are modeled by Gaussian distributions, while concepts are modeled with multinomial distributions. In [Barnard and Forsyth, 2001, Barnard et al., 2003], several methods to model the joint distribution of words and blobs are discussed. Subsequent to the learning of joint distributions, image annotation in these approaches is regarded as a problem of likelihood estimation between the blobs and the words. However, the accuracy of these models is sensitive to the quality of image segmentation. In the same context, [Monay and Gatica-Perez, 2003, Monay and Gatica-Perez, 2004] proposed to use Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) for image annotation.

2.4.2.3 Semi-Supervised Learning

Some recent work for image annotation have proposed to use semi-supervised learning to draw benefits from both, labeled and unlabeled (or weakly labeled) images. Indeed, due to the lack of available labeled training data, semi-supervised learning is a good

alternative for automatically annotating large-scale image databases with semantic concepts. In semi-supervised learning, a number of labeled examples are usually required for training an initial *weakly useful predictor* which is in turn used for exploiting the unlabeled examples [Guillaumin et al., 2010]. Therefore, these approaches try to tackle the image annotation problem by incorporating the information gathered from labeled images in a large amount of unlabeled data. [Zhu, 2006] proposed a good introduction to Semi-Supervised Learning (SSL).

Current semi-supervised learning based approaches can be categorized into three main paradigms.

- In the first, a generative model, such as Naive Bayes classifiers or a mixture of Gaussians, is used to model the joint probability distribution over observations and label sequences. The EM algorithm is then performed in order to model the label estimation or the parameter estimation process [Guillaumin et al., 2010].

- The second paradigm deals with the regularization of the learning process using the unlabeled data [Kück et al., 2004, Fergus et al., 2009, Wang et al., 2006]. For instance, some approaches use a graph where the nodes are set as the image visual descriptors and the edges are used to encode the similarity between the image samples. Therefore, the label smoothness can be enforced over the graph as a regularization term. For example, [Fergus et al., 2009] specified the label prediction function using smooth eigenvectors of the graph Laplacian which are calculated by a numerical method.

- The third paradigm, called co-training or also multiview learning, works under a two-view[4] setting. It assumes that features can be split into two sets and that the two sets are conditionally independent given the class. This paradigm requires training two separate classifiers with the labeled data, on the two sub-feature sets respectively. Then, the unlabeled data is classified by each classifier, which must agree on the much larger unlabeled data as well as the labeled one. Each classifier is also able to 'teach' the other one with the unlabeled examples using the most confident prediction. For example, [Zhou et al., 2007] proposed to take advantages of the correlations between the views using canonical component analysis. Their proposed method can perform semi-supervised learning with only one labeled training example.

Discussion

In order to compare the previously defined categories of image classification approaches, we propose in Table 2.1 a list of advantages and limitations with respect to each category. A more detailed comparison of these approaches is provided in Section 2.6. As one can deduce from this table, none of these categories is optimal in general terms, but each can be suitable with respect to a given image annotation problem. Given a specific image annotation task, the criteria for choosing between these categories can be summarized as follows:

[4] two views of an item, for example: image and HTML text

- Availability of a sufficient amount of images for training the classifiers.
- The size of the annotation vocabulary (scalability).
- The need for a natural ranking of keywords.
- The required level of precision for the image annotation task.

Image Classification Based Approaches		
Approaches	Pros	Cons
Supervised Learning	✓Are more intuitive to implement. ✓Allow designing features and tuning the learning algorithm for each classification task.	✗ Do not scale well with a large vocabulary (require training separated classifier). ✗ Do not allow natural ranking of keywords since the probabilities obtained from the different classifiers are not comparable.
Unsupervised Learning	✓Are more scalable in terms of training (when image database is large and/or concept number is high). ✓Produce a natural ranking of keywords.	✗ Do not guarantee that the semantic annotations are optimal under a recognition or retrieval sense (since keywords are not explicitly treated as classes).
Semi-Supervised Learning	✓Require few labeled data to train efficient classifiers. ✓Draw benefits from weakly labeled images.	✗ Scale polynomially with the number of images. ✗ Are impractical for use on gigantic collections with hundreds of millions of images and thousands of classes [Rohrbach et al., 2010].

Table 2.1: Comparison of image classification approaches.

2.4.3 Hierarchical Image Classification

Many recent approaches have proposed to use hierarchical structures in order to address the scalability problem of automatic image annotation. In this dissertation we have also addressed this problem, i.e. the use of semantic hierarchies for image annotation and hierarchical image classification. Therefore, in order to put our contributions into perspective and to compare to existing approaches, we provide a detailed description of the hierarchical image classification approaches in Section 5.2. Moreover, we believe that these hierarchies are bearers of knowledge, either, visual or conceptual, and consequently we provide a discussion on the used hierarchy structures

for image annotation in Section 3.2. In the following, we only introduce the formal definition of hierarchical image classification.

From a formal point of view, the hierarchical image classification problem can be defined as follows.

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.
  - Each textual annotation $\mathscr{A}_i$ consists of a set of concepts $\{c_j \in \mathcal{C}, j = 1..n_{i_i}\}$ associated with a given image $i_i \in \mathcal{DB}$.
- $\mathcal{C}' = \langle c'_1, c'_2, \cdots, c'_{\mathcal{N}'} \rangle$ a set of visual classes[5] (or concepts, depending on the nature of the hierarchy) that link all the concepts of $\mathcal{C}$ in a hierarchical structure.
- We define:
  - $\mathscr{C}$ as a set of classes (or concepts) such that $\mathcal{C} \subseteq \mathscr{C}$ and $\mathcal{C}' \subset \mathscr{C}$,
  - X a random vector of visual features (i.e. $X = \{x_k^v \mid \forall k \in [1, \mathcal{L}]$, x_k^v is a visual representation of the image $i_k\}$),
  - $(\mathscr{C}, \sqsubseteq)$ a class hierarchy, where $\sqsubseteq$ is a partial order representing the subsumption relationship, i.e. $(\forall c_i, c_j \in \mathscr{C} : c_i \sqsubseteq c_j \iff c_i$ is subsumed by $c_j)$,
  - T a set of examples (x_k^v, S_k), with $x_k^v \in X$ and $S_k \subseteq \mathscr{C}$ such that $c_i \in S_k \Rightarrow \forall c_i \sqsubseteq c_j : c_j \in S_k$.

Objective: Find an objective function $f : X \to \mathcal{C}$ such that:

$$f(x_k^v) = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}} \left(\sum_{c_j \in S_k} P_{X|Y_i}(x_k^v | c_j) \right) \tag{2.3}$$

Therefore, the aim of hierarchical image classification is to find an objective function (or a decision function) $f(x_k^v)$ that maximizes the probability of association between visual representation of images and the concepts from the annotation vocabulary ($c_j \in \mathcal{C}$), while taking into account the local probabilities throughout a given path in the hierarchy. These local probabilities correspond to the belonging likelihood

[5] In this definition, we refer to a visual class as an abstract node in the hierarchy without any associated label, whereas concept is a node in the hierarchy with a given associated label.

of the visual representation of a given image to the intermediate classes (or concepts) of the hierarchy. Thus, hierarchical image classification introduces a dependency between the concept nodes of a given path in the hierarchy in order to achieve a relevant decision about the belonging of images to the leaf concepts. For more information, see Section 5.5.

2.4.4 Discussion

Current approaches for automatic image annotation have participated to reduce partially the semantic gap problem by providing a set of methods allowing to link the visual content of images to semantic concepts. Within this context, many approaches based on machine learning techniques have been proposed in order to model the correlation between image features and the concepts. Indeed, automatic image annotation has been considered in the last decade as a multi-class classification problem. However, these approaches, even if they adequately describe the visual content of images, are often limited to detect perceptual manifestations of semantics, and then are unable to model the image semantics as it is perceived by humans (cf. Section 2.3). They also have many limitations when dealing with broad content image databases [Liu et al., 2007], i.e. the obtained performances vary significantly according to the considered concept number and the targeted image datasets as well [Hauptmann et al., 2007, Deng et al., 2010]. 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 conflicting annotations [Fan et al., 2008b]. Thus, it is clear that there is a lack of coincidence between the high-level semantic concepts and the low-level features of images, and that the image semantics is not always correlated with its visual appearance.

Consequently, and always in the quest for models that could help to map successfully low-level features into high-level semantic concepts, some approaches proposed to use **knowledge-driven frameworks** for image annotation. These approaches rely on the use of (explicit or implicit) *contextual knowledge* in order to improve the image annotation accuracy. Two categories of approaches can be distinguished:

1. using information fusion stemming from multiple sources, and
2. using explicit semantic structures, such as semantic hierarchies and ontologies, for modeling knowledge about image context.

This knowledge is thereafter injected into the process of image analysis/annotation in order to achieve a relevant decision about the image content. We refer to these approaches as *Semantic Image Annotation* approaches. In the next section we introduce these approaches and we review some of the relevant work.

2.5 Semantic Image Annotation: Towards Knowledge -Driven Approaches for Image Annotation

Objects in the real world are always seen embedded in a specific context, and the representation of this context is essential for the analysis and the understanding of images. Contextual knowledge may stem from multiple sources of information, including knowledge about the expected identity, size, position and relative depth of an object within a scene [Gronau et al., 2008]. For instance, topological knowledge can provide information about which objects are most likely to appear within a specific visual setting, e.g. an office typically contains a desk, a phone, and a computer, but it is unlikely that it contains a bed. Spatial information can also provide information about which locations within a visual setting are most likely to contain objects, e.g. in a beach scene, the sky is usually placed at the top, while the sea is below. Given a specific context, this kind of knowledge can help reasoning on data to improve image annotation [Hudelot et al., 2008, Neumann and Möller, 2008].

Definition 7 (Contextual information/knowledge)**.** *By contextual information, we mean the collection of relevant conditions and surrounding influences that make a situation unique and comprehensible. While contextual knowledge is the information, and/or skills that have particular meaning because of the conditions that form part of their description.*

Consequently, it is of prime interest to make efficient use of contextual knowledge in order to narrow the semantic gap, and to improve the accuracy of image annotation. In the following, we introduce semantic image annotation approaches which make use of contextual knowledge, either implicitly or explicitly. Firstly, we introduce the set of approaches for image annotation using information fusion as a tentative to (implicitly) capture the image semantics. Thereafter, we introduce the second category of approaches which use explicit semantic structures, such as semantic hierarchies and ontologies, in order to narrow the semantic gap.

2.5.1 Information Fusion for Image Annotation

The first notable attempt for using contextual information (perceptual context) for image annotation was proposed by [Lavrenko et al., 2003]. The authors proposed a statistical generative model which looks at the probability of associating words with image regions. They used the surrounding visual context by computing a joint probability of image features over different regions in an image using a training set. This joint probability is thereafter used to annotate and retrieve images. Thus in this model, the association of different regions provides context while the association of words with image regions provides meaning.

Recent work in computer vision have also stressed the importance of contextual information for improving object recognition in real world images [Oliva and Torralba, 2007, Galleguillos and Belongie, 2010]. These approaches are based on the Biederman's semantic relations [Biederman, 1972] to achieve robust

object categorization in real world scenes. These semantic relations, called *contextual features*, can be classified into three:

1. Semantic context, which can be defined as the likelihood of finding an object in some scenes but not others [Rabinovich et al., 2007, Galleguillos et al., 2008, Divvala et al., 2009]

2. Spatial context, which can be defined as the likelihood of finding an object in some spatial locations with respect to other objects in the scene [Russell et al., 2007, Heitz and Koller, 2008, Galleguillos et al., 2010]

3. Scale context, which is a contextual relation based on the scales of an object with respect to others [Torralba, 2003, Torralba et al., 2010, Gould et al., 2009]

In all these approaches, the used contextual information is about image features, i.e. contextual features, and is incorporated at the image processing level. Therefore, these approaches do not fall within the scope of our research, and we will not discuss further on them. A comprehensive survey about this category of approaches can be found in [Galleguillos and Belongie, 2010].

Subsequently, several approaches have proposed to combine information collected from the different image modalities in order to improve the image annotation accuracy. These approaches use a set of techniques, called *semantic combination*, in order to efficiently fuse these information (or modalities). Merged information usually stems from a couple (or more) of multimedia modalities, i.e. fusion of visual and textual information [Tollari, 2006, Clinchant et al., 2011], fusion of textual information and user tags [Guillaumin et al., 2010, Rohrbach et al., 2010, Znaidia et al., 2012], fusion of multiple images [Tommasi et al., 2008], etc. These approaches, called cross-media fusion approaches, are based on the following paradigms:

- Exploitation of the diversity between the different information sources.
- Exploitation of the dependency between the multimedia modalities.
- Combination of multiple complementary information sources and generalizing over them.
- Cooperative combination of features in order to get a more precise representation of the world.

As introduced by [Clinchant et al., 2011], three types of information fusion techniques have been proposed:

1. *Early fusion*: which consists in representing the multimedia objects in a multimodal feature space (e.g. concatenation of features stemming from images and textual information) [Lavrenko et al., 2003, Tollari, 2006, Rasiwasia et al., 2010].

2. *Late fusion*: consists in merging the monomedia similarity profiles by means of aggregation functions (e.g mean average of both modalities) [Escalante et al., 2008, Choi et al., 2010, Kulkarni et al., 2011].

3. *Transmedia fusion*: which consists in a diffusion processes that act as a transmedia pseudo-relevance mechanism [Jeon et al., 2003, Bruno et al., 2008, Ah-Pine et al., 2009].

Although, [Maillot et al., 2007, Müller et al., 2010] have reported that the late fusion and the transmedia fusion techniques are performing better than the early fusion one, there is no fusion strategy which is optimal in a general way. The fusion strategy has to be selected according to the targeted task and the data structure [Clinchant et al., 2011].

Discussion

The image annotation approaches based on information fusion are motivated by the use of the correlations between the different multimedia modalities in order to reduce the semantic gap problem. Indeed, the fusion of information stemming from multiple sources can reduce considerably the uncertainty of image annotation, and as a consequence, significant improvements on the accuracy of image annotation have been reported by these approaches.

However, these approaches do not necessarily provide a reliable solution to solve the semantic gap problem in its entirety. Actually, these approaches do not allow establishing explicit links between image features and image semantics. Therefore, they only use implicit inference mechanisms for improving the image annotation results and remain closely sensitive to machine learning accuracy. The use of *explicit* semantic structures, such as semantic hierarchies and ontologies, seems then to be essential to tackle the semantic gap problem in an effective way.

2.5.2 Ontology-Driven Approaches for Image Annotation

An important issue to solve the semantic gap problem is to make use of explicit and formal methods to represent contextual knowledge. This will help taking into account general and specific context of the image, and allow reasoning in order to improve the image annotation. [Kompatsiaris and Hobson, 2008] underlines that among the possible representations of knowledge in the multimedia domain, ontologies are the most useful, and have considerable advantages. Indeed, ontologies provide a formal framework that may contain explicit semantic definitions, which can be directly processed by a machine, and allow at the same time to derive implicit knowledge by automatic inference.

2.5.2.1 Ontology and Multimedia Ontologies

The concept of Ontology emerged in the early 1990's in the artificial intelligence and the knowledge engineering communities. The term Ontology was borrowed from philosophy where it means a theory about the nature of being, the study of the kinds of things that exist. Nowadays, ontologies have become the new standard for knowledge representation [Horrocks et al., 2003, Baader, 2011].

Ontologies are a formal, explicit specification of a shared conceptualization [Gruber, 1995]. "Formal" reflects that ontology is machine-readable and allows reasoning about its content from the human and the machine side. "Explicit" means that the type of concepts used, and the constraints on their use are explicitly defined. "Shared" refers to the common knowledge embodied in ontology. "Conceptualization" refers to the model obtained by abstracting some phenomena existing in the real world by identifying the relevant concepts of those phenomena. Thus, ontologies allow capturing the relevant knowledge of a domain, provide a common understanding of this domain knowledge, determine acknowledged vocabulary of this domain, and give the explicit definition of the vocabulary(terms) and the relations between these vocabularies in formal models at different levels [Ren and Cheng, 2008].

Recent advances in ontological engineering have motivated several work in the field of image retrieval. As a result, a considerable number of multimedia ontologies have been proposed in order to define standards for the description of low-level multimedia content [Hunter, 2001, Simou et al., 2005, Arndt et al., 2007, Dasiopoulou et al., 2010]. Other domain ontologies have been proposed to allow the semantic interpretation of images and reasoning over the extracted descriptions [Bagdanov et al., 2007, Neumann and Möller, 2008, Peraldi et al., 2007, Hudelot et al., 2008, Hudelot et al., 2010].

Usually, ontologies were used in the image retrieval field to target the following goals:

1. A unified description of low level features: where ontologies are used to provide standards of description of low-level features - e.g. [Simou et al., 2005, Dasiopoulou et al., 2010].

2. Visual description ontology: where ontologies are used to represent the different types of relations among image features such as edges, lines and region - e.g. [Maillot and Thonnat, 2008, Yao et al., 2010]. Typically the use of these ontologies comes during the image analysis process, and targets to optimize or to reason on this task.

3. Semantic mapping: ontologies are used to help the mapping between the visual level and the semantic level of images. For instance, the use of semantic hierarchies to reduce the semantic gap, e.g. [Fan et al., 2008a, Fan et al., 2008b].

4. Knowledge description: ontologies are used to model the concepts (objects) and relations among them. Typically, these are all approaches that use reasoning on concepts or on contextual information, i.e. after the image analysis process. These approaches most often tackle the problem of image annotation and interpretation - e.g. [Hollink et al., 2004, Simou et al., 2008, Hudelot et al., 2008, Dasiopoulou et al., 2009, Hudelot et al., 2010, Hamadi et al., 2012].

In a general way, the use of ontologies contributes to improve the image retrieval by incorporating contextual knowledge. The integrated knowledge can help in various spots of the image retrieval process, such as: i) image analysis, ii) mapping visual

features into semantic concepts, iii) assigning a meaning to tags, iv) disambiguation, and v) annotation enrichment. However, since the major challenge with the semantic gap is to provide effective tools that allow the mapping between low-level features and semantic concepts, semantic mapping is almost getting all the lights and it has been one of the most active issues. Indeed, while common methods for image annotation are limited to provide a latent correlation between semantic and visual spaces, ontology-driven approaches can make explicit this relationship. Consequently, ontologies provide an effective way to map low-level features into semantic concepts by building rules that supply semantic association between image features and concepts. This is achieved while maintaining a semantic structure to this process which will allow reasoning in order to check the consistency of this mapping. This domain is still in its infancy, but early results seem promising.

Moreover, despite a widely collaborative annotation effort, and the large amount of available goundtruth, the provided annotation vocabulary can never reach the richness of human-know vocabularies. Besides, different persons may use different terms to describe/search a same image. To address this problem, some ontologies [Fellbaum, 1998, Liu and Singh, 2004, Naphade et al., 2006] were proposed/used to structure the annotation vocabulary and the concepts that can be used by users to express their needs in a multimedia information retrieval system. The use of these ontologies allows then to make more precise and consistent the description of multimedia content. Ontologies allow disambiguating the various interpretation of a multimedia content by providing a semantic structure for the annotation vocabulary.

With respect to the image annotation field, explicit knowledge structures (or ontologies under their general form) have been introduced under three different shapes[6] for modeling contextual knowledge, as: i) *Heavy-Weight Ontologies*, ii) *Light-Weight Ontologies*, and iii) *formal Ontologies*. Figure 2.9 illustrates these categories and their underlying methods. In the following, we review the use of these explicit semantic structures for image annotation, and we discuss their roles in narrowing the semantic gap.

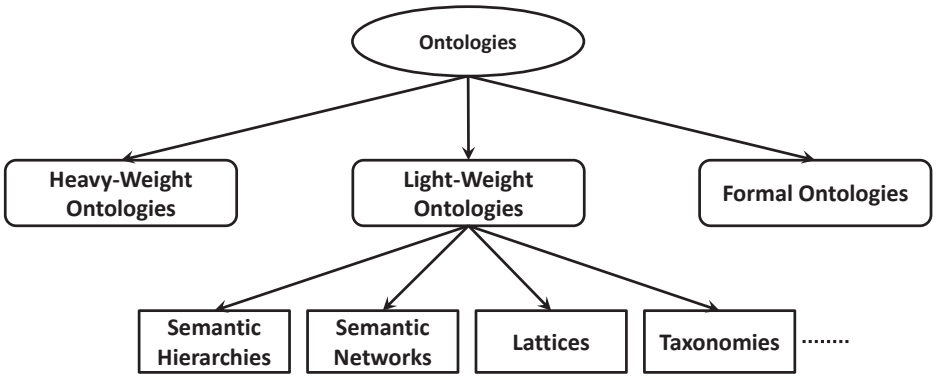


Figure 2.9: Explicit knowledge structures used for image annotation.

[6]This categorization is based on the level of expressiveness of the used knowledge model.

2.5.2.2 Heavy-Weight Ontologies (HWO)

Heavy-Weight Ontologies are fully described ontologies, including concept definitions and relations. These ontologies make intensive use of axioms to model knowledge about concepts and to restrict domain semantics. Heavy-Weight Ontologies, tailored to the image semantics understanding, have been used to attach meaning to the produced annotations and to help extracting, querying, analyzing and interpreting these annotations.

For example, in M-OntoMat-Annotizer [Petridis et al., 2006], low level MPEG-7 visual descriptions are linked to conventional Semantic Web ontologies and annotations. M-OntoMat-Annotizer is used in order to construct an ontology that includes prototypical instances of high-level domain concepts together with a formal specification of corresponding visual descriptors. Thus, it formalizes the interrelationships of high-level and low-level multimedia concept descriptions, allowing therefore for new kinds of multimedia content analysis and reasoning.

[Dong and Li, 2006] proposed a multi-ontology based multimedia annotation model. They proposed to integrate a domain-independent multimedia ontology with multiple domain ontologies in an effort to provide multiple domain specific views of multimedia content. Thus, accessing multimedia content can be less subjective to users background knowledge and their need of information. [Hare et al., 2006] suggested to use an ontology as an extra level in between the search query and keywords. So when performing a concept-based search, the search engine automatically performs inference to find all narrower concepts of the query concept. However, this method focuses on the query understanding and does not take into account image content. [Mezaris et al., 2003] proposed an image retrieval methodology where low-level features are extracted from image regions, and mapped automatically to intermediate level descriptors called 'object ontology', which is used for the definition of high-level concepts. Nevertheless, the suggested 'object ontology' is purely visual as it defines a simple vocabulary to describe perceptual manifestation of semantics (objects or regions).

Discussion

The aforementioned approaches are interesting, in particular, since they tried to narrow the semantic gap by using ontologies as a mid-level between image features and the semantic concepts in a tentative to make these relationships explicit. However, these approaches have not provided a formalism allowing reasoning over these explicit relationships, which make them impractical for improving the image annotation. Moreover, as stated previously in this chapter, there is no guarantee that there is a direct relationship between image features and the semantic concepts, but it rather seems that, often, this relationship is implicit and inferred from the context (of the image or of the observer).

It is therefore more appropriate to provide ontologies that model the image context, and to build tools allowing reasoning on this knowledge from both sides, the

visual appearance of concepts and their conceptual definition. This will certainly help narrowing the semantic gap in an effective manner.

2.5.2.3 Light-Weight Ontologies (LWO)

Ontologies have varying degrees of expressiveness. In this dissertation, we refer to ontologies with restricted expressiveness as Light-Weight Ontologies. These are partially described ontologies, like taxonomies, thesauri, semantic hierarchies, lattices and semantic networks. With respect to image retrieval field, many approaches have proposed to use Light-Weight Ontologies in order to perform image classification, or to provide multiple levels of abstraction image annotation [Marszalek and Schmid, 2007, Wu et al., 2008, Tousch et al., 2008, Fan et al., 2009, Martinet et al., 2011, Wu et al., 2012].

Definition 8 (Light-Weight Ontologies)**.** *A Light-Weight Ontology is a directed graph whose nodes represent concepts. The links between the nodes indicate associations (or untyped relationships) between the corresponding concepts. These associations express semantic nearness [Reimer et al., 2011].*

Light-Weight Ontologies were used under different shapes in order to improve the image annotation. For instance, *concept hierarchies* [Naphade et al., 2006, Marszalek and Schmid, 2007, Deng et al., 2009], *visual taxonomies* [Yao et al., 2010, Fei-Fei and Perona, 2005, Griffin and Perona, 2008, Bart et al., 2008], and *semantic hierarchies* [Fan et al., 2008a, Li et al., 2010] have been used for image annotation. These hierarchies[7] are typically ontologies limited to the subsumption (*is-a*) relationship, i.e. they are a collection of classes ordered by the transitive closure of explicitly declared subclass or subtype relations. Being $\mathcal{A}$ a subclass of $\mathcal{B}$, captures the fact that the state and the behavior of the elements of $\mathcal{A}$ are coherent with '*the intended meaning*' of $\mathcal{B}$ (or '*the visual features*' of $\mathcal{B}$, depending on the nature of the hierarchy), while disregarding the additional features and functionalities that characterize the subclass [Logozzo and Cortesi, 2006]. One of our contributions in this dissertation deals with the building of semantic hierarchies dedicated to image annotation. Therefore, we discuss in more details the use of these hierarchies for image annotation in Section 3.2.

Other approaches have proposed to use semantic networks, built upon both conceptual and visual information, for image annotation [Fan et al., 2009, Wu et al., 2008, Wu et al., 2012, Tousch et al., 2012]. '*Flickr distance*' is proposed in [Wu et al., 2008], which is a new semantic similarity measure between the concepts in the visual domain. A Visual Concept NETwork (*VCNet*) based on the *Flickr distance* is also proposed in [Wu et al., 2008, Wu et al., 2012]. [Fan et al., 2009] proposed an algorithm to integrate the visual similarity and the contextual similarity

[7] In this dissertation, we make difference between concept hierarchies and semantic hierarchies within the image retrieval context. We refer to the former as a hierarchy illustrating the subsumption relationship between concepts from the conceptual perspective, while the second illustrates the same relationship from the (image) semantics perspective and should therefore take account of the visual similarity of concepts. More details are given in Chapter 3.

for a topic network generation. An image parsing to text description (*I2T*) framework is proposed in [Yao et al., 2010], which generates text descriptions for images and videos. *I2T* is mainly based on an And-or Graph for visual knowledge representation. Semantic lattices have also been explored in order to improve the image annotation [Belkhatir et al., 2004, Tousch et al., 2008, Martinet et al., 2011]. For instance, [Tousch et al., 2008] used a handmade semantic lattice describing the "Car" domain with the aim of classifying images according to a trade-off between semantic precision and accuracy. [Martinet et al., 2011] proposed to merge the vector space model of information retrieval with the conceptual graph (which has a lattice structure) formalism in order to provide a framework for relational information retrieval dedicated to images.

Discussion

Light-Weight Ontologies have shown to be very useful to narrow the semantic gap. Indeed, the aforementioned approaches have reported significant improvement in the image annotation accuracy. However, besides few work that addressed the problem of building semantic structures dedicated to image annotation, a basic problem with most of recent approaches is that they used either a specification based on textual information extracted from Wordnet, or a specification based on image features. Consequently, while these semantic structures are useful to provide a meaningful structure (organization) for image concepts, this organization does not necessarily reflect image semantics, since these structures were not originally built for this specific task. This is one of the fundamental assumptions of this thesis work and we will argue in more detail on this problem throughout this dissertation.

Moreover, most of recent work have attempted to use the incorporated knowledge (about image concepts) in these semantic structures in an implicit manner. Nevertheless, it is more convenient to use the explicit knowledge provided by these structures, and therefore, it is of prime importance to focus on this issue in order to exploit the strong inter-concepts correlation, and also the subsumption relationship that can help reasoning on the proper decision-making for image annotation.

2.5.2.4 Formal Ontologies

While ontologies often play a passive taxonomic role, some approaches consider ontologies as an active inference framework for image annotation and interpretation. These approaches focus on the use of formal semantics by constructing precise mathematical models of image semantics, i.e. explicit and formal modeling of knowledge about image context and the concepts.

Given the high expressivity and the well-defined inference services coming with them, Description logics appear today as an excellent candidate to support ontology representation and ontological reasoning. Description logics (DL) are a family of formal language for representing knowledge. They provide facilities for implementing knowledge bases, reasoning about their content, and manipulating them [Baader et al., 2003].

Definition 9 (Description logics)**.** *Description logics (DLs) are a family of knowledge representation formalisms that represent the knowledge of an application domain (the "world") by first defining the relevant concepts of the domain (its terminology), and then using these concepts to specify properties of objects and individuals occurring in the domain (the world description) [Baader et al., 2003].*

As illustrated in Figure 2.10, a DL knowledge base usually consists of three components, the *TBox*, the *RBox* and the *ABox*. The *TBox* and the *RBox* introduce the terminology, i.e. the vocabulary of an application domain, while the *ABox* contains assertions about named individuals in terms of this vocabulary. These notions are introduced in details in Section 4.4.

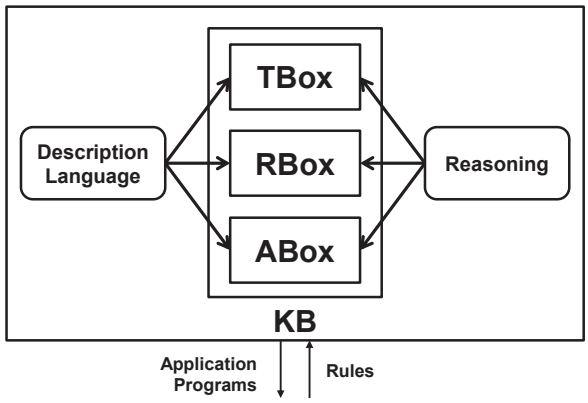


Figure 2.10: Architecture of a knowledge representation system based on Description Logics.

Formal ontologies[8] based approaches have successfully managed to exploit some interesting reasoning properties in the context of high-level image interpretation. These approaches make intensive use of contextual knowledge and inference rules with the aim of performing semantic image analysis, annotation and/or interpretation, ensuring thus the acquisition of interpretations that can match human cognition. In order to achieve these inference tasks, these approaches are based on sets of objects (or concepts) and relationships between them (called roles in DL). Axioms, which are a set of appropriate statements, are used to capture the conditions that need to be met by coherent and consistent states of the domain (interpretations). Image annotation/interpretation tasks are therefore formalized as:

1. Deduction: where the interpretation is an instantiation of formal knowledge consistent with evidence about the real-world domain [Hotz and Neumann, 2005, Hartz and Neumann, 2007, Hudelot et al., 2008, Dasiopoulou et al., 2009], or as

2. Abduction: where the interpretation is an instantiation of formal knowledge which allows to deduce the evidence [Shanahan, 2005, Town, 2006, Peraldi et al., 2007, Atif et al., 2011, Atif et al., 2013].

[8] A formal ontology is an ontology which is defined by axioms in a formal language.

Deductive and abductive reasoning were introduced as inference standards, where for deductive reasoning: if Σ is a logical theory and α a set of facts, through deduction is verified whether φ is logically entailed, that is whether $\Sigma, \alpha \models \varphi$. For abductive reasoning: given Σ and φ, abduction consists in finding an '*explanation*' α so that the entailment $\Sigma, \alpha \models \varphi$ is true. For example, let us say Σ is (visual or contextual) contextual knowledge regarding concept "Car", α is (visual or contextual) information extracted from an image containing a "Car" and φ is an instance of concept "Car", deduction is then, given $\Sigma, \alpha \models \varphi$.

For instance, in [Hudelot et al., 2008, Hudelot et al., 2010] an ontology of spatial relations is proposed to facilitate image interpretation. Indeed, [Hollink et al., 2004, Maillot et al., 2004, Hudelot et al., 2005] have reported that spatial relationships among objects and regions have appeared to be crucial in the concept detection process. [Mylonas et al., 2009] proposed an approach based on a visual thesaurus and visual context to improve concept detection. The authors introduce local (topological and unified) context in the analysis, to refine the confidence values of regions before taking decision. [Town, 2006] proposes an iterative process where low-level detections (induction) are compared with high-level models to derive new hypotheses (deduction). These can in turn guide the search for evidence to confirm or reject the hypotheses on the basis of expectations defined over the lower level features. In [Simou et al., 2008], a knowledge-assisted analysis architecture is proposed to perform the refinement of an initial set of over-segmented regions. They also used a fuzzy reasoning engine for the extraction of additional implicit knowledge and the improvement of region-based classification by incorporating spatial relations and neighborhood information. A comprehensive survey on the use of Description Logics for image annotation/interpretation is proposed in [Neumann and Möller, 2008, Möller and Neumann, 2008, Dasiopoulou and Kompatsiaris, 2010].

In this dissertation, we propose a new approach for building a formal multimedia ontology dedicated to image annotation and we also propose an approach for image annotation using ontological reasoning. Therefore, in order to highlight our contributions, we provide a more detailed discussion on related work in Section 4.2 and in Section 6.2.

Discussion

So far, we have seen that many recent work have proposed to use ontologies with the aim of improving the automatic image annotation, and knowledge based models have also been proposed in a tentative to achieve a consistent representation of image semantics. Specifically, formal ontologies are of great interest for image annotation, as they provide a formal framework containing explicit semantic definitions of concepts and their relations. These ontologies allow modeling contextual knowledge in a formal way, which allow deriving implicit knowledge by automatic inference.

However, most of current approaches for image annotation have not used all the expressiveness, nor the reasoning ability provided by ontologies. Most of them have proposed to use ontologies in order to define standards for the description of low-level multimedia content [Simou et al., 2005, Dasiopoulou et al., 2010], or for

providing semantic structures for the annotation vocabulary [Wei and Ngo, 2007], or as a hierarchical image classification frameworks [Marszalek and Schmid, 2007, Fan et al., 2008a, Deng et al., 2009], or also to define the semantic relationships between image concepts [Hollink et al., 2004, Hudelot et al., 2008]. Nevertheless, for the purpose of reducing the semantic gap in an effective manner, ontologies should be considered as an active inference framework for reasoning about image annotation and interpretation. Although interesting attempts have recently emerged [Dasiopoulou et al., 2008, Straccia, 2009, Hudelot et al., 2010], ontology-driven approaches for semantic image annotation are still in their infancy, and significant efforts have to be done before achieving effective systems for image annotation. In particular, the automatic building of such ontologies is rarely tackled.

2.6 General Discussion

As previously introduced, automatic image annotation is still an important challenge despite more than a decade of research. Indeed, many recent work have addressed this issue and have proposed new approaches with the aim of narrowing the semantic gap problem. In order to summarize the advantages and limitations of existing approaches for image annotation, we propose in Table 2.2 a comparison between the different families of approaches. As shown in this table, *text-based approaches* for image annotation provide a rich description of image content which may include the visual dimension, but also, subjective, spatial, temporal and social ones. However, text-based approaches are sensitive to the surrounding context which may be irrelevant with respect to image content. In particular, these approaches do not consider image content and therefore do not guarantee a consistent annotation, nor allow to process raw images which are widely spread nowadays (e.g. in photo sharing services).

As regards of *automatic image annotation approaches*, significant solutions have been proposed to reduce partially the semantic gap problem thanks to machine learning algorithm. These approaches have successfully managed to provide methods allowing to link the visual features of image to semantic concepts. However, it seems that the only use of machine learning algorithms is not sufficient to solve the image annotation problem. Indeed, these approaches face significant problems when dealing with large image databases (large in terms of image and concept number). Moreover, automatic image annotation approaches are limited to detect perceptual (visual) manifestations of semantics, i.e. they do not provide an explicit link between image features and concepts and they only consider concepts as classes (without any associated semantics).

Finally, *semantic image annotation approaches* have proposed to use either implicit knowledge (as the correlation between image modalities) or explicit knowledge (using knowledge models) about image context in order to achieve relevant decisions on the image annotation. Specifically, ontology-driven approaches for image annotation have shown to be appropriate for modeling image semantics as they provide explicit relationships between image concepts, between concepts and their meaning and sometimes between image features and concepts. These explicit relationships

Approaches	Pros	Cons
Text-Based Approaches for Image Annotation		
Text-Based Approaches	✓May describe the visual content of images, and also subjective, spatial, temporal and social dimensions. ✓Description closer to the human one.	✗ Can only process images with surrounding context, ✗ Sensitive to the surrounding context, ✗ Sensitive to the subjectivity of human perception, ✗ Do not take into account image content (may be incoherent w.r.t image content).
Content-Based Approaches for Image Annotation		
Automatic Image Annotation Approaches	✓Allow processing raw images without any surrounding information. ✓Retrieval is achieved using semantic concepts. ✓These concepts are extracted automatically from low-level features. ✓Sometimes description is less subjective than textual approaches. ✓Allow to narrow the semantic gap.	✗ Sometimes sensitive to image segmentation. ✗ Suffer from the scalability problem. ✗ Do not allow semantic interpretation of images. ✗ Do not adapt to the user background.
Semantic Image Annotation Approaches	✓May describe the visual content of images, and also subjective, spatial, temporal and social dimensions. ✓Allow processing raw images without any surrounding information. ✓Retrieval is achieved using abstract and semantic concepts. ✓Description less subjective to human perception. ✓Allow reasoning on image annotation and interpretation. ✓Allow to narrow the semantic gap. ✓Scale well with large databases.	✗ Difficult to implement. ✗ Not well explored. ✗ The reasoning power has not yet showed its efficiency. ✗ Ontological reasoning may be subject to the scalability problem.

Table 2.2: Comparison of the different families of approaches for image annotation.

	Semantic level (visual vs. semantics)	Reasoning capabilities (implicit vs. explicit Knowledge)	Scalability (image number)	Scalability (concept number)
Visual CBIR	-	-	++++	-
Concept Detector	++	+	++++	+
Supervised Learning	++	+	++	++
Unsupervised Learning	+	+	+++	+++
Semi-Supervised Lear.	+	+	+++	++
Hierarchical Classif.	++	++	++++	+++
Information Fusion	++	++	+++	+++
LWO	+++	+++	+++	+++
HWO	++++	+++	+	+
Formal Ontologies	++++	++++	++++	++

Table 2.3: Comparison of the different categories of approaches for image annotation.

are of great interest, since if they are used correctly could help to formalize the link between images and their semantic, and consequently allow reasoning on image annotation in order to achieve a consistent decision-making. However, the ontology building and ontological reasoning may be subject to the scalability problem.

In Table 2.3 and in Figure 2.11 we propose a comparison between the different categories of approaches for image annotation. The performance evaluation of each approach is quite subjective to our standpoint about existing approaches in the image annotation field, since these were not developed following experiments. However, this quantification is motivated by the results and the discussions reported in several significant work in our domain, and represents our synthesis of the state of the art. The Comparison criteria were the following:

- Semantic level: visual content vs. semantic description.
- Reasoning capabilities: implicit knowledge vs. explicit knowledge.
- Scalability problem according to image number.
- Scalability problem according to concept number.

Figure 2.11: Comparison of the different categories of approaches for image annotation.

As one can see, semantic image annotation approaches appear to be promising to help tackling the semantic gap problem in an efficient manner. Indeed, the semantic level provided by these approaches, and specifically by formal ontologies, is rich

enough to match user expectations in an image retrieval system. Formal ontologies allow also to dispose of good reasoning capabilities, which is useful to reduce the uncertainty problem introduced by machine learning algorithms and to reason about the consistency of image annotation. However, as shown in Table 2.3, ontology-driven approaches may suffer from the scalability problem. Indeed, in these knowledge models, 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 [Spaccapietra et al., 2004]. Consequently, an important issue in these approaches is to provide methods allowing to build ontologies dedicated to image annotation in an automatic manner.

According to our review of the state of the art, we present in the following our general standpoint about existing approaches for semantic image annotation and the main issues that motivated our contributions. These issues will be discussed deeply in the different chapters of this dissertation.

- Existing knowledge models used for image annotation are based either on a visual or textual specification (e.g. based on WordNet). However, these knowledge models, are they representative enough of image semantics? Would we not need a more elaborated specification (for example, **by combining visual and textual specification**) for modeling the image semantics?

- Ontology-driven approaches seem promising for narrowing the semantic gap. However, knowledge representation in these models is a hard task[9] and faces a major scalability problem. Therefore, should we not rather focus on providing methods that allow building ontologies (HWO, LWO or formal ontologies) **in automatic manner** in order to scale with a large number of concepts?

- Currently, concept hierarchies (and respectively visual taxonomies) are the most widely explored for semantic image annotation. These, constrain the reasoning to the inheritance relationships. Do we not need to enrich these structures with other semantic relationships as composition relationships, spatial, topological, etc., in order to benefit from a stronger reasoning power on **contextual knowledge**?

2.7 Conclusion

In this chapter, we have proposed a comprehensive survey of the recent work for image annotation. Our aim was not to provide an exhaustive survey of the state of the art approaches, nor to state that an approach is better than the others, but to introduce the different categories of approaches for image annotation and to underline the benefits and limits of each of them. This chapter also has highlighted the importance of contextual knowledge and explicit semantic structures for the image annotation problem.

[9]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 [Spaccapietra et al., 2004].

Indeed, the use of explicit semantic structures, such as semantic hierarchies and ontologies, seems to be essential to improve the image annotation and in order to narrow the semantic gap. In this dissertation, we are interested in semantic image annotation. These approaches aim at improving the image annotation by the use of images content and structured knowledge models about image semantics. This involves first to capture and to model suitable knowledge models about image context, and subsequently to build effective tools using these knowledge models in order to improve the image annotation process. The overall goal remains to narrow the semantic and sensory gaps using the available visual features of images and relevant knowledge sources, ultimately to satiate the user. Specifically, our concerns in this dissertation are as follows. Firstly, we address the problem of building explicit knowledge models dedicated to image annotation. Subsequently, we focus on the use of these knowledge models in order to produce a semantically consistent image annotation.

Part I

Building Structured Knowledge Models Dedicated to Image Annotation

Chapter 3

Building Semantic Hierarchies Faithful to Image Semantics

Contents

3.1 Introduction

In this chapter, we propose a new approach for automatically building semantic hierarchies dedicated to image annotation. Our approach is based on a new image-semantic measure, named "Semantico-Visual Relatedness of Concepts" (*SVRC*), which allows estimating the semantic similarity between image concepts. The proposed measure incorporates visual, conceptual and contextual information in order to provide a measure which is meaningful and representative of image semantics. We also propose a new methodology based on the previously proposed measure *SVRC* and on a new heuristic, named *TRUST-ME*, to connect the concepts with higher relatedness till the building of the final semantic hierarchy. The built hierarchy explicitly encodes, a general to specific, concept relationships, and therefore provides a semantic structure to concepts which facilitates the semantic interpretation of images. An evaluation of the effectiveness of the produced semantic hierarchy is presented in Chapter 5.

The rest of this chapter is structured as follows. In Section 3.2 we present the motivations of our proposal. We review some existing approaches and we emphasize their limitations. A discussion is also presented, where we provide a definition to what is meant by a semantic hierarchy suitable for image annotation. In Section 3.3 we introduce the proposed measure for computing the semantic relatedness between image concepts. Section 3.4 introduces the proposed rules for the building of semantic hierarchies dedicated to image annotation. Section 3.5 presents an application of the proposed method on the Pascal VOC'2010 dataset and illustrates the obtained hierarchy. Finally, the chapter is concluded in Section 3.6.

3.2 Motivation for Using Semantic Hierarchies

As stated in Chapter 2, current approaches for automatic image annotation face significant problems to narrow the semantic gap. Indeed, these approaches allow to adequately describe the visual content of images but are unable to extract image semantics like humans do, i.e. they are limited to describe the perceptual manifestation of image semantics. Moreover, they are subject to the scalability problem when dealing with broad content image databases [Liu et al., 2007, Deng et al., 2010].

A new trend to overcome the aforementioned problems is to use explicit semantic structures, such as semantic hierarchies and ontologies. Specifically, semantic hierarchies have shown to be very useful to narrow the semantic gap [Deng et al., 2010]. In particular, they provide a hierarchical semantic structure for image concepts that can be used as a framework for image classification. They also can be used as a formal framework for reasoning about the consistency of image annotation, and thus for reducing the uncertainty introduced by machine learning algorithms.

Several approaches have been proposed to use, or also to build, semantic hierarchies in order to improve image annotation. As introduced in Chapter 2, these approaches can be mainly categorized in three:

1. Language-based hierarchies: based on textual information[1,2] [Wei and Ngo, 2007, Marszalek and Schmid, 2007, Torralba et al., 2008],
2. Visual hierarchies (called also visual taxonomies): based on low-level image features [Marszalek and Schmid, 2008, Griffin and Perona, 2008, Sivic et al., 2008, Bart et al., 2008], and
3. Semantic hierarchies: based on both textual and visual features [Li et al., 2010, Fan et al., 2007].

3.2.1 Language-Based Hierarchies

Language-based hierarchies are concept hierarchies built using only conceptual information, i.e. based on a similarity measure between concepts calculated from textual data, or from a conceptual specification. These include commonsense ontologies, such as WordNet and Wikipedia, or any other concept hierarchy which does not take account of the visual information of images.

Currently, many existing approaches are based on WordNet [Fellbaum, 1998] for extracting the concept hierarchy that will serve thereafter for image annotation. For instance, [Marszalek and Schmid, 2007] proposed a semantic hierarchy classifier based on WordNet. Their hierarchy is built by extracting the relevant subgraph of WordNet that links all the concepts of the annotation vocabulary. The structure of this hierarchy is then used to train a set of hierarchical classifiers in order to perform image

[1] Examples of textual information used for building hierarchies are: tags, surrounding context, WordNet, Wikipedia, etc.

[2] In the rest of this dissertation, we refer to the information extracted from these textual sources as conceptual information or also conceptual semantics.

annotation. [Torralba et al., 2008] proposed a classification scheme using the Wordnet tree. Given a query image, they look for neighbors using some visual similarity measure. Then, each neighbor votes for its branch within the Wordnet tree. Votes are accumulated, from the entire sibling set, across a range of semantic levels. Classification is therefore performed by assigning to the query image the label with the most votes at the desired height within the tree, where the number of votes is acting as a measure of confidence in the decision. *ImageNet* is proposed in [Deng et al., 2009], which is a large-scale ontology of images built upon the backbone of WordNet. ImageNet aims at populating the 80 000 synsets[3] of WordNet with an average of 500-1000 manually selected images. [Deng et al., 2010] proposed a measure of similarity between categories based on WordNet, and claimed that the obtained results indicate a correlation between the structure of the concept hierarchy (of WordNet) and visual confusion between the categories. [Deng et al., 2011a] proposed a visual similarity function defined on semantic attributes of trained images using the WordNet hierarchy. The similarity between two concepts was defined according to the lowest common ancestor of these concepts. The reported results have shown that adding hierarchical knowledge has significantly increased retrieval performance.

Some other approaches have proposed to use the *LSCOM*[4] ontology, or to automatically build a binary concept hierarchy using an agglomerative algorithm on concept vectors [Wei and Ngo, 2007]. For instance, [Naphade et al., 2006] proposed *LSCOM*, a multimedia ontology which aims at designing a taxonomy with a coverage of around a 1000 concepts for broadcast news video retrieval. An Ontology-enriched Semantic Space (OSS) was built in [Wei and Ngo, 2007] to ensure globally consistent comparison of semantic similarities.

While these hierarchies are useful to provide a meaningful structure (organization) for concepts, they ignore the visual information which is an important part of image semantics. Furthermore, this kind of hierarchies are usually deep, and may contain many intermediate concepts which are not necessarily relevant in the image domain. For instance, it is may be hard to discriminate, at the perceptual level, the images of the following classes: ruminant vs. non-ruminant animals, or carnivore vs. herbivore animals, etc. Moreover, as it will be shown in our experiments, the hierarchy depth has a significant impact on the accuracy of the hierarchical classifiers - cf. Section 5.6. Figure 3.1 illustrates a hierarchy of concepts that we built by the extraction of the relevant subgraph of WordNet which links the 20 concepts of the Pascal VOC'2010 dataset. The building of the hierarchy is achieved by connecting first the most related concepts according to the shortest path in WordNet, then by connecting their hypernyms until reaching the root of WordNet. A disambiguation step is performed previously in order to retrieve the good synset (sense) that corresponds to each word (this method is quite similar to [Torralba et al., 2008], with the difference that they took the most common meaning of each word to transform the initial graph-structured relationships between words into a tree-structured one).

[3]Synonym Set: atomic component of WordNet, composed of a group of interchangeable words denoting a particular purpose or meaning. The meaning of the synsets is further clarified with short defining glosses. A concept may correspond to one or more synsets.

[4]http://www.lscom.org/

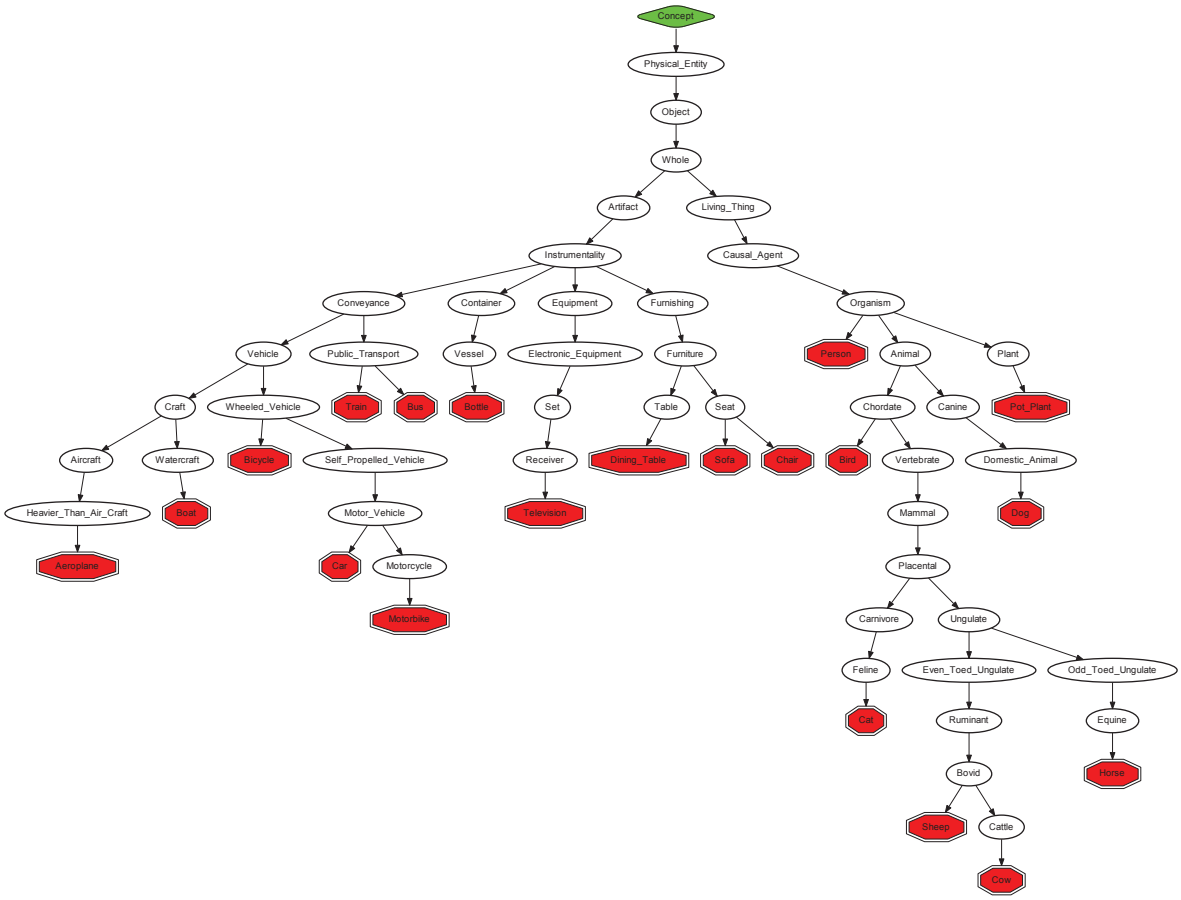


Figure 3.1: A concept hierarchy built by extracting the relevant subgraph of WordNet that links the 20 concepts of the VOC'2010 dataset. Double octagon nodes are the initial concepts and the diamond one is the root of the hierarchy.

3.2.2 Visual Hierarchies

Some other approaches have proposed to use visual information in order to build visual hierarchies (called also visual taxonomies) for the end-task of image classification [Fei-Fei and Perona, 2005, Sivic et al., 2008, Marszalek and Schmid, 2008, Bart et al., 2008, Gao and Koller, 2011]. For instance, [Fei-Fei and Perona, 2005, Griffin and Perona, 2008] proposed to automatically build a visual taxonomy for image classification, which consists in a hierarchy of classifiers built directly from the confusion matrix. The authors suggested using this taxonomy to increase the classification speed instead of performing a multi-class classification on all the categories. [Sivic et al., 2008] proposed to group visual objects using a multi-layer hierarchy tree that is based on common visual elements. The clustering is achieved by adapting to the visual domain, the generative Hierarchical Latent Dirichlet Allocation (hLDA) model [Blei et al., 2004]. [Bart et al., 2008] proposed a Bayesian method to organize a collection of images into a tree shaped hierarchy. [Marszalek and Schmid, 2008, Gao and Koller, 2011] adopted a relaxed hierar-

chy structure, where a set of binary classifiers are organized in a tree or DAG (Directed Acyclic Graph) structure. [Pujol et al., 2006] used a greedy sequential forward floating search algorithm to find two subsets of classes with the maximum mutual information between the feature distributions.

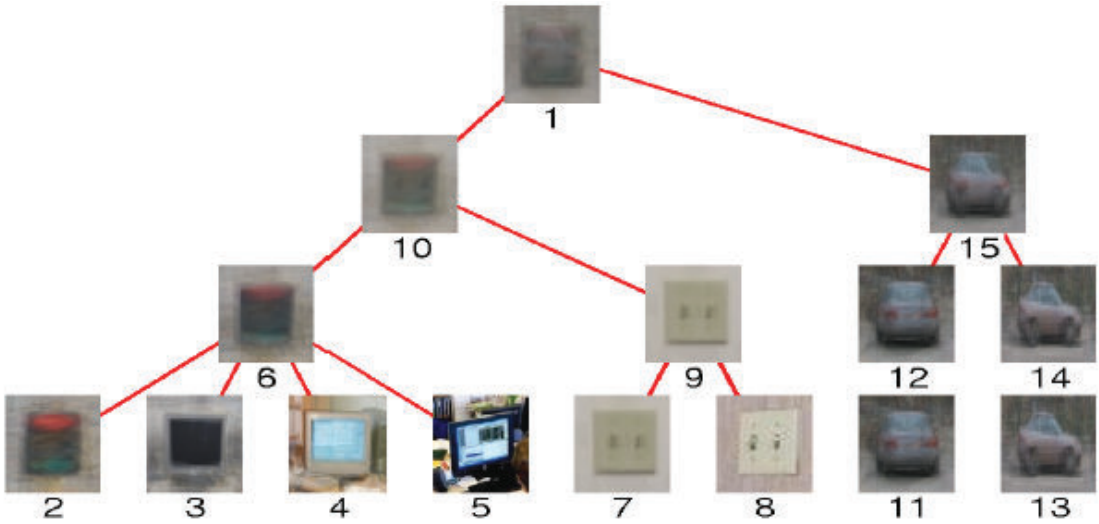


Figure 3.2: Obtained result by the method of [Sivic et al., 2008] for the building of visual object class hierarchies. The illustrated hierarchy is built on a subset of the LabelMe dataset [Russell et al., 2008].

However, these visual hierarchies face a major problem which is the lack of expressiveness at the semantic level. Indeed, these hierarchies are not usable at the semantic level since they do not carry any semantics, and are uninterpretable as illustrated in Figure 3.2. Consequently, these hierarchies serve only as visual taxonomies, and can only be used for hierarchical image classification in order to improve the classification accuracy. Hence, a good direction for building meaningful and suitable semantic hierarchies for the purpose of image annotation would be to make use of both information sources, conceptual and visual.

3.2.3 Semantic Hierarchies

Although the two first categories of hierarchies have received more attention, they showed a limited success in their general usage. Indeed, the conceptual semantics is not always consistent with the perceptual (visual) semantics of images, and is then insufficient to build good semantic structures for the purpose of image annotation [Wu et al., 2012]. Whereas the perceptual semantics cannot lead by itself to have a meaningful semantic hierarchy (cf. Figure 3.2), i.e. it is difficult to interpret these hierarchies in higher levels of abstraction. Therefore, it seems mandatory to combine the both components of image semantics, i.e. perceptual and conceptual, in order to build semantic hierarchies suitable for image applications.

For instance, a semantic hierarchy based on a visual similarity (between images) and a contextual similarity (between their tags) is proposed in [Fan et al., 2007,

Shen and Fan, 2010]. [Fan et al., 2008a] proposed an approach for hierarchical concept learning, by incorporating concept ontology and multi-task learning in order to exploit the strong inter-concept correlations. [Li et al., 2010] proposed a method based on visual features and image tags to automatically build a '*semantivisual*' image hierarchy.

To summarize, semantic hierarchies seem to have great potentials to improve the image annotation, mostly through their explicit representation of concepts relationships which may help understanding image semantics. However, to our knowledge, none of the existing approaches have used the reasoning power of these hierarchies in this context. Most of the proposals have focused on the problem of hierarchies building, or have proposed to use them as a hierarchical framework for image classification.

3.2.4 Discussion

As we have seen in the previous sections, many approaches have proposed to use WordNet in order to extract/build a hierarchical structure that will serve thereafter for hierarchical image classification. However, the use of WordNet in this way implicitly assumes that conceptual and visual information are tightly correlated. While this assumption might be true for many concepts, it does not hold for all the concepts of WordNet, and respectively for a large annotation vocabulary. Indeed, we believe that WordNet is not necessarily appropriate for modeling image semantics. The organization of concepts in WordNet follows a psycholinguistic structure, which may be useful for reasoning about the concepts and understanding their conceptual meaning, but it is limited and inefficient for reasoning about the image context or its content. More precisely, the distances between related concepts in WordNet do not necessarily reflect an appropriate semantic measure for annotating images or reasoning about them, i.e. the distances between concepts is not proportional to their semantic relatedness with respect to the image domain.

(a) Dolphin (b) Human (c) Shark

Figure 3.3: An example illustrating that conceptual measures are not always relevant with respect to the image domain. According to the WordNet density measure [Deng et al., 2010], dolphins (a) are more similar to humans (b), than to sharks (c).

Although [Torralba et al., 2008, Deng et al., 2010, Deselaers and Ferrari, 2011] assert have find a surprisingly strong correlation between purely linguistic metrics and

the performance of visual classification algorithms, we will argue in the rest of this chapter that their statement does not always hold, specifically in broad domain applications. For instance, let us consider the example depicted in Figure 3.3. For fairness, we used the same conceptual (linguistic) measure than [Deng et al., 2010], i.e. WordNet density $h(c_i, c_j)$, defined as the height of the lowest common ancestor of two concepts c_i and c_j. According to the WordNet density measure: $h(Human, Dolphin) = 5$, $h(Shark, Human) = 5$ and $h(Shark, Dolphin) = 7$. This implies that concept "Dolphin" is closer to "Human" than to "Shark", which is consistent from a biological point of view since "Dolphin" and "Human" are mammal while "Shark" is not. However, in the image domain it is more accurate to have higher similarity between concepts "Shark" and "Dolphin" as they live in the same environment, share similar visual features, and for sure share a higher value in the confusion matrix. Therefore, a suitable semantic hierarchy should represent this information or allow deducing it in order to help understanding the image semantics. To sum up , we propose in Table 3.1 a comparison of the different types of hierarchies used for annotating images.

3.3 Proposed Measure: Semantico-Visual Relatedness of Concepts (*SVRC*)

Based on the previous discussion, we define the following assumptions which motivated the design of our approach:

Assumption 1. *A suitable semantic hierarchy for image annotation should:*

A1- model image context, and consequently incorporates all aspect of image semantics, i.e. at least the visual and conceptual information between image concepts.

A2- reflect image semantics, i.e. the organization of concepts into the hierarchy and their semantic relatedness reflect image semantics - cf. Chapter2 - Section 2.3.

A3- allow grouping visually similar concepts in order to obtain better performance of classifiers.

Following the above assumptions, we propose in the remaining of this chapter a new method for building semantic hierarchies dedicated to image annotation. Our approach is based on a new measure to estimate the semantic relatedness between concepts, which is more faithful to image semantics since it is based on its different modalities. As illustrated in Figure 3.4, this measure, named *SVRC*, is based on 1) a visual similarity which represents the visual correspondence between concepts, 2) a conceptual similarity which defines a relatedness measure between target concepts, based on the concept definition in WordNet, and 3) a contextual similarity which measures the distributional similarity between each pair of concepts. *SVRC* is then used in *TRUST-ME*, a set of heuristic rules that allow deciding about the likelihood of the semantic relatedness between concepts, and help building the final semantic hierarchy.

	Visual Hierarchies	Language-Based Hierarchies	Semantic Hierarchies
Used Information	Visual	Conceptual	Visual + (Conceptual +/or Contextual)
Significance (Semantics)	Meaningless	Meaningful, but not for the image domain	Meaningful
Uses	Image classification	Image classification , Concepts disambiguation, Reasoning (consistency checking), Multi-level image annotation.	Image classification , Concepts disambiguation, Reasoning (consistency checking), Multi-level image annotation.
Pros	Easy to implement. Efficient with a limited vocabulary.	Improve the classification accuracy. Can produce explanation of the classification decision. Useful (see the cell above).	*Misclassified images are usually affected to (visually and semantically) related concepts.* Improve the classification accuracy. Can produce explanation of the classification decision. Useful (see the cell above).
Cons	Classification decisions are hard to justify. Do not allow reasoning. Subject to the scalability problem. Highly dependent on the training dataset.	Misclassified images may be affected to semantically unrelated concepts. Scalability with a large annotation vocabulary/large scale databases not demonstrated.	Scalability with a large annotation vocabulary/large scale databases not demonstrated.
		Lack of metrics for comparing these hierarchies from a semantic standpoint.	

Table 3.1: Comparative table of the different hierarchies used for image annotation.

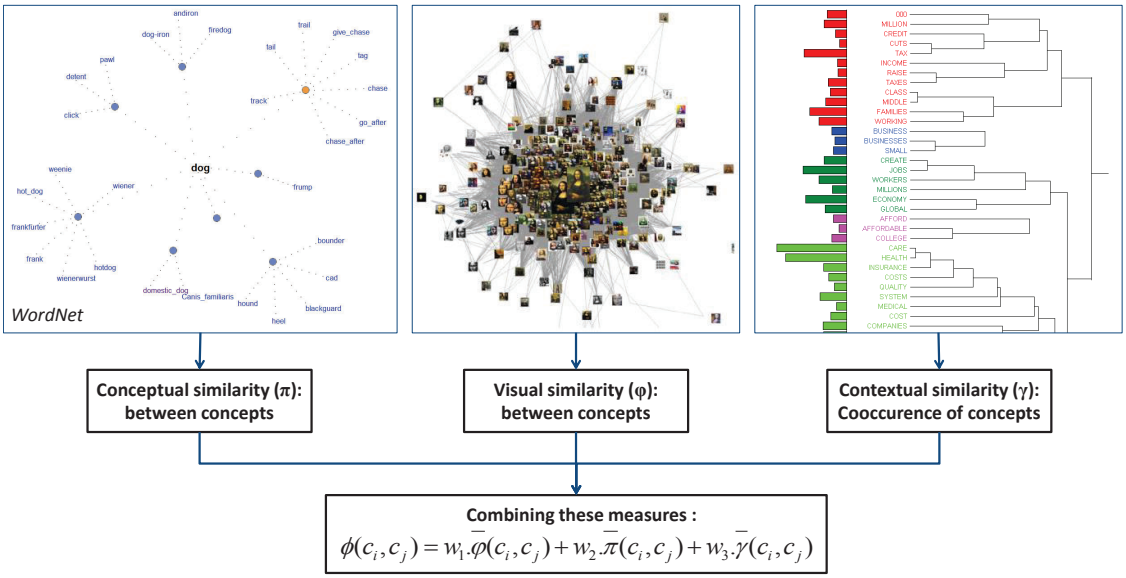


Figure 3.4: Overview of the *SVRC* measure which is based on visual, conceptual and contextual similarities.

3.3.1 Problem Formalization

In the following we introduce a formal description of our problem.

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 a set of concepts $\{c_j \in \mathcal{C}, j = 1..n_{i_i}\}$ associated with a given image $i_i \in \mathcal{DB}$.
- $\mathcal{CO}$, a generic commonsense ontology containing $\mathcal{N}'$ concepts ($\mathscr{C}$), such that $\mathcal{C} \subseteq \mathscr{C}$. For this work, we used WordNet as a commonsense ontology.

Our objective is to build a semantic hierarchy ($\mathcal{SH}$), 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. The built semantic hierarchy should consider the previously defined assumptions 1. The proposed approach consists therefore in identifying $\mathcal{C}'$ new concepts that link (and subsume) all the concepts of $\mathcal{C}$ in a hierarchical structure that accurately represents image semantics. Figure 3.5 illustrates the architecture of our approach for building semantic hierarchies dedicated to image annotation.

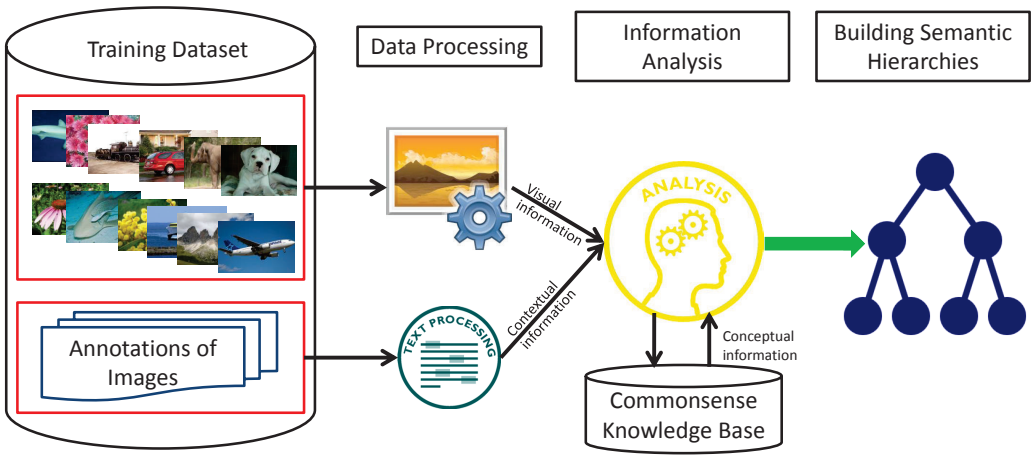


Figure 3.5: From image data to structured knowledge models: Architecture of our approach for building semantic hierarchies dedicated to image annotation.

3.3.2 Visual Similarity Between Concepts

The visual similarity between concepts allows estimating the visual correlation between two concepts. Many recent approaches have proposed to measure this similarity using:

- the confusion matrix [Griffin and Perona, 2008, Bengio et al., 2010, Deng et al., 2010],
- Kullback-Leibler (KL) divergence [Fan et al., 2008b], or its variant Jensen-Shannon (JS) divergence [Wu et al., 2008],
- the common features shared across the classes [Torralba et al., 2007], or using
- canonical correlation analysis between the image sets of targeted concepts [Fan et al., 2009].

In our approach, we use a simple but effective method for estimating the visual correlation between concepts. Indeed, we propose to compute for each visual concept a centroid assumed to be representative of its visual appearance. These centroids are computed using the set of support-vectors defining each SVM classifier associated with a given concept. Therefore, the visual similarity between two given concepts is computed as the inverse Euclidean distance between their centroids. In the following, we detail the proposed visual similarity.

Let x_i^v be any visual representation of an image i_i (a visual feature vector), we train for each concept c_j a classifier that can associate this concept with its visual features. For this, we use $\mathcal{N}$ binary Support Vector Machines (SVM) [Cortes and Vapnik, 1995] (One-Versus-All) with a decision function $\mathcal{G}(x_i^v)$:

$$\mathcal{G}(x_i^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x_i^v) + b \tag{3.1}$$

where $\mathcal{N} = |\mathcal{C}|$ is the size of the annotation vocabulary, $\mathbf{K}(x_k^v, x_i^v)$ 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. Note that the training samples x_k^v with weight $\alpha_k > 0$ are the *support vectors*.

After several tests performed on the training samples in order to find the kernel function which gives the best result for defining our classifiers, we decided to use a radial basis function kernel:

$$\mathbf{K}(x_k^v, x_i^v) = exp\Big(\frac{\|x_k^v - x_i^v\|^2}{\sigma^2}\Big) \tag{3.2}$$

Now, given these $\mathcal{N}$ trained SVM where the inputs were the visual features of images and the outputs were the concepts $c_i \in \mathcal{C}$, we want to define a centroid $\vartheta(c_i)$ for each concept class c_i that best represents it. These centroids should then minimize the sum of squares within each set S_i as illustrated in Figure 3.6:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{\mathcal{N}} \sum_{x_j^v \in S_i} \|x_j^v - \mu_i\|^2 \tag{3.3}$$

where S_i is the set of *support vectors* of class c_i, $S = \{S_1, S_2, \cdots, S_{\mathcal{N}}\}$, and μ_i is the arithmetic mean of vectors in S_i.

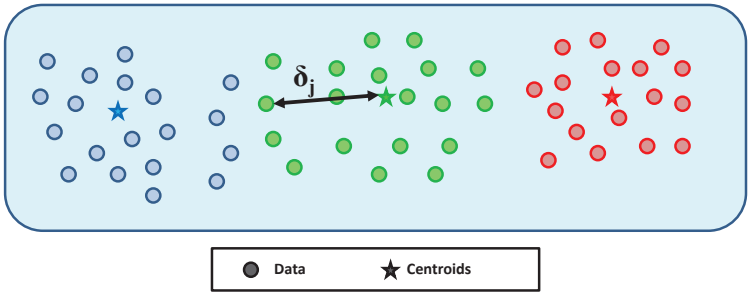


Figure 3.6: Objective: for each concept $c_i \in \mathcal{C}$, find a centroid $\vartheta(c_i)$ that minimizes δ_j (δ_j is the intra-class variance) for all $x_j^v \in S_i$.

Our objective being to estimate a distance between these concepts $c_i \in \mathcal{C}$ in order to assess their visual similarities, we compute the centroid $\vartheta(c_i)$ of each visual concept c_i as follows:

$$\vartheta(c_i) = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j^v \tag{3.4}$$

The visual similarity between two concepts c_i and c_j, is then inversely proportional to the distance between their visual features (centroids) $\vartheta(c_i)$ and $\vartheta(c_j)$:

$$\varphi(c_i, c_j) = \frac{1}{1 + d(\vartheta(c_i), \vartheta(c_j))} \tag{3.5}$$

where $d(\vartheta(c_i), \vartheta(c_j))$ is the Euclidean distance between $\vartheta(c_i)$ and $\vartheta(c_j)$.

3.3.3 Conceptual Similarity

Conceptual similarity reflects the semantic relatedness between two concepts from a linguistic and/or a taxonomic point of view. Several conceptual similarity measures have been proposed [Resnik, 1995, Banerjee and Pedersen, 2003, Budanitsky and Hirst, 2006, Popescu and Grefenstette, 2011]. Most of them are based on a lexical resource, such as WordNet [Fellbaum, 1998]. A first family of approaches is based on the structure of this external resource (often used as a semantic network or a directed graph), and the similarity between concepts is computed according to the distances of the paths connecting them in this structure [Budanitsky and Hirst, 2006]. However, as introduced in Section 3.2.4, the structure of these resources does not necessarily reflect image semantics, and therefore such measures do not seem suited to our problem.

An alternative approach to measure the semantic relatedness between concepts is to use their provided definitions. With respect to the WordNet resource, these definitions are known as *glosses*[5] and are provided by the *synsets*[3] associated with each concept. For example, [Banerjee and Pedersen, 2003] proposed a measure of semantic relatedness between concepts that is based on the number of shared words (overlaps) in their definitions (glosses).

In this work we used the gloss vector relatedness measure proposed by [Patwardhan and Pedersen, 2006], in which they suggest to use "second order" co-occurrence vector of glosses rather than matching words that co-occur in it. Specifically, in a first step a word space of size $\mathcal{P}$ is built by taking all the significant words used to define all synsets of WordNet. Thereby, each concept c_i is represented by a context vector $\overrightarrow{w}_{c_i}$ of size $\mathcal{P}$, where each n^{th} element of this vector represents the number of occurrences of n^{th} word in the word space in the gloss of c_i. The semantic relatedness of two concept c_i and c_j is therefore measured using the cosine similarity between $\overrightarrow{w}_{c_i}$ and $\overrightarrow{w}_{c_j}$:

$$\eta(c_i, c_j) = \frac{\overrightarrow{w}_{c_i} \cdot \overrightarrow{w}_{c_j}}{|\overrightarrow{w}_{c_i}||\overrightarrow{w}_{c_j}|} \tag{3.6}$$

Some concepts definitions in WordNet are very concise which could make this measure unreliable. Consequently, in the same spirit as [Patwardhan and Pedersen, 2006], we propose to extend the glosses of the target concepts with the glosses of their adjacent concepts (located in their immediate neighborhood). Hence, for each concept c_i, we define Ψ_{c_i} as the set of all adjacent glosses connected to c_i, i.e. (Ψ_{c_i}=$\{$gloss(c_i), gloss(hyponyms(c_i)), gloss(meronyms(c_i)), etc.$\}$). Then, each element (gloss), denoted x, of Ψ_{c_i} is represented by $\overrightarrow{w}_x$ as explained above. The similarity measure between two concepts c_i and c_j is therefore defined as the sum of the individual

[5]Glosses: each synset contains a short defining glosses (definitions and/or example sentences).

cosines of the corresponding gloss vectors:

$$\theta(c_i, c_j) = \frac{1}{|\Psi_{c_i}|} \sum_{x \in \Psi_{c_i}, y \in \Psi_{c_j}} \frac{\overrightarrow{w}_x \cdot \overrightarrow{w}_y}{|\overrightarrow{w}_x||\overrightarrow{w}_y|} \ , \quad where |\Psi_{c_i}| = |\Psi_{c_j}|. \tag{3.7}$$

Finally, each concept $c_i \in \mathcal{C}$ may be associated to one or several synsets in WordNet. Usually, these synsets have different meanings, and thus differ from each other in their definitions (glosses) and in their positions in the hierarchy of WordNet. A disambiguation step is then mandatory in order to associate to each concept $c_i \in \mathcal{C}$ the good synset with respect to its semantics. For instance, the similarity between "Mouse" (Animal) and "Keyboard" (device) widely differs from the one of "Mouse" (device) and "Keyboard" (device). In our approach, the disambiguation is performed as follows. First, we compute the conceptual similarity between the different senses (synsets) of c_i and c_j. The maximum likelihood of these similarities is then used to identify the most likely meaning of these two concepts, i.e. disambiguate c_i and c_j. Consequently, the conceptual similarity is calculated as following:

$$\pi(c_i, c_j) = \underset{\delta_i \in s(c_i), \delta_j \in s(c_j)}{\operatorname{argmax}} \theta(\delta_i, \delta_j) \tag{3.8}$$

where $s(c_x)$ is "the set of all synsets that can be associated to the meanings of c_x".

3.3.4 Contextual Similarity

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). The information related to the context of appearance of concepts, called *contextual information*, is used to connect concepts that often appear together in images although semantically distant from the taxonomic point of view. Moreover, this contextual information can also help to infer higher-level knowledge from images. For example, if a photo contains "Sea" and "Sand", it is likely that the scene depicted in this photo is the one of "Beach". It is therefore important to measure the contextual similarity between concepts.

However, unlike the visual and the conceptual similarity, the contextual similarity is a *corpus-dependent* measure, and more precisely depends on the distribution of concepts in the corpus. Therefore, it is important to ensure an equivalent distribution of concepts within the training set, or to normalize the produced measure in order to perform a good contextual measure. Otherwise this measure will be biased from the outset and will be unreliable.

In our approach, we define the contextual similarity between two concepts c_i and c_j as the Pointwise Mutual Information [Church and Hanks, 1990] (PMI), denoted $\rho(c_i, c_j)$, and computed as follows:

$$\rho(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \tag{3.9}$$

where: $P(c_i)$ is the probability of occurrence of c_i, and $P(c_i, c_j)$ is the joint probability of c_i and c_j. These probabilities are estimated by computing the frequency of occurrence and co-occurrence of concepts c_i and c_j in the database.

Given $\mathcal{N}$ the total number of concepts in the database, $\mathcal{L}$ the total number of images, n_i 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 and c_j, the above probabilities can be estimated by:

$$\begin{aligned} \widehat{P(c_i)} &= \frac{n_i}{\mathcal{L}} \\ \widehat{P(c_i, c_j)} &= \frac{n_{ij}}{\mathcal{L}} \end{aligned}$$

As a consequence:

$$\rho(c_i, c_j) = \log \frac{\mathcal{L} * n_{ij}}{n_i * n_j} \tag{3.10}$$

$\rho(c_i, c_j)$ 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 and c_j are negatively correlated. Otherwise, $\rho(c_i, c_j) > 0$ and quantifies the degree of dependence between these two concepts.

In this work, we only want to measure the positive dependence between the concepts and therefore we set negative values of $\rho(c_i, c_j)$ to 0. Finally, to normalize the contextual similarity between two concepts c_i and c_j into [0,1], we compute it in our approach by:

$$\gamma(c_i, c_j) = \frac{\rho(c_i, c_j)}{-\log[\max(P(c_i), P(c_j))]} \tag{3.11}$$

3.3.5 Computation of the Semantico-Visual Relatedness of Concepts

For two given concepts c_i and c_j, their similarity measures: visual $\varphi(c_i, c_j)$, conceptual $\pi(c_i, c_j)$ and contextual $\gamma(c_i, c_j)$ are first normalized into the same interval using the Min-Max Normalization [Jain et al., 2005]. Then, the Semantico-Visual Relatedness $\phi(c_i, c_j)$ of these concepts c_i and c_j is computed as:

$$\phi(c_i, c_j) = \omega_1 \cdot \overline{\varphi}(c_i, c_j) + \omega_2 \cdot \overline{\pi}(c_i, c_j) + \omega_3 \cdot \overline{\gamma}(c_i, c_j) \tag{3.12}$$

where $\sum_{i=1}^{3} \omega_i = 1$. $\overline{\varphi}(c_i, c_j)$, $\overline{\pi}(c_i, c_j)$, and $\overline{\gamma}(c_i, c_j)$ are the normalized visual similarity, conceptual similarity and contextual similarity.

The choice of the weights ω_i is very important. According to the target application, some would prefer to build a domain-specific hierarchy (that best represents a specific-domain or corpus), and can therefore assign a higher weight to the contextual similarity ($\omega_3 \nearrow$). Others would be conducted to build a generic hierarchy, and will therefore assign a higher weight to the conceptual similarity ($\omega_2 \nearrow$). However, if the purpose of the semantic hierarchy is rather to build a hierarchical framework to

image classification, it may be advantageous to assign a higher weight to the visual similarity ($\omega_1 \nearrow$).

The building of semantic hierarchies using our approach remains flexible, depending on the sought image semantics and the targeted application. However, the recommended method to find the best weights ($w_i, i \in [1,3]$) needed to compute the *SVRC* measure is to use cross-validation. Indeed, using the cross-validation, it is possible to assess the impact of weights distribution as a function of the system performance. Thus, finding the best weights to achieve the best performance of the system. More details on the impact of these weights on our "Semantico-Visual" measure, and the choice of the weighting factor by cross-validation are proposed in Section 3.5.2.

3.4 Rules for the Hierarchy Building

Once we have estimated the semantic relatedness between each pair of concepts, it is important to group them in a more comprehensive hierarchy despite the uncertainty introduced by semantic similarity measurements. In the following, we propose a heuristic named *TRUST-ME*, that allows to infer hypernym relationships between concepts, and to bring together these various concepts in a hierarchical structure.

Let us define the following functions to understand the reasoning rules we used for the building of our hierarchy:

- $Closest(c_i)$ returns the closest concept to c_i according to the *SVRC* measure:

$$Closest(c_i) = \underset{c_k \in \mathcal{C} \setminus \{c_i\}}{\operatorname{argmax}} \ \phi(c_i, c_k) \tag{3.13}$$

- $LCS(c_i, c_j)$ allows to find the *Least Common Subsumer* of c_i and c_j in WordNet:

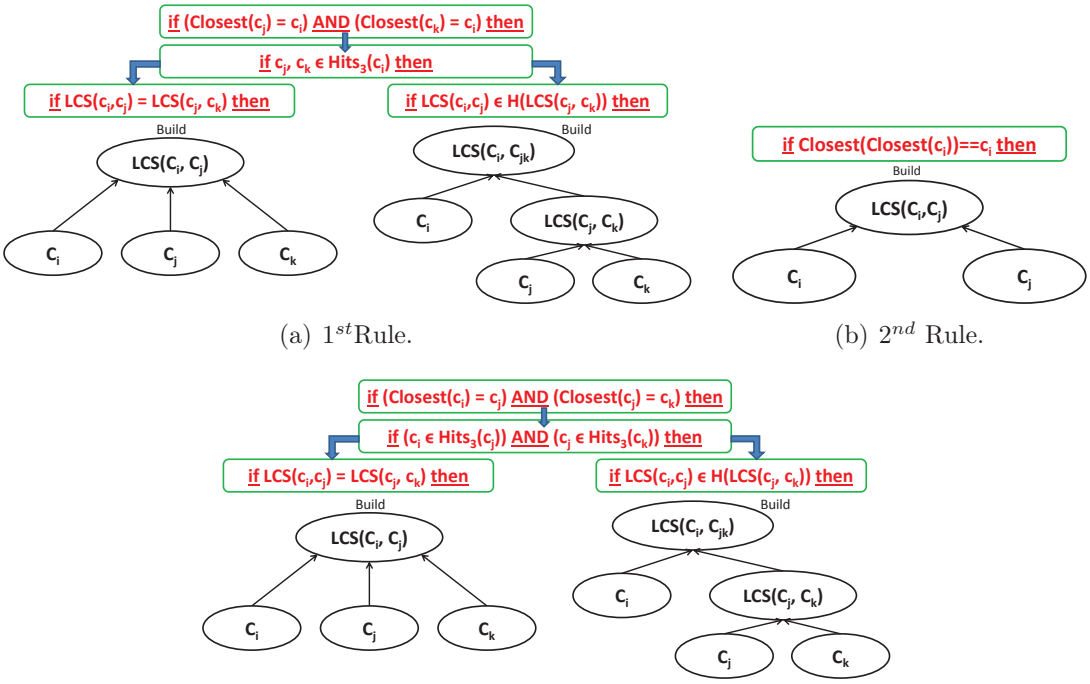
$$LCS(c_i, c_j) = \underset{c_k \in \{H(c_i) \cap H(c_j)\}}{\operatorname{argmax}} len(c_k, root) \tag{3.14}$$

  where $H(c_i)$ is a function allowing to find the set of all hypernyms of c_i in WordNet, *root* is the root node of WordNet, and $len(c_x, root)$ returns the length of the shortest path in WordNet between the concept c_x and the *root*.

- $Hits_3(c_i)$ is a function returning the 3 closest concepts to c_i within the meaning of $Closest(c_i)$.

Basically, *TRUST-ME* consists of three rules which are based on the *SVRC* measure and on the reasoning about the Least Common Subsumer (LCS) to select the concepts to be connected to each other. These rules are illustrated and executed in the order described in Figure 3.7.

Rule 1. The first rule checks whether a concept c_i is classified as the closest relative to more than one concept, i.e. $(Closest(c_j) = c_i), \forall j \in \{1, 2, \cdots\}$. In this case, if these concepts $\{c_j, \forall j \in \{1, 2, \cdots\}\}$ are reciprocally in $Hits_3(c_i)$, then according to their LCS they will be connected either directly to their LCS or in a two levels structure as illustrated in Figure 3.7(a).



(a) 1^{st}Rule. (b) 2^{nd} Rule.

(c) 3^{rd} Rule.

Figure 3.7: Rules in *TRUST-Me* allowing to infer the relatedness relationships between the different concepts. Preconditions (in red) and actions (in black).

Rule 2. In the second rule, if $(Closest(c_i) = c_j)$ and $(Closest(c_j) = c_i)$ (can also be written as $Closest(Closest(c_i)) = c_i$), then c_i and c_j are actually related and are connected to their LCS - cf. Figure 3.7(b).

Rule 3. The third rule covers the case when $(Closest(c_i) = c_j)$ and $(Closest(c_j) = c_k)$. In this case, if $c_i \in Hits_3(c_j)$ and $c_j \in Hits_3(c_k)$, then according to the LCS of c_i, c_j and c_k, these concepts will be connected either directly to their LCS or in a tow levels structure as illustrated in Figure 3.7(c).

The building of the semantic hierarchy is a bottom-up process (starts from the leaf concept nodes), and uses an iterative algorithm until reaching the root node. Given a set of concepts associated to a set of images in a dataset, our method compute the *SVRC* $(\phi(c_i, c_j))$ between all pairs of concepts, then links most related concepts to each other while respecting the defined rules in *TRUST-ME* - cf. Algorithm 1. Thus, we obtain a new set of concepts in a higher level resulted by the linked concepts in the lower level. We iterate the process until all the concepts are linked to a root node. In Section 3.5, we illustrated some semantic hierarchies built using our method, in particular on VOC'2010 dataset.

The complexity of our method for building hierarchies is of $\mathcal{O}(\mathcal{N}^2)$, where $\mathcal{N}$ is the size of the initial annotation vocabulary. Indeed, our method requires two steps: i) computing the similarity between each pair of concepts, which is performed in $\mathcal{O}(\mathcal{N}^2)$, and ii) connecting each pair of concepts to the *least common subsumer* using the rules illustrated in Figure 3.7, which is performed in the worst-case in $\mathcal{O}(\mathcal{N}log(\mathcal{N}))$.

Algorithm 1: Semantic Hierarchy Building

Data: Images and their annotation
Result: A semantic hierarchy
a set of trained classifiers for each concept node in the hierarchy
begin
- $\forall\ c_i \in \mathcal{C}$, $\top(concepts) \leftarrow c_i$
foreach $(c_i, c_j \in \top(concepts))$ **do**
- Compute the visual similarity $\varphi(c_i, c_j)$
- Compute the conceptual similarity $\pi(c_i, c_j)$
- Compute the contextual similarity $\gamma(c_i, c_j)$
- Compute the SVRC $\phi(c_i, c_j)$
end
while $|\top(concepts)| > 1)$ **do**
$\bot(concepts) \leftarrow \top(concepts)$
$\top(concepts) \leftarrow \emptyset$
- Connect one level of the semantic hierarchy using *TRUST-ME*:
$\top(concepts) \leftarrow$ TRUST-ME$(\bot(concepts))$
foreach $(c_i, c_j \in \top(concepts))$ **do**
- Compute $\phi(c_i, c_j)$ by averaging the distances of hyponyms of c_i to hyponyms of c_j
end
end
end
$*\ \bot$: stands for lower level in the hierarchy
$*\ \top$: stands for higher level in the hierarchy

3.5 Experimental Results

As part of this chapter, we will only illustrate the obtained hierarchies using the above described method on the Pascal VOC'2010 dataset [Everingham et al., 2010], and the impact of the weighting factors on the consistency of the produced hierarchies. We also evaluate the distribution of the proposed similarities on the different image modalities, i.e. visual, conceptual and contextual similarity. An experimental evaluation of the produced hierarchies within a framework of hierarchical image classification is performed in Chapter 5.

3.5.1 Visual Representation of Images

To compute the visual similarity of concepts, we used in our approach the Bag-of-Features (BoF) model, also known as bag-of-visual words, which is a widely known method [Li and Perona, 2005]. The BoF model has shown excellent performances and became one of the most widely used techniques for image classification and object recognition. The BoF model consists of three steps: feature detection, feature description and the codebook generation. The feature detection process is performed

by the extraction of several local patches (or regions), which will be considered as visual words. In our approach, the used BoF model is built as follows. Lowe's DoG Detector [Lowe, 1999] is used for detecting a set of salient image regions. A signature of these regions is then computed using SIFT descriptor [Lowe, 1999]. The SIFT descriptor has proofed to be invariant to intensity, rotation, scale and affine variations to some extent, which makes it very robust to feature description. This descriptor converts each patch to 128-dimensional vector describing gradient magnitudes and orientations around the keypoint. Each image in the database is then represented as a collection of non-ordered vectors of the same dimension (128 for SIFT).

Afterwards, given the collection of detected patches (or regions) from the training dataset of all categories, we generate a codebook of size $K = 1000$ by performing the k-means algorithm. Since the K-means algorithm is sensitive to initialization, we control clusters after each iteration to ensure that we do not have empty ones. Empty clusters are again randomly initialized. Finally, the generated codebook is a set of features assumed to be representative of all images features. Thus, each patch (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 D visual words, where each bin in the histogram correspond to the occurrence number of a visual word in that image.

3.5.2 Impact of Weighting on the Building of Hierarchies

This work aims at building meaningful semantic hierarchies for the purpose of image annotation. Thus, for our experiments we set the weighting factors in an **experimental** way as follows: $\omega_1 = 0.4$, $\omega_2 = 0.3$, and $\omega_3 = 0.3$. Indeed, we used cross-validation in order to retrieve the best set of weights that allows to maximize the accuracy of hierarchical image classification. The adjustment of weighting was performed using a step of 0.1. In fact, during our experiments we observed that the use of a lower step for adjusting weights ($\triangle_{\omega_i} < 0.1$) does not have a significant impact on the produced hierarchy, but the counterpart is an execution time which grows exponentially. We also underline that the accuracy of image classification depends mostly on the structure of the hierarchy and not on the used weights.

Furthermore, our experimentations on the impact of weights (ω_i) showed also that the visual similarity is more representative of concepts similarity, as it will be illustrated with the produced hierarchies in Figure 3.8, Figure 3.11 and Figure 3.10, and with the heat maps in Figure 3.12.

In order to illustrate our approach, we show in Figure 3.8 the built semantic hierarchy using the weighting factors described above, and in Figure 3.9 the buit binary hierarchy based on our semantic-visual relatedness of concepts (SVRC) and the second rule of TRUST-ME. The semantic hierarchy is built under the shape of a taxonomy, i.e. a concept c_i may have several hyponyms: $hyponyms(c_i) = \{c_f,\ 0 \leq f \leq |\mathcal{C}|\}$. The binary hierarchy is built by restricting the number of child nodes of each concept to 2, i.e. $hyponyms(c_i) = \{c_f,\ 0 \leq f \leq 2\}$. As can be seen in these two figures, the structure of the hierarchy and the relationships between concepts seem to be coherent at both, conceptual and visual point of view.

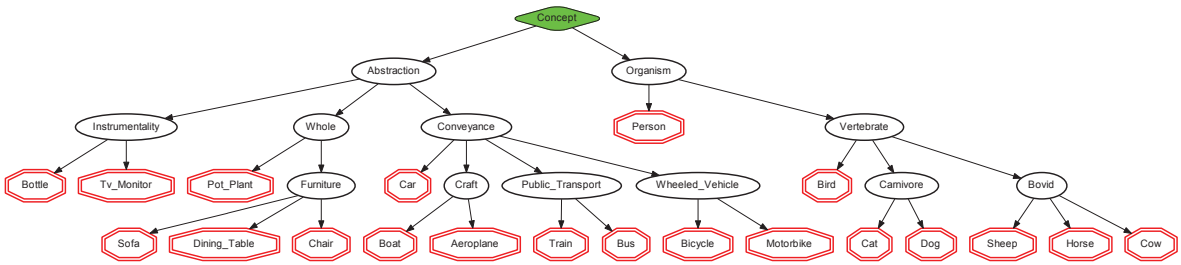


Figure 3.8: The semantic hierarchy built on Pascal VOC'2010 dataset using the proposed method. Double octagon nodes are the original concepts and the diamond one is the root of the produced hierarchy.

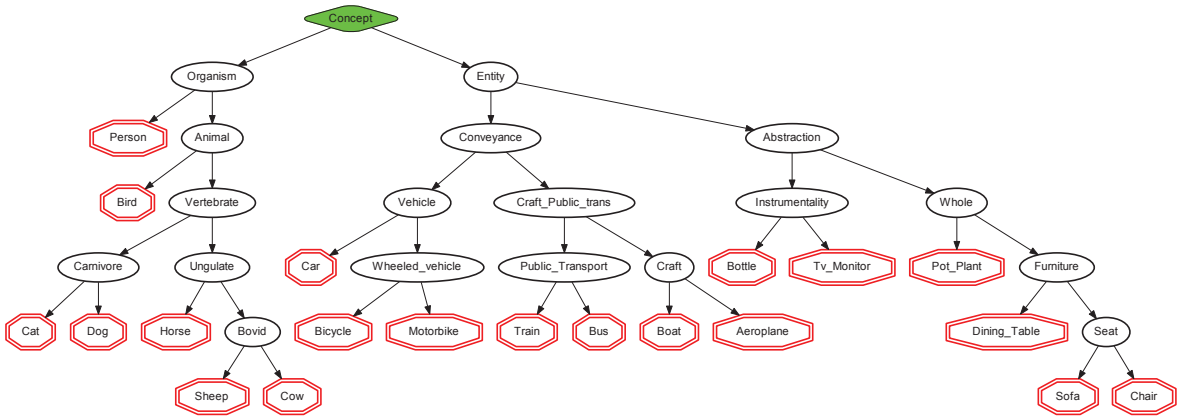


Figure 3.9: A *binary* semantic hierarchy built on Pascal VOC'2010 dataset using our measure *SVRC* and *TRUST-ME*. Double octagon nodes are the original concepts and the diamond one is the root of the produced hierarchy.

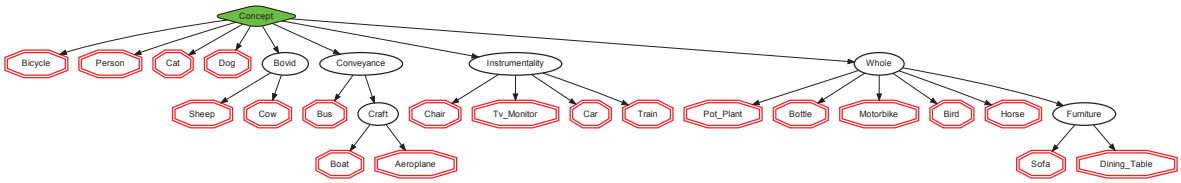


Figure 3.10: A hierarchy of concepts built on Pascal VOC'2010 dataset using *TRUST-ME* and the visual similarity ($\phi(c_i, c_j) = \varphi(c_i, c_j)$).

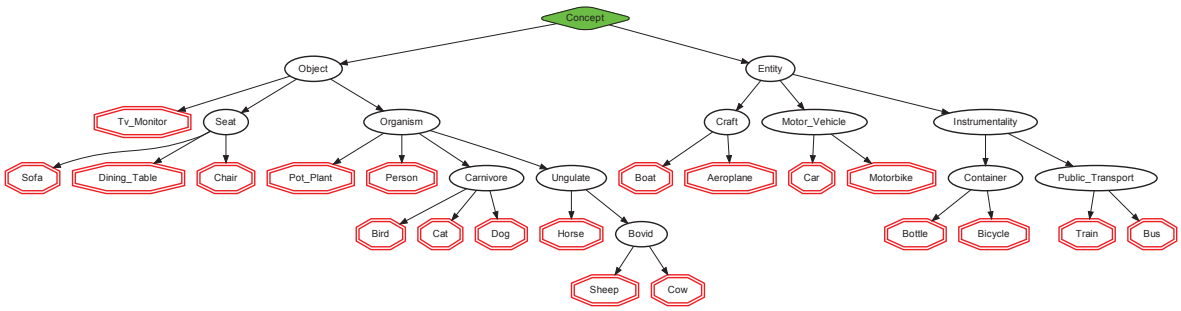


Figure 3.11: A hierarchy of concepts built on Pascal VOC'2010 dataset using *TRUST-ME* and a similarity measure based on the shortest path in WordNet between the considered concepts.

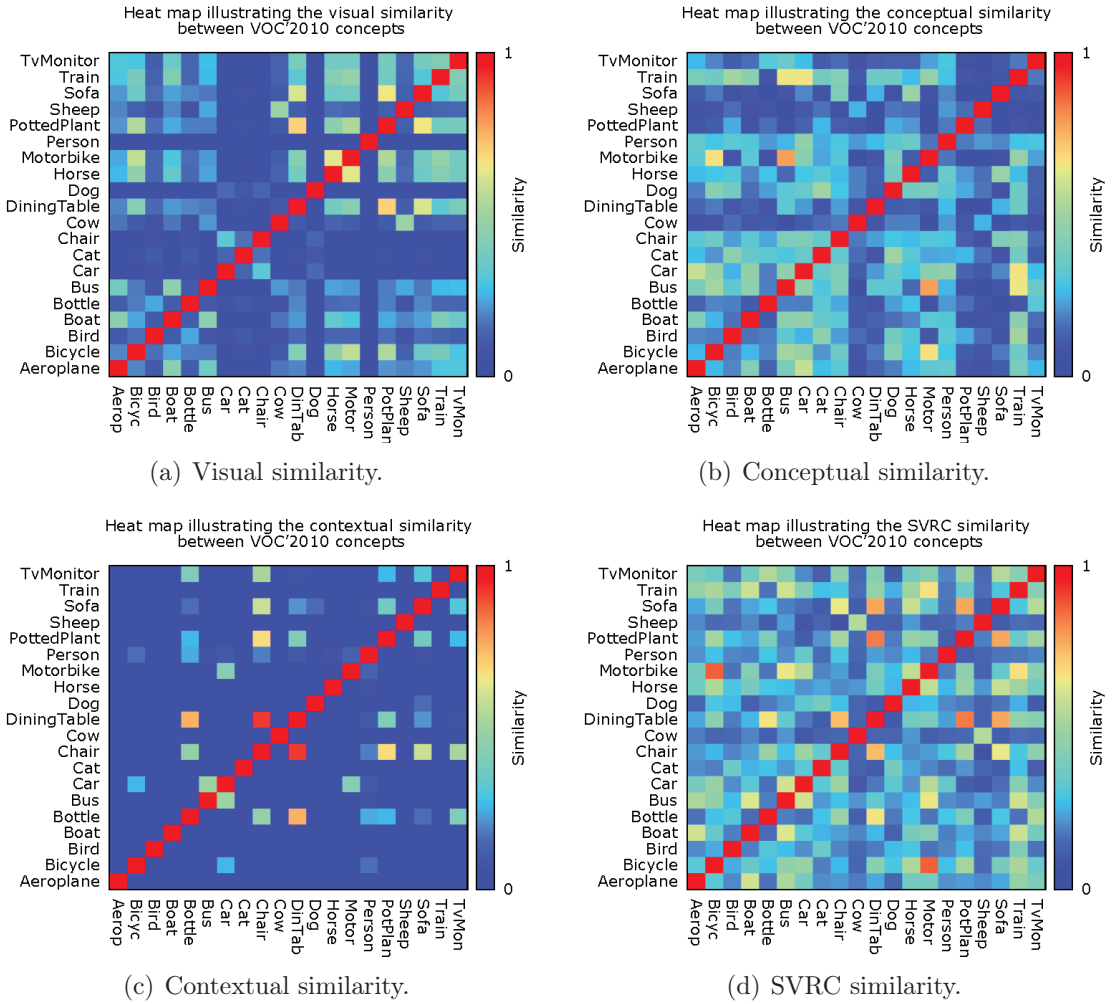
Figure.3.11 illustrates the built hierarchy on Pascal VOC'2010 dataset, using a similarity measure based on the shortest path in WordNet between the considered concepts. Figure.3.10 illustrates the built hierarchy on the same dataset by using only the visual similarity and our heuristic *TRUST-ME*. Compared to the produced semantic hierarchy in Figure 3.8, these hierarchies seem to be less consistent and less efficient. Unfortunately, to our knowledge there is no method currently available that would allow to compare these hierarchies in terms of structures nor in terms of informative modelling. However, we can easily note some inconsistency with the hierarchies built using only the conceptual similarity (respectively the visual similarity), as for example:

- in Figure 3.11, "Bicycle" was considered as a hyponym of "Container" instead of "Vehicle" (both are hypernyms of "Bicycle" in WordNet). Indeed, according to the shortest path in Wordnet, the distance between "Bicycle" and "Bottle" is less than the distance between "Bicycle" and "Motorbike" (respectively less than the distance to all the other concepts of Pascal VOC'2010). Consequently, "Bicycle" and "Bottle" have been considered as very close and have shared "container" as hypernym, which is senseless with respect to the context of Pascal VOC dataset. An explanation to this result is expressed in Definition 10.

- in Figure 3.10, the concepts "Car" and "Train" are considered as closer to "Tv-Monitor" and "Chair" than to "Motorbike", "Boat" and "Aeroplane", which is not coherent. Also, "Car" and "Train" are considered as hyponyms of "Instrumentality" instead of "Conveyance", which could make sense in some context, but in the context of VOC'2010 dataset we believe that "Conveyance" sounds better.

Definition 10 (Conceptual Semantics)**.** *The precise semantic meaning of a concept can only be understood in a specific-context*[6]*, and therefore, it is difficult to gain high accuracy without modeling this specific-context. Consequently, conceptual similarity is context-sensitive.*

[6]A specific-context is the precise domain application in which data may be interpreted. The interpretation may be subjective to this domain.



(a) Visual similarity.

(b) Conceptual similarity.

(c) Contextual similarity.

(d) SVRC similarity.

Figure 3.12: Heat maps illustrating the semantic affinity matrix of the visual similarity (a), conceptual similarity (b), contextual similarity (c) and SVRC similarity (d), between VOC'2010 concepts.

In Figure 3.12 we illustrate the semantic affinity matrix of the VOC'2010 concepts according to the different image modalities, i.e. visual, conceptual, contextual and semantic modality. For each modality, the semantic affinity matrix (also called distance matrix) between concepts is represented as a heat map [7]. This choice is motivated by the easiness provided by heat maps to compare the distribution of individuals. We can see in this figure that the distribution of correlation between each pair of concepts is widely different according to the image modality, and respectively the used measure. This makes sense if we consider the problems related to each modality. Remember, the visual similarity reflects only the perceptual similarity between con-

[7] A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heat maps allow to visualize the proximity of the individuals based on a range of preset color.

cepts, the contextual similarity is a "corpus-dependent" measure (i.e. depends on the concept distribution in the corpus), and the conceptual similarity is context sensitive (cf. Definition 10). This observation has motivated our choice for integrating the different image modalities in order to decrease the limitation of each of them.

According to the obtained results in Figure 3.12, we make the statement that the visual similarity, the conceptual similarity and the contextual one are not always correlated and have different distributions with respect to the targeted semantic or visual space, partly due to the aforementioned problems. Therefore, we do not agree with the statements of [Deselaers and Ferrari, 2011] who reported that: i) the visual variability within a category grows with its semantic domain and ii) visual similarity grows with semantic similarity. At least, we believe that these statements cannot be generalized to any concept.

Finally, we can see in Figure 3.12 that, with respect to each similarity measure, there are some relevant results and many non-relevant ones with respect to image semantics (image context). In Table 3.2, we illustrated some examples of the obtained results by the visual, conceptual, contextual and SVRC measures, and where we have

	Closest(Concept)			
Concept	Visual	Conceptual	Contextual	*SVRC*
Aeroplane	Boat	Car	-	Boat
Bicycle	Motorbike	Car	Motorbike	Motorbike
Bird	Bottle	Train	-	Train
Boat	Bus	Car	-	Bus
Bottle	Bird	Cat	Dining_Table	Dining_Table
Bus	Boat	Motorbike	Car	Motorbike
Car	Chair	Train	Bus	Bus
Cat	Chair	Dog	Sofa	Dog
Chair	Car	Train	Dining_Table	Dining_Table
Cow	Sheep	Sheep	-	Sheep
Dining_table	Pot_Plant	Train	Chair	Pot_Plant
Dog	Car	Cat	-	Cat
Horse	Motorbike	Sofa	Person	Sofa
Motorbike	Bicycle	Bus	Car	Bicycle
Person	-	Car	Bottel	Car
Pot_plant	Dining_Table	Cat	Chair	Dining_Table
Sheep	Cow	Cow	-	Cow
Sofa	Dining_Table	Horse	Chair	Dining_Table
Train	Motorbike	Car	-	Motorbike
Tv_monitor	Train	Bus	Chair	Sofa

Table 3.2: Top correlated concepts according to image modalities.

colored in red irrelevant results and in green sufficiently relevant results[8]. We can see for instance, that according to the conceptual similarity *Closest(Horse)=Sofa*, or also according to the visual similarity *Closest(Dog)=Car*. For cons, as illustrated in Figure 3.12(d), the combination of the different measures, i.e. visual, conceptual and contextual, allows usually recovering from these inconsistencies. Our heuristic *TRUST-ME* is supposed to make the final adjustment in order to achieve a more reliable decision on the relatedness of the concepts, and to link together in a semantic hierarchy the concepts that are semantically related.

3.6 Conclusion

In this chapter we proposed a new approach to automatically build a suitable semantic hierarchy for image annotation/classification. Our approach is based on a new measure of image-semantic relatedness, named "Semantico-Visual Relatedness of Concepts" (*SVRC*), which takes into account the visual similarity and also the conceptual and the contextual ones. As illustrated in our experiments, *SVRC* provides a measure that is more faithful to the image semantics, and consequently allows to dispose of more meaningful and consistent semantic hierarchies with respect to image context. A new heuristic, named *TRUST-ME*, is also proposed for reasoning about concepts relatedness, and to link together in a semantic hierarchy the concepts that are semantically related. Our experiments showed that the built semantic hierarchy using our semantico-visual relatedness measure (SVRC) and the proposed heuristic (TRUST-ME) is more coherent compared with the hierarchies produced using only visual information, and only conceptual information. An experimental evaluation of the produced hierarchies within a framework of hierarchical image classification is proposed in Chapter 5.

Despite the improvement achieved using semantic hierarchies as a hierarchical framework for image classification, there is still a gap between the low-level visual features and the high-level concepts, i.e. **the semantic gap**. Therefore, there is a need to make use of the explicit knowledge of these hierarchies by supplying a formal framework to reason about the coherence of extracted information from images. In Chapter 4 we propose a new approach to build multimedia ontologies that serve as a framework for reasoning about annotation consistency, and thus allow an effective use of explicit knowledge extracted from images.

[8]In these examples, we consider a result as sufficiently relevant if it remains coherent with respect to the all these criteria: i) general context of appearing in images, ii) the visual similarity, and iii) the conceptual consistency.

Chapter 4

Building Fuzzy Multimedia Ontologies for Image Annotation

Contents

4.1 Introduction

In Chapter 3, we have proposed a new method for building semantic hierarchies suitable for image annotation. As aforementioned, semantic hierarchies enable only the description of the subsumption relationships between concepts, which is actually insufficient for modeling the rich semantics of multimedia content and reasoning about it. Indeed, more sophisticated knowledge structures, i.e. *richer in terms of semantic relationships*, are required in order to deal with big data currently available everywhere, and in order to provide efficient systems for storing, indexing and retrieving in this amount of data while satisfying semantic user expectations.

This chapter proposes an approach to build an ontology of spatial and contextual information suitable for reasoning about the coherence of image annotation. Our approach uses firstly the visual and conceptual information in order to build a semantic hierarchy that will serve as a backbone of our multimedia ontology. Contextual and spatial information about image concepts are then computed and incorporated in the ontology to model other semantic relationships between concepts. We also have chosen to use fuzzy description logics as a formalism to represent our ontology in order to take into account the uncertainty and the imprecision of these kinds of information, but also in order to enable formal reasoning as it will be introduced in Chapter 6.

The rest of this chapter is structured as follows. In Section 4.2, we highlight our motivations for proposing a new approach for building multimedia ontologies dedicated to image annotation. Section 4.3 presents an overview of the proposed approach. Section 4.4 introduces the proposed formalism for our multimedia ontology, and the set of axioms and inferences rules allowing to perform the reasoning tasks. Section 4.5 introduces the design of our multimedia ontology as a knowledge base, i.e. the construction of the *TBox*, *RBox*, and *ABox*. In Section 4.6, we illustrates the obtained multimedia ontology on the Pascal VOC dataset. A discussion about the proposed approach and the usefulness of our multimedia ontology for computer vision tasks is presented in Section 4.7. The chapter is concluded in Section 4.8.

4.2 Context and Motivations

4.2.1 Context and General Problems

As introduced in Chapter 2, most approaches for image annotation rely on machine learning techniques to provide a mapping function that allows classifying images in semantic classes using their visual features [Barnard et al., 2003, Lavrenko et al., 2003, Carneiro et al., 2007]. However, these approaches face the scalability problem when dealing with broad content image databases [Liu et al., 2007], i.e. their performances decrease significantly when the concept number is high and depend on the targeted datasets as well [Hauptmann et al., 2007]. 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 a multimedia retrieval system. Consequently, current techniques are struggling to scale up. Therefore, the only use of machine learning seems to be insufficient to solve the problem of image annotation. Firstly, because of the lack of a computer (or statistical) 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 or ontologies to improve image annotation [Dasiopoulou et al., 2009, Straccia, 2010, Wu et al., 2012, Dong et al., 2012].

Indeed, ontologies defined as a formal, explicit specification of a shared conceptualization [Gruber, 1995] 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 interpretation tasks [Hudelot, 2005]. For instance, in [Dasiopoulou et al., 2008], 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 [Hudelot et al., 2005, Kompatsiaris and Hobson, 2008, Hudelot et al., 2008]. In most of these approaches, only the descriptive part of ontologies is used as a common multi-level language to describe image content [Simou et al., 2008], or more recently as semantic concept networks to refine image annotation [Wei and Ngo, 2007, Fan et al., 2008a, Wu et al., 2012], or to perform image classification [Marszalek and Schmid, 2007, Torralba et al., 2008, Dong et al., 2012].

However, much remains to be done in order to achieve more expressive ontologies of image semantics. In the following we describe the noted problems.

GP1 Firstly, most of existing approaches for building multimedia ontologies start from an existing specification (defined by an expert, or inferred from a generic commonsense ontology, etc.) of a domain in order to design their ontologies. However, any given specification, as well-defined it may be, remains 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. These approaches remain useful, but it is more interesting to try to solve the problem in a more generic way by building automatically knowledge models using data mining techniques.

GP2 Secondly, most of approaches for building multimedia ontologies are based either on a conceptual specification [Jaimes and Smith, 2003, Hoogs et al., 2003, Snoek et al., 2007], or on a visual one [Mezaris et al., 2003, Maillot et al., 2004, Yao et al., 2010]. However, as previously pointed out in Chapter 3, conceptual semantics/specification is not always consistent with the perceptual semantics of images (cf. Figure 3.12), and is then insufficient to build ontologies dedicated to image annotation. In the other hand, the perceptual (visual) semantics does not in itself lead to significant semantic structures. Therefore, these approaches do not allow to model image semantics accurately.

GP3 Finally, 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.

4.2.2 Motivations for Building Ontologies as a Knowledge Base

According to [Spaccapietra et al., 2004], ontologies can be classified as taxonomic ontologies, descriptive ontologies, or as knowledge bases, according to the way they are designed. Early efforts on building ontologies have focused on *taxonomic ontologies*, i.e. ontologies acting as thesauri or sophisticated dictionaries. These approaches have focused on defining the used terms in a given domain, and the organization of these terms into generalization/specialization hierarchies, enriched by semantic links (relationships) commonly used in linguistics (e.g., synonymy, antonymy, meronymy, etc.). WordNet [Fellbaum, 1998] and LSCOM [Naphade et al., 2006] are two examples of taxonomic ontologies.

Some other approaches addressing the ontology building problem, have focused on supporting richer ontology models that enable sharing more complex information. They tend to enrich the description of the concepts semantics by associating to each concept a structured description of its properties and its relationships with the others. These ontologies, called *descriptive ontologies*, allow modeling some given domain and the appropriate terms to talk about its concepts. DOLCE[1] is an example of such ontologies [Gangemi et al., 2002].

The final category of approaches have proposed to build ontologies as a knowledge bases. These ontologies consist of conceptual knowledge, i.e. description of concepts

[1] http://www.loa.istc.cnr.it/DOLCE.html

and their relations, and knowledge about the instances. Indeed, in this dissertation we refer to an ontology with associated instances as a *knowledge base*.

With respect to multimedia domain, a core problem with most of existing approaches for ontologies building is that they are limited to propose taxonomic or descriptive ontologies, i.e. they have only focused on describing the concepts (of a given domain) and their relationships. Even the supplied rules[2] and axioms[3] by these approaches are only used to provide a better description of the concepts and their roles. However, an efficient ontology would be rather built as a knowledge base, i.e. by modeling conceptual knowledge about the concepts and their relationships, while including instances and inference rules allowing reasoning on them in an effective manner.

Such ontologies allow *terminological* reasoning (on concepts and inter-concepts relationships) and *assertional* reasoning (on instances). Within this meaning, we are interested in this chapter on building multimedia ontologies as a knowledge base. Specifically, we propose an approach that aims at modeling image semantics through the representation of the semantic relationships between concepts, while including valuable knowledge about the image context under the shape of *instances* of these concepts and relationships. These instances are automatically recovered from image databases, which allows achieving a more representative knowledge base of the image semantics (and respectively the image context).

4.2.3 Motivations for Building Fuzzy Formal Ontologies

This work aims at building (automatically) a multimedia ontology which models knowledge about image semantics, and which will subsequently allow performing reasoning tasks in order to achieve a semantically consistent image annotation. With respect to the knowledge engineering domain, the reasoning tasks require to dispose firstly of a formal ontology.

In this regard, Description Logics (DLs) appeared to us as an excellent candidate to support ontology representation and ontological reasoning. Indeed, as introduced in Section 2.5.2.4, description logics are a family of formal languages for knowledge representation (respectively ontologies representation) and reasoning about them. The goal of reasoning, within this context, is to infer implicitly represented knowledge from the knowledge that is explicitly contained in the knowledge base.

However, despite the rich expressiveness of DLs, they lack the ability to deal with vague and uncertain information which is very common in multimedia content/annotation [Simou et al., 2008]. Indeed, as often stressed in this dissertation, multimedia applications, and specifically image annotation, face a major problem of uncertainty introduced by machine learning algorithms, but also the semantic gap problem, the subjectivity of human perception, etc.

[2]Rules: statements in the form of an *if-then* sentence that describe the logical inferences that can be drawn from an assertion in a particular form.

[3]Axioms: assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application.

To cope with the problem of uncertainty of image annotation, we have chosen to use *Fuzzy-DLs* as a formalism for representing our multimedia ontology. Indeed, in DLs a statement s is either true or false ($s \in \{0, 1\}$), whereas in fuzzy DL, $s \in [0, 1]$ is the degree of truth of the statement. Consequently, the value of s is used to model the uncertainty of an assertion. Thereby, we propose in this chapter a new approach for building a *fuzzy formal* ontology dedicated to image annotation. The proposed ontology is built using a highly expressive formalism, i.e. $f\text{-}\mathcal{SROIQ}(D)$, which provides many axioms for building the ontology and reasoning about its instances. The formalism of our multimedia ontology is introduced in Section 4.4.2.

However, it is important to note that DL strength is on the theoretical level. From a practical standpoint, DL ontologies face significant problems and their effectiveness is still to be proved. For instance:

DLP1 Ontology modeling in DL is not at all 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 [Spaccapietra et al., 2004].

DLP2 This makes the design by humans of a well-defined ontology a big challenge, with no guarantee of success (scalability problem of ontology building).

DLP3 Querying functionality are limited in Description logics, i.e. a more expressive ontology query language is needed.

DLP4 Scalability of DL reasoners, when dealing with very large knowledge bases, is not demonstrated .

4.2.4 Discussion

In this dissertation, we propose to go deeper in the use of ontologies in order to improve image annotation. Our objective is twofold. We first propose an approach to automatically build a fuzzy multimedia ontology suitable for image annotation. Our multimedia ontology includes visual, conceptual, contextual and spatial knowledge. Indeed, spatial and contextual knowledge are two valuable sources of information for image annotation and interpretation [Hollink et al., 2004, Hudelot et al., 2008, Wu et al., 2008]. Secondly, we propose in Chapter 6 a generic approach for image annotation, combining both machine learning techniques, such as hierarchical image classification, and fuzzy ontological reasoning. The proposed approach uses explicit knowledge stored in our multimedia ontology, and ontological reasoning on the produced image annotations by machine learning techniques in order to lift the ambiguity of these annotations and to achieve a better accuracy on the description of images.

In this chapter, we propose a new approach for building multimedia ontologies, while trying to handle some of the aforementioned issues. Specifically, our main contributions can be summarised as follows:

1. Our approach allows for building multimedia ontologies as a knowledge base, and consequently allows for *terminological* reasoning and *assertional* reasoning. Therefore, we answer the problem stated in Section 4.2.2.

2. In our approach, we use data mining techniques in order to automatically discover from image databases the adequate specification (or ontology) for representing image content. This specification is assumed to be faithful to data semantics. Thus, we also show that ontologies are not necessarily limited to be defined by humans. Therefore, we reduce the following problems: **DLP1**, **DLP2**, **GP1** (introduced above).

3. Our approach uses different modalities of the image semantics in order to build an ontology that best represents the data semantics. Thus, we solve the problem **GP2**.

4. Our framework allows both, multimedia ontologies building and ontological reasoning in order to produce a semantically consistent image annotation. The obtained results have shown a significant improvement in the accuracy of image annotation. Thus, our approach allows to answer to the limitation **GP3** stated before.

5. In recent work, the uncertainty was about the truth-degree of a given assertion with respect to the real world facts, e.g. how much it is true that the earth is round, or how much it is true that the left hemisphere is on the left of the right hemisphere (depending on the actual angle between them). In our approach, every assertion is subject to uncertainty with the respect to *image semantics*, e.g. how much it is true that a mouse appears on the right of a keyboard (given an image database and according to the appearance context).

4.3 Overview of our Approach for Building Multimedia Ontologies

We propose in the rest of this chapter, a new approach for building a fuzzy multimedia ontology dedicated to image annotation. Our ontology incorporates several types of knowledge about image semantics. Furthermore, this knowledge is extracted automatically from image databases using data mining techniques, which make it faithful to image semantics. In our approach, three steps are required for knowledge discovery from image databases:

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 the spatial and the contextual knowledge.

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 Chapter 3.

Thereafter, the building of our multimedia ontology is fully automatically performed, i.e. without any human intervention, using the formalism described in Section 4.4.2, and the extracted knowledge from image databases which will be detailed in the following.

Problem Formalization.

Given:

- $\mathcal{DB}$, an 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 $(c_{j_{xmin}}, c_{j_{ymin}}, c_{j_{xmax}}, c_{j_{ymax}})$, where $c_{j_{xmin}}$ and $c_{j_{ymin}}$ are the coordinates of the low left corner of the bounding box (and respectively $c_{j_{xmax}}$ and $c_{j_{ymax}}$ 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 $\mathcal{C} \subseteq \mathscr{C}$. For this work, 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. This ontology should not only incorporate the subsumption relationships between the different concepts, but also richer semantic relations, such as contextual and spatial relationships. Figure 4.1 illustrates the architecture of our approach for building multimedia ontologies.

As introduced in Section 4.2, current knowledge-driven approaches for semantic image annotation have many limitations. Firstly, they were limited to propose either taxonomic ontologies or descriptive ontologies. Secondly, most of them are based on either a conceptual specification or on a visual one, which make them unsuitable for image annotation. Furthermore, most of these approaches have used a handmade built

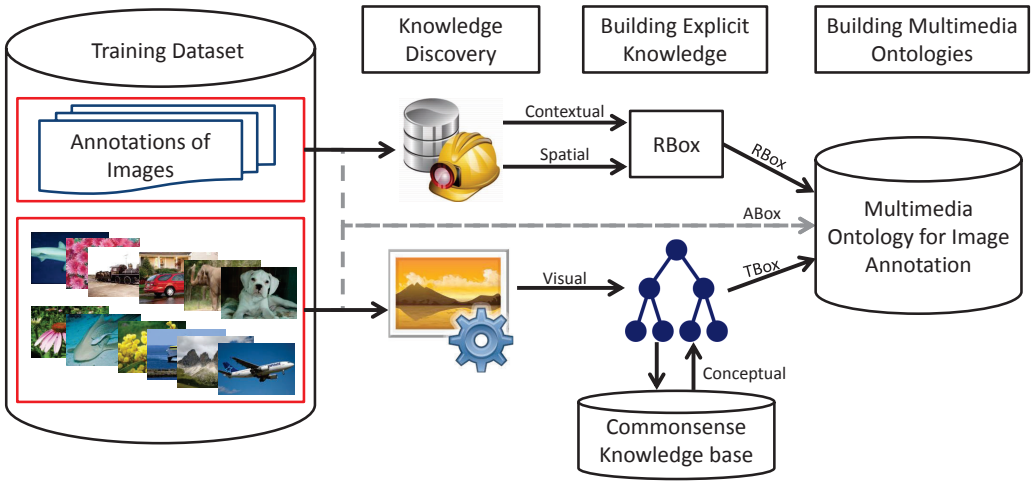


Figure 4.1: From image data to structured knowledge models: Architecture of our approach for building multimedia ontologies dedicated to image annotation.

ontology, which raises the problem of scalability of the building of these knowledge structures.

In order to answer these specific limitations, we propose in this work to build our ontology as a knowledge base that contains explicit and structured knowledge about image context. So 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. Thus, the depicted knowledge within our ontology 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 to perform reasoning tasks on the image annotation and therefore to achieve an effective decision-making toward that goal.

The design of our ontology, and of a well defined formal ontology in a general way, passes exclusively through the following main steps which will be detailed in the remaining of this chapter:

- ⋆ Defining the DL formalism of the proposed ontology, i.e. the expressiveness of the ontology.
- ⋆ Defining the set of axioms and inferences rules allowing to perform the reasoning tasks on the proposed ontology.
- ⋆ Defining the main concepts of the ontology.
- ⋆ Defining the *RBox*, i.e. definition of the key roles (relationships between concepts) and their properties.

- ⋆ Defining the *TBox*, i.e. definition of the subsumption hierarchy, and consequently the subsumption relationships between the ontology concepts.
- ⋆ Defining the *ABox*, i.e. the instances of concepts and the relations between them with respect to the roles defined in the *RBox*.

4.4 Formalism of our Multimedia Ontology

4.4.1 Preliminaries

The Web Ontology Language (OWL) is the current standard language for creating 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/tradeoff between expressiveness and reasoning complexity [Bobillo and Straccia, 2009, Bobillo and Straccia, 2011].

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*[4] [Horridge and Bechhofer, 2011], 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 4.1 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 [Straccia, 2001]. 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 $\alpha \in [0, 1]$.

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 degree of membership $A(x)$ 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

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

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 unique $y \in X$ where $R(x,y)$ is defined.

4.4.2 Expressiveness of our Ontology

As introduced in the previous section, for the concern of providing a highly expressive multimedia ontology with a decidable reasoning, we used the fuzzy DL $f\text{-}\mathcal{SROIQ}(D)$ for designing our ontology. Based on the work of [Straccia, 2006, Stoilos and Stamou, 2007], we introduce in the following the specific formalism (constructors and axioms) used for defining our multimedia ontology.

The $f\text{-}\mathcal{SROIQ}(D)$ 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 $f\text{-}\mathcal{SROIQ}(D)$ 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 $d^D : \Delta_D^n \rightarrow [0,1]$ [Straccia, 2012].

In $f\text{-}\mathcal{SROIQ}(D)$, 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 $f\text{-}\mathcal{SROIQ}(D)$, a fuzzy concept is defined by the following assertions[5]:

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

For more details about the semantics of these assertions cf. Table 4.1 - Constructor C1-C16.

Fuzzy $\mathcal{KB}$. A $f\text{-}\mathcal{SROIQ}(D)$ 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.

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

C	Constructor	Syntax	Semantics
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	$\geq m\ S.C$	$(\geq m\ S.C)^{\mathcal{I}}(a) = sup_{b1,\ldots b_m \in \Delta^{\mathcal{I}}}\left((\otimes_{i=1}^{m}\{S^{\mathcal{I}}(a,b_i) \otimes C^{\mathcal{I}}(b_i)\}) \otimes (\otimes_{j<k}\{b_j \neq b_k\})\right)$
12		$\geq m\ T.d$	$(\geq m\ T.d)^{\mathcal{I}}(a) = sup_{v1,\ldots v_m \in \Delta_D}\left((\otimes_{i=1}^{m}\{T^{\mathcal{I}}(a,v_i) \otimes d_D(v_i)\}) \otimes (\otimes_{j<k}\{v_j \neq v_k\})\right)$
13	At-most restriction	$\leq n\ S.C$	$(\leq n\ S.C)^{\mathcal{I}}(a) = inf_{b1,\ldots b_{n+1} \in \Delta^{\mathcal{I}}}\left((\otimes_{i=1}^{n+1}\{S^{\mathcal{I}}(a,b_i) \otimes C^{\mathcal{I}}(b_i)\}) \rightarrow (\oplus_{j<k}\{b_j = b_k\})\right)$
14		$\leq n\ T.d$	$(\leq n\ T.d)^{\mathcal{I}}(a) = inf_{v1,\ldots v_{n+1} \in \Delta_D}\left((\otimes_{i=1}^{n+1}\{T^{\mathcal{I}}(a,v_i) \otimes d_D(v_i)\}) \rightarrow (\oplus_{j<k}\{v_j = v_k\})\right)$
15	Local reflexivity	$\exists S.Self$	$(\exists S.Self)^{\mathcal{I}}(a) = S^{\mathcal{I}}(a,a)$
16	Fuzzy nominals	$\bigcup_{i=1}^{m}\{(o_i,\alpha_i)\}$	$\{(o_1,\alpha_1),\ldots,(o_m,\alpha_m)\}^{\mathcal{I}}(a) = sup_{i \mid a \in \{o_i^{\mathcal{I}}\}}\alpha_i$
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	Syntax	Semantics
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	$\langle R_1 R_2 \cdots R_n \sqsubseteq R \rhd \alpha \rangle$	$sup_{b_1 \ldots b_{n+1} \in \Delta^{\mathcal{I}}} \bigotimes [R_1^{\mathcal{I}}(b_1,b_2),\ldots,R_n^{\mathcal{I}}(b_n,b_{n+1})] \rightarrow R^{\mathcal{I}}(b_1,b_{n+1}) \rhd \alpha$
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 $S^{\mathcal{I}}(b,a) = 0$

Table 4.1: Syntax and semantics of the Fuzzy Description Logic $f\text{-}\mathcal{SROIQ}(D)$ used for designing our multimedia ontology. $a,b \in \Delta^{\mathcal{I}}$ 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, $\rhd \in \{>, \geq\}$, $\bowtie \in \{>, <, \geq, \leq\}$.

Fuzzy ABox. The fuzzy *ABox* consists of a finite set of fuzzy concept and fuzzy role assertion axioms. Typically, these assertions include: concept assertion ($\langle a : C \bowtie \alpha \rangle$)[6], 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 4.1 - Axioms A1-A5.

Fuzzy TBox. The fuzzy *TBox* is a finite set of General Concept Inclusions (GCIs) constrained with a truth-value and of the form $\langle C \sqsubseteq D \rhd \alpha \rangle$ between two $f\text{-}\mathcal{SROIQ}(D)$ concepts C and D. Concept equivalence $\langle C \equiv D \rangle$ can be captured by two inclusions $C \sqsubseteq D$ and $D \sqsubseteq C$ - (cf. Table 4.1 - Axioms A6-A7).

Fuzzy RBox. The fuzzy *RBox* consists of a finite set of role axioms which are illustrated in Table 4.1 - 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 4.4.3, we have defined the fuzzy operators used in Table 4.1 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 $\alpha \leq \beta, \beta$ otherwise.
5. KD implication (for other constructors): $\alpha \rightarrow \beta = max(1 - \alpha, \beta)$.

The Semantics of $f\text{-}\mathcal{SROIQ}(D)$ is defined in terms of *fuzzy interpretations* [Straccia, 2001].

Fuzzy interpretation. 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 $f\text{-}\mathcal{SROIQ}(D)$ 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: $\mathcal{I} \models \mathcal{KB}$.

[6] $\rhd \in \{>, \geq\}$, $\bowtie \in \{>, <, \geq, \leq\}$

4.4.3 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 [Baader et al., 2003]. 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) [Straccia, 2001].

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

$$glb(\mathcal{KB}, \tau) = \sup\{n \mid \mathcal{KB} \models \langle \tau \geq n \rangle\}, \quad where \ \ \sup \emptyset = 0 \tag{4.1}$$

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:

$$bsd(\mathcal{KB}, C) = \sup_{\mathcal{I} \models \mathcal{KB}} \sup_{x \in \Delta^{\mathcal{I}}} \{C^{\mathcal{I}}(x)\} \tag{4.2}$$

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 4.4.2. 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)**.**

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

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

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

Both explanations and membership degrees are meaningful with respect to a given application. However, the product semantics allows having a more meaningful membership value for our target application than the one produced by Zadeh semantics. Suppose for example, 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. This property is reachable by the use of product semantics.

4.5 Building our Multimedia Ontology

4.5.1 Main Concepts of our Ontology

Proposed concepts. For the building of our multimedia ontology and for achieving the reasoning tasks about image annotations, we propose to define the following main concepts:

- "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.

4.5.2 Definition of the *RBox*

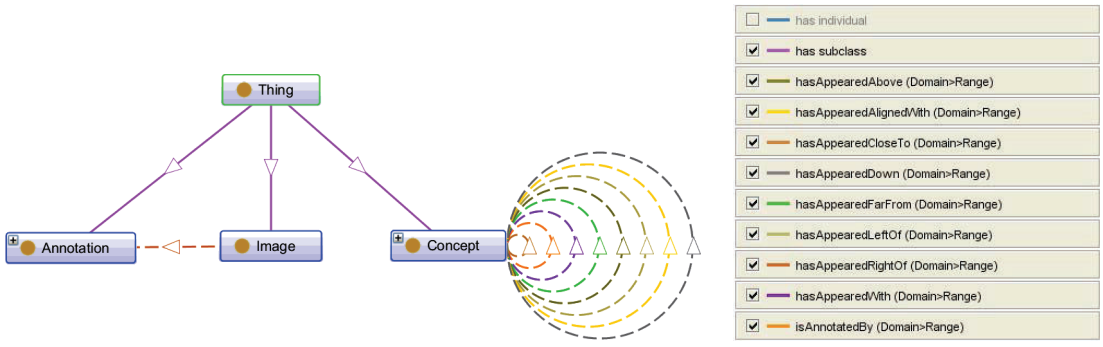
As stated previously, our intent is to design an ontology of spatial and contextual information dedicated to reasoning about the consistency of image annotation. We therefore define in Table 4.2 the proposed roles, and their properties, that constitute the *RBox* of our multimedia ontology. These roles can be categorized as contextual relationships (the light-orange rows) and spatial relationships (the light-blue rows), and are detailed, respectively in Section 4.5.4.1 and in Section 4.5.4.2.

The choice of these specific roles is motivated by the expected (and conceived) reasoning scenarios in order to improve the problem of image annotation - cf. Section 4.7 and Chapter 6. For the clarity of presentation, the semantics of each of the proposed roles in Table 4.2 will be introduced in the section that corresponds to it.

Role name	Domain	Range	Symetric	Reflexive	Functional	Inverse
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	Domain	Range	Symetric	Reflexive	Functional	Inverse
hasFrequency	Concept	Float	-	-	Yes	-
hasAppearedAlone	Concept	Float	-	-	Yes	-

Table 4.2: Roles and functional roles used for defining concept relationships in our ontology ($RBox$). Light-orange rows correspond to contextual relationships and light-blue rows correspond to spatial relationships.



(a) (b)

Figure 4.2: Illustration of the used roles for defining concept relationships in our ontology ($RBox$). Figure 4.2(a) illustrates the main concepts of our multimedia ontology, and the used fuzzy roles (the dashed arrows) for defining the relationships between concepts. Figure 4.2(b) illustrates the roles names.

4.5.3 Building the Semantic Hierarchy and Definition of the *TBox*

The subsumption hierarchy (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 (child) concepts. Consequently, these subsumption relationships allow defining the *Terminological Box* of ontologies.

In order to define the *TBox* of our multimedia ontology, we propose to 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 Chapter 3. Indeed, we propose in this work to automatically build the semantic hierarchy using a *Semantico-Visual* similarity computed between concepts. The proposed *Semantico-Visual* similarity incorporates:

i) a *visual similarity* which represents the visual affinity between concepts, and

ii) a *conceptual similarity* which defines a relatedness measure between target concepts based on their definitions in WordNet.

For more information about how to compute these (visual and conceptual) similarities, the reader is suggested to refer to Section 3.3.2 and 3.3.3.

Afterwards, the building of the subsumption hierarchy is *bottom-up*, and is based on the set of heuristic rules defined in Section 3.4 in order to link together the concepts that are semantically most related, with respect to 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.

Once the subsumption hierarchy is built, and therefore we dispose of the set of concepts ($C \cup C'$) of our multimedia ontology, the subsumption relationships between all these concepts are added to our ontology according the hierarchy structure. This is achieved automatically using the axiom A6 illustrated in Table 4.1.

Example 3 (Definition of the *TBox*)**.**

$$\langle Sofa \sqsubseteq Seat \geq 1 \rangle$$
$$\langle Chair \sqsubseteq Seat \geq 1 \rangle$$
$$\langle Seat \sqsubseteq Furniture \geq 1 \rangle$$
$$\dots$$

4.5.4 Definition of the *ABox*

Subsequent to 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 much representative knowledge base about image semantics. This information, mainly consisting of contextual and spatial relationships between image concepts, will form the *ABox* of our ontology and will serve for reasoning about image annotation. Moreover, our intent is to design a fuzzy multimedia ontology and to represent the uncertainty about the concepts relationships. consequently, we introduce in the following how to compute the truth-degree for each of the proposed roles (relationships).

4.5.4.1 Contextual Relationships

As previously introduced in Section 3.3.4, 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, as aforementioned 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 defined three contextual relations that we estimated important in order to reason about image annotations. These are: $\mathcal{CON} = \{$*"hasFrequency", "hasAppearedWith", "isAnnotatedBy"*$\}$. However, nothing prevents the enrichment of our ontology with other contextual relationships in the future. 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:

$$\widehat{P(c_i)} = \frac{n_i}{\mathcal{L}} \tag{4.3}$$

$$\widehat{P(c_i, c_j)} = \frac{n_{ij}}{\mathcal{L}} \tag{4.4}$$

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 $\mathcal{DB}$:

$$\mu_{\text{hasFrequency}(c_i)} = P(c_i) \tag{4.5}$$

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:

$$\rho(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} = \log \frac{\mathcal{L} * n_{ij}}{n_i * n_j} \tag{4.6}$$

As previously introduced in Section 3.3.4, $\rho(c_i, c_j)$ quantifies the amount of information shared between the two concepts c_i and c_j. In this work, we only want to estimate the positive correlation between two given concepts and therefore we set the negative values of $\rho(c_i, c_j)$ to 0. Moreover, in order to normalize into [0,1] the membership degree of the fuzzy role '*hasAppearedWith*' between two concepts c_i and c_j, we compute it in our approach as:

$$\mu_{\text{hasAppearedWith}(c_i,c_j)} = \frac{\rho(c_i, c_j)}{-\log[\max(P(c_i), P(c_j))]} \tag{4.7}$$

Example 4 (Contextual Relationship: '*hasAppearedWith*')**.**

$$\begin{aligned} \langle a \quad &: \quad Tv\_Monitor \geq 1 \rangle \\ \langle b \quad &: \quad Sofa \geq 1 \rangle \\ \langle (a:b) \quad &: \quad hasAppearedWith \geq 0.26 \rangle \\ &\cdots \end{aligned}$$

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 likeliness that a given annotation '$Annotation_j$' is associated with an image $i_i \in \mathfrak{I}$. To this end, all possible annotations in $\mathcal{DB}$ are extracted and are added to our ontology as subconcepts of concept "Annotation". For instance, we illustrated in Example 5 some inputs of the added assertions to our $ABox$:

The confidence value of this relationship is computed as follows:

$$\mu_{\text{isAnnotatedBy}(Image_1, Annotation_j)} = \frac{n_{Annotation_j}}{\mathcal{L}} \tag{4.8}$$

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 $\mathcal{DB}$.

Example 5 (Contextual Relationship: '*isAnnotatedBy*')**.**

$$
\begin{aligned}
\langle Annotation_1 &\equiv Bus \sqcap Motorbike \sqcap Person\rangle \\
\langle Annotation_1 &\sqsubseteq Annotation \geq 1\rangle \\
\langle Annotation_2 &\equiv Aeroplane \sqcap Car \sqcap Person\rangle \\
\langle Annotation_2 &\sqsubseteq Annotation \geq 1\rangle\rangle \\
\langle Annotation_3 &\equiv Dining\_Table \sqcap Chair \sqcap Bottle \sqcap Dog\rangle \\
\langle Annotation_3 &\sqsubseteq Annotation \geq 1\rangle \\
\langle a &: Image \geq 1\rangle \\
\langle b &: Annotation_1 \geq 1\rangle \\
\langle (a:b) &: isAnnotatedBy \geq 0.004823\rangle \\
&\cdots
\end{aligned}
$$

4.5.4.2 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 [Bloch, 2005]. For instance, using object detectors if one have detected in an image that "Sky" has appeared bellow "Sea", we can easily 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 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. Figure 4.3. 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 Figure 4.4.

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 [Bloch, 1999, Bloch, 2005].

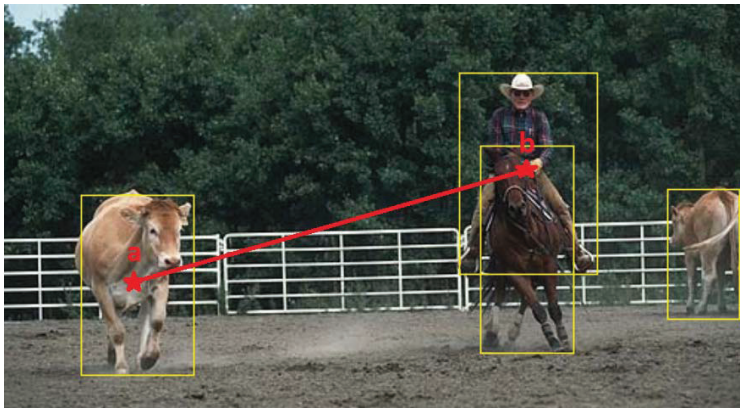
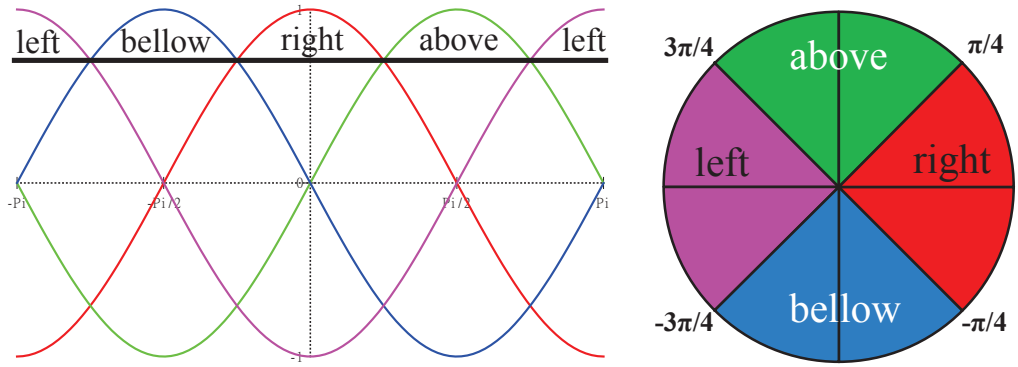


Figure 4.3: Spatial primitives are computed according to the angle between the segment joining two points '*a*' and '*b*' and the *x-axis* of the image. '*a*' and '*b*' are the centroids of two given objects (here "Cow" and "Person") in a given image.



(a) (b)

Figure 4.4: Directional relationships are computed according to an angle α with the *x-axis*.

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

$$\mu_{\mathcal{X}(c_i,c_j)} = \frac{\sharp \text{ of instances where } \mathcal{X}(c_i,c_j)}{n_{ij}} \tag{4.9}$$

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

In our approach, the distance relationships are defined 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 their bounding box $(O_{xmin}, O_{xmax}, O_{ymin}, O_{ymax})$ and $(P_{xmin}, P_{xmax}, P_{ymin}, P_{ymax})$. We define then the

following primitives:

$$distance(O,P) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{4.10}$$

$$size(O) = \sqrt{(O_{xmax} - O_{xmin})^2 + (O_{ymax} - O_{ymin})^2} \tag{4.11}$$

$$close(O,P) = \begin{cases} 1 & \text{if } distance(O,P) < 2(size(O) + size(P)) \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

$$farfrom(O,P) = \begin{cases} 1 & \text{if } distance(O,P) \geq 2(size(O) + size(P)) \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

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

$$\mu_{\chi(c_i,c_j)} = \frac{\sharp \text{ of instances where } \chi(c_i,c_j)}{n_{ij}} \tag{4.14}$$

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

Example 6 (Spatial Relationships)**.**

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

4.6 Experiments

In order to illustrate our approach, we propose in the following to test our proposal on the Pascal VOC dataset. As part of this chapter, we only present the built multimedia ontology on the considered dataset. An empirical evaluation of the effectiveness of the proposed multimedia ontology within an image annotation framework is proposed in Chapter 6.

Figure 4.5 illustrates the built semantic hierarchy on the Pascal VOC'2010 dataset. Indeed, in Section 4.5.3 we proposed to build a semantic hierarchy using a variant (based on the visual and conceptual similarities) of our approach introduced in Chapter 3. 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 our method described in Section 3.4. 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

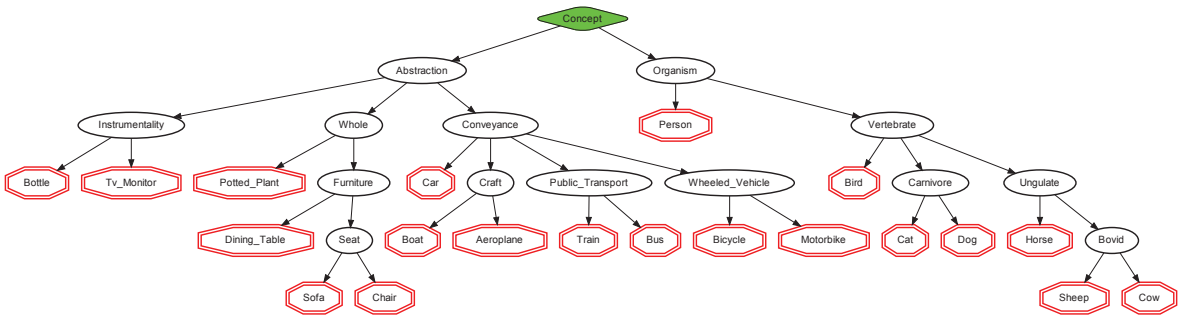


Figure 4.5: The built semantic hierarchy on the Pascal VOC'2010 dataset. Double octagon nodes are original concepts (i.e. concepts in $\mathcal{C}$), and the diamond one is the root of the produced hierarchy.

concur with the assumption that a suitable semantic hierarchy for representing image semantics should incorporate visual and conceptual (semantic) modalities during the building process - cf. Assumption 1.

Figure 4.6 illustrates the built ontology on Pascal VOC dataset and the used roles for defining the concepts 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).

4.7 Discussion and Usage Scenarios

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,

(a)

(b)

Figure 4.6: The built multimedia ontology on Pascal VOC dataset is illustrated in Figure 4.6(a). Dashed arrows represent the fuzzy roles used for defining the contextual and spatial relationships between concepts. Figure 4.6(b) illustrates the roles names.

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 7 (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, as introduced in Chapter 2 - Section 2.4.1, 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 3000 concepts, as for instance the SUN database [Xiao et al., 2010]. Thus, we will dispose of 3000 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 8 (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 image annotation, 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.

4.8 Conclusion

In this chapter, 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.

For illustrating the efficiency of the proposed multimedia ontology in an experimental way, we propose in Chapter 6 a new method for image annotation. Our approach is based on the hierarchical image classification and a multi-stage reasoning framework (based on our multimedia ontology) for reasoning about the consistency of the produced annotation. An empirical evaluation of our approach on the Pascal VOC'2010 dataset is also proposed.

Part II

Application: Using Structured Knowledge Models for Image Annotation

Chapter 5

Hierarchical Image Classification using Semantic Hierarchies

Contents

5.1 Introduction

As introduced in Chapter 3, semantic hierarchies have been introduced recently to improve image annotation. They were mostly used as a framework for hierarchical image classification, and thus to improve the classifiers accuracy and to reduce the complexity of managing large scale data.

This chapter aims primarily to experimentally validate the proposed approach in Chapter 3, which allows building semantic hierarchies dedicated to image annotation. We propose therefore to investigate the contribution of semantic hierarchies, and specifically our proposed approach for building hierarchies, for hierarchical image classification. Thus, we propose first a method based on the hierarchy structure to train efficiently hierarchical classifiers. Our method, named One-Versus-Opposite-Nodes, allows decomposing the problem into several independent tasks and therefore scales well with large databases. We also propose two methods for computing a hierarchical decision function that serves to annotate previously unseen images. The former is performed by a top-down classifiers voting, while the second is based on a bottom-up score fusion. The experiments on Pascal VOC'2010 dataset showed that our methods improve well the accuracy of image annotation.

The rest of this chapter is structured as follows. Section 5.2 presents the motivations of our proposal. In Section 5.3, we present an overview of the proposed approach for hierarchical image classification. Section 5.4 introduces the proposed method for training the hierarchical classifiers based on the semantic hierarchy structure. In Section 5.5, we present the proposed hierarchical decision functions for computing the belonging of a given test image to the classes from the annotation vocabulary. Section 5.6 reports our experimental results on Pascal VOC'2010 dataset. The chapter is concluded in Section 5.7.

5.2 Context and Motivations

In the last decade, many approaches have considered image annotation as an image classification problem and very often as a multi-class classification problem. As introduced by [Marszalek and Schmid, 2008, Cevikalp, 2010], multi-class classification problem is often handled by the combination of binary SVM classifiers. The traditional combination strategies, which do not exploit any hierarchical structure, are

the One-Versus-All (OVA) strategy (through competition) and the One-Versus-One (OVO) strategy (through voting). However, these strategies do not scale well with a large number of classes, since they only offer linear (OVA) or square (OVO) complexity depending on the class number [Marszalek and Schmid, 2008, Bengio et al., 2010].

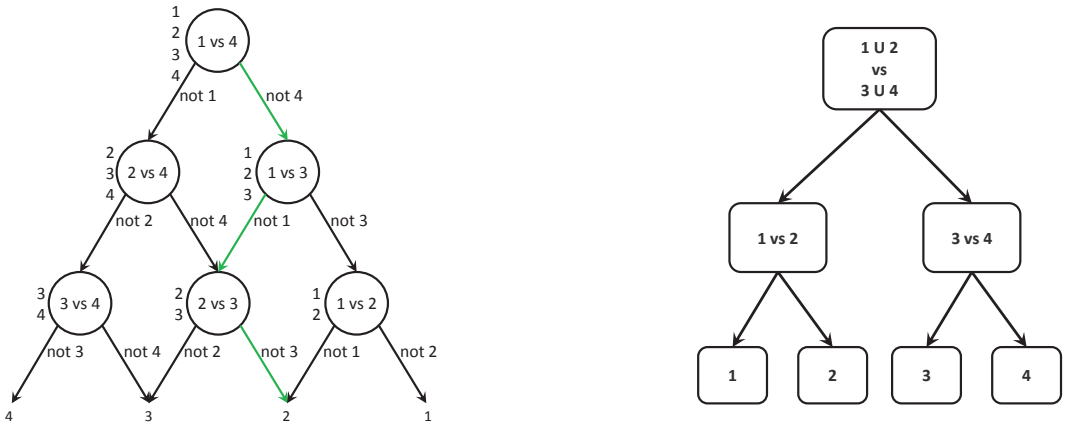
As discussed in [Deng et al., 2010], to cope with a large semantic space, i.e. a large number of concept categories, the classification process can be improved by the use of semantic or visual hierarchies. This assumption has motivated several recent work [Marszalek and Schmid, 2007, Griffin and Perona, 2008, Fan et al., 2007, Fan et al., 2008a, Cevikalp, 2010, Li et al., 2010]. These approaches can be classified into tops-down methods or bottom up methods according to the way the hierarchy is built. In the top-down approach, the class hierarchy is built by the recursive partitioning of the set of classes [Griffin and Perona, 2008, Marszalek and Schmid, 2008, Cevikalp, 2010, Gao and Koller, 2011]. In the bottom-up approach, the class hierarchy is built by the agglomerative clustering of the classes [Marszalek and Schmid, 2007, Fan et al., 2008a, Li et al., 2010, Bengio et al., 2010, Deng et al., 2011b]. These hierarchies are thereafter combined with a set of binary classifiers in order to reduce the complexity of the classification problem. Indeed, this usually results in a logarithmic complexity, which is of interest, especially when dealing with a large number of classes. We refer to these as hierarchical image classification approaches.

Recently, two main directions have been explored for computing a decision function for hierarchical image classification: i) using Decision Directed Acyclic Graphs (*DDAG*) [Platt et al., 2000, Marszalek and Schmid, 2008, Gao and Koller, 2011], and ii) using Binary Hierarchical Decision Trees (*BHDT*) [Marszalek and Schmid, 2007, Griffin and Perona, 2008, Cevikalp, 2010].

Given, $\mathcal{C} = \langle c_1, c_2, \cdots, c_{\mathcal{N}} \rangle$ the annotation vocabulary of the image database, where $\mathcal{N} = |\mathcal{C}|$, the DDAG based approaches train $\mathcal{N}(\mathcal{N}-1)/2$ binary classifiers and use a Directed Acyclic Graph (DAG) to decide about the belonging of an image i to a class $c_j \in \mathcal{C}$. These methods allow at each node in a distance d from the rooted DAG to eliminate d candidate classes from C, resulting in a $\mathcal{N}-1$ decision nodes to be evaluated for the labeling of a test sample - cf. Figure 5.1(a).

On the other side, BHDT based approaches build and use hierarchies as a binary tree, i.e. data are divided hierarchically into two subsets until each subset consists of only one class. Data partition is often achieved using a clustering algorithm. Thus, one SVM is trained for each node of the tree, resulting in a $\log_2 \mathcal{N}$ SVM runs to label a test sample - cf. Figure 5.1(b). BHDT based approaches target to optimize the efficiency of SVM classifiers by reducing the unnecessary comparisons while maintaining a high classification accuracy [Cevikalp, 2010].

The aforementioned approaches (BHDT and DDAG based methods) focus on optimizing the classification accuracy and do not model in anyway the image semantics. Although they allow achieving a higher image classification accuracy compared to traditional approaches, they constrain the used hierarchies to binary structures, resulting in a significant deadlock when the concept number is large. For instance, the method of [Marszalek and Schmid, 2007] is in a deadlock when the concept number exceeds 34, since their method requires $\mathcal{N}-1$ intermediate concepts extracted from



(a) A decision DAG

(b) A four classes BHDT

Figure 5.1: Hierarchical classification using Decision Directed Acyclic Graphs (DDAG) and Binary Hierarchical Decision Trees (BHDT). 5.1(a) illustrates the decision *DAG* for finding the best class out of four classes {1,2,3,4} [Platt et al., 2000]. 5.1(b) illustrates a simple *BHDT* with four classes: {1,2,3,4} [Griffin and Perona, 2008].

WordNet according to the hypernymy relationship ($\mathcal{N}$ being the number of initial concepts), and knowing that WordNet depth is limited to 17 levels.

Moreover, the image semantics does not correspond necessarily to binary items, and many concepts do not have binary opposites[1]. For example, animal or plant species have no binary opposites. Also, data is by nature more complex than binary items. For instance, animals can be classified as 'Herbivores', 'Carnivores' or also 'Omnivores' (both herbivore and carnivore). Consequently, binary structures are not always faithful to data semantics and can lead to inconsistencies in representing knowledge about images or a given domain.

Some other approaches have proposed to use n-ary tree-based models to address the large-scale image classification problem [Zweig and Weinshall, 2007, Bengio et al., 2010, Deng et al., 2011b]. In particular, label embedding trees proposed by [Bengio et al., 2010] have been shown to achieve high performance on the image classification problem (outperforms other tree-based or embedding approaches). Nevertheless, the learning of the tree structure is based on the confusion matrix and involves the training of OVA classifiers for all of the $\mathcal{N}$ classes of the annotation vocabulary. Moreover, their method may lead to an unbalanced tree structure. In order to overcome these problems, [Deng et al., 2011b] have proposed some improvements to the original method of [Bengio et al., 2010]. However, in these approaches, the tree structure over the set of classes is motivated by the reduction of the misclassification loss of the tree. Consequently, the resulted tree structure does not explicitly carries

[1]In lexical semantics, opposites are words that lie in an inherently incompatible binary relationship. The notion of incompatibility here refers to the fact that one word in an opposite pair entails that it is not the other pair member [Wikipedia-Opposite_Semantics, 2012].

any semantic information and thus, does not fall in the main scope of our researches - cf. Section 3.2.2. Hierarchical topic models have also been applied for image classification [Sivic et al., 2008, Li et al., 2010]. But, as stated above, in this chapter we are interested in the SVM based approaches for hierarchical image classification.

Alternative approaches have emerged recently and have proposed the use of semantic relationships between concepts for the building or the use of hierarchies [Fan et al., 2007, Fan et al., 2008a, Torralba et al., 2008, Maillot and Thonnat, 2008, Deng et al., 2009][2]. For instance, [Fan et al., 2008a] proposed to incorporate concept ontology and a multi-task learning algorithm for hierarchical concept learning. A hierarchical boosting algorithm is also proposed to learn their ensemble classifiers hierarchically. The labeling of a new sample is obtained by a voting procedure at all levels of the hierarchy, i.e. $|\mathcal{C} + \mathcal{C}'|$ SVM runs are necessary for labeling new images (where $\mathcal{C}'$ is the intermediate concept nodes in the hierarchy). In [Deng et al., 2009], a 'tree-max classifier' based on ImageNet hierarchy is proposed. The authors proposed to train a classifier at each node of the ImageNet tree. Then, a decision function is computed according to a target class and all its child nodes in order to decide whether an image contains a given object class or not.

To sum up, we illustrate in Figure 5.2 and Table 5.1 the different strategies for hierarchical image classification described above, and we give some information about their complexity according to the number of classes.

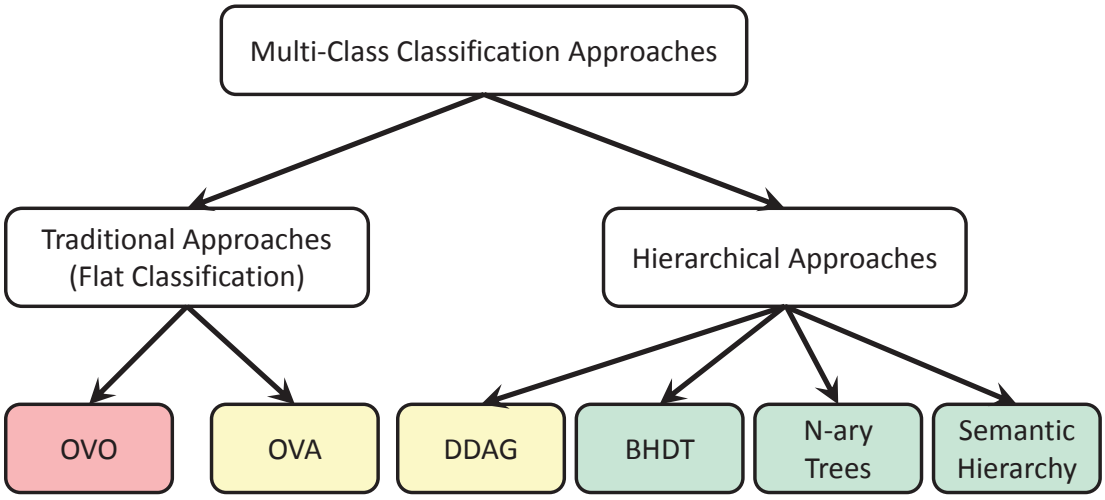


Figure 5.2: Existing approaches for multi-class image classification. The approaches in red boxes present a square complexity, those in yellow boxes present a linear complexity, and those in green boxes present a logarithmic complexity.

Our concern in this chapter is to survey how to draw profits from semantic hierarchies within the context of image classification frameworks. We therefore propose an approach for hierarchical image classification that scales well with the number of classes, and that is less sensitive to unbalanced data.

[2][Maillot and Thonnat, 2008] is a One-Class classification approach.

	Training	Labeling
OVO	$\mathcal{O}(\mathcal{N}^2)$	$\mathcal{O}(\mathcal{N}^2)$
OVA	$\mathcal{O}(\mathcal{N})$	$\mathcal{O}(\mathcal{N})$
DDAG	$\mathcal{O}(\mathcal{N}^2)$	$\mathcal{O}(\mathcal{N})$
BHDT	$\mathcal{O}(\mathcal{N})$	$\mathcal{O}(\log_2 \mathcal{N})$
N-ary Trees	$\mathcal{O}(\mathcal{N})$	$\mathcal{O}(\log_2 \mathcal{N})$
Semantic Hierarchies	$\mathcal{O}(\mathcal{N})$	$\mathcal{O}(\log_2 \mathcal{N})$

Table 5.1: Complexity of existing approaches for image classification.

5.3 Overview of our Approach

In the following, we propose an approach for training hierarchical classifiers in an effective manner. Our approach relies on the structure of the semantic hierarchy in order to train a set of classifiers used for image classification. Subsequently, we propose two methods for computing a decision function in order to achieve image classification. The first one is a *bottom-up* approach, and performs the hierarchical image classification by score fusion. Fusion is achieved by the spreading of scores starting from leaf nodes until reaching the root node. The second is a *top-down* approach and is performed by classifiers voting. Starting from the root node and according to the classifier votes, the hierarchy is traversed until reaching the leaf nodes. The proposed approach for hierarchical classification is independent from the visual representation of images.

For the building of the semantic hierarchy, we rely on the proposed method in Chapter 3. As aforementioned, the building of the hierarchy is based on a *semantico-visual* relatedness measure which incorporates: i) visual information, ii) conceptual information, and iii) contextual information about image concepts. Thus, the built hierarchy is suitable for image classification as it combines the different modalities of images. Figure 5.3 illustrates the semantic hierarchy that we will use for the remaining

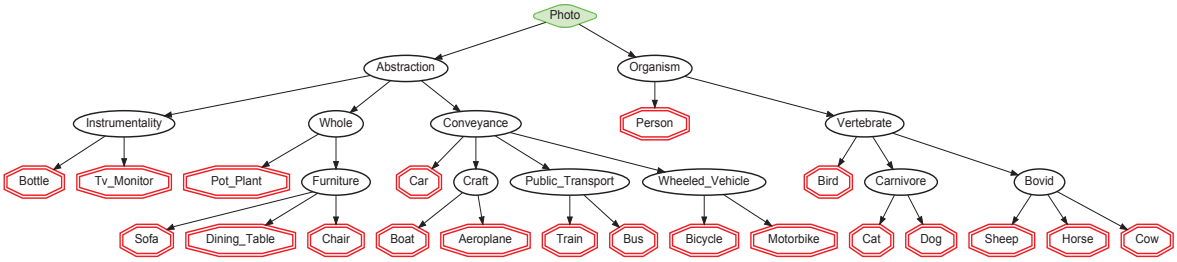


Figure 5.3: The built semantic hierarchy on VOC'2010 dataset by the method proposed in Chapter 3. Double octagon nodes are the original concepts, the diamond one is the root of the produced hierarchy, and other nodes are intermediate concepts discovered by our method for building semantic hierarchies.

of our experiments. This hierarchy is a *N-ary tree* like structure where leaf nodes are the initial concepts. Indeed, one of our motivations for using this method is that the semantics of data (and specifically images) is much more complex than binary items.

5.4 Proposed Approach for Training Hierarchical Classifiers

Based on the hierarchy structure and the subsumption relationships, 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 subsumption relationships, and thus represent the same information with different levels of details. Therefore, the results of these classifiers can be merged in order to achieve relevant decision on the membership of an image to a given class. For instance, according to the semantic hierarchy illustrated in Figure 5.3, a "Cow" $\sqsubseteq$[3] "Bovid" $\sqsubseteq$ "Vertebrate" $\sqsubseteq$ "Organism". Therefore, a good classification system should be able when it recognizes that an image contains a "Cow", to say that it also contains all the concepts that are subsumers of this concept, i.e. "Bovid", "Vertebrate" and "Organism". Reciprocally, if an image does not contains all the subsumers of a given concept, then it should not be annotated by this concept.

Concretely, given a semantic hierarchy, a classifier for each concept node of the hierarchy is trained by performing a One-Versus-Opposite-Nodes (OVON) SVM. Indeed, in order to propose a method that scales well with large image databases, a good strategy would be to decompose the problem into several independent tasks based on the hierarchy structure. Thus, instead of considering all images of a given database for training the classifiers, we only consider the images of a given target concept and its *sibling* concept nodes. This is similar to cut the target concept node from its upper part of the hierarchy and to focus only on its children and sibling concepts. Therefore, for training the classifier of a given target concept, we took as positive samples all images associated with its children leaf nodes. Negative samples are all images of leaf nodes of its sibling concepts as illustrated in Figure 5.4. So, if an image is annotated by "Cow" it will also serves to train the classifiers for "Bovid", "Vertebrate" and "Organism".

The proposed approach is of interest, especially in our case where we only dispose of the images of leaf concept nodes. In particular, this approach allows recognizing other (abstract) concepts that were not initially included in the annotation vocabulary (allows extending the annotation vocabulary of the image database). Furthermore, decomposing the image classification problem into such a complementary and independent sub-tasks will allow a better scalability, since only few classes will be considered for each iteration and with more balanced data (the ratio of positive/negative samples is closer to 0.5).

[3] $\sqsubseteq$: stands for the subsumption ("*is-a*") relationship.

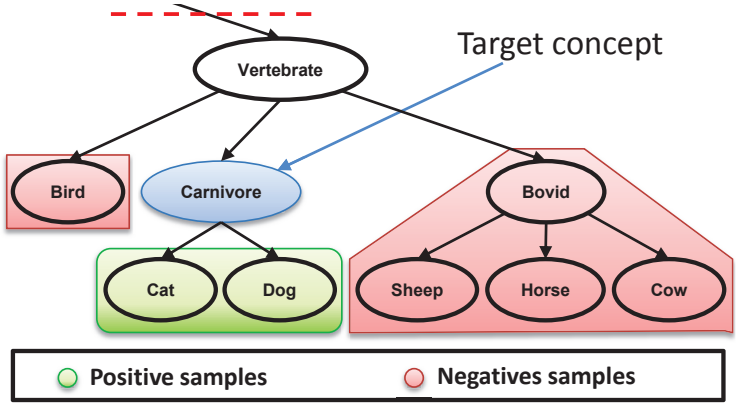


Figure 5.4: Proposed method for training hierarchical classifiers: One-Versus-Opposite-Nodes (OVON).

Problem Formalization.
In a formal way, given:

- $\mathcal{DB}$, an 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 a set of concepts $\{c_j \in \mathcal{C}, j = 1..n_{i_i}\}$ associated with a given image $i_i \in \mathcal{DB}$.
- $\mathcal{SH}$, a semantic hierarchy consisting of a set of $|\mathcal{C}| + |\mathcal{C}'|$ concepts, such that $\mathcal{C} \subseteq \mathcal{C}'$ and $\{\mathcal{C} \cup \mathcal{C}'\} \subseteq \mathcal{SH}$,.

Our objective is to train a set of $(|\mathcal{C} \cup \mathcal{C}'| - 1)$ classifiers, where each classifier is associated with a concept $c_i \in \{\mathcal{C} \cup \mathcal{C}'\}$, and is trained using the following image samples:

- $(\text{Positive samples})_{c_i} = \{i_+ \in \mathfrak{I} \mid \exists\ c_j\ s.t.\ c_j \sqsubseteq c_i,\ \perp \sqsubseteq c_j,\ c_j \in \text{label}(i_+)\}$.
- $(\text{Negative samples})_{c_i} = \{i_- \in \mathfrak{I} \mid \exists\ c_j, c_k\ s.t.\ c_i \sqsubseteq c_k, c_j \sqsubseteq c_k,\ c_i \neq c_j,\ \perp \sqsubseteq c_j\ ,\ c_j \in \text{label}(i_-)\}$.

- where $\text{label}(i_i)$ is a function that returns the set of concepts associated with the image $i_i \in \mathcal{DB}$, $\perp = \emptyset$.

5.5 Proposed Methods for Computing the Decision Function

5.5.1 Bottom-Up Score Fusion (BUSF)

The Bottom-Up Score Fusion (BUSF) method is based on the fusion of classifier scores in order to achieve the final decision about the belonging of an input image to one or more classes from the annotation vocabulary. The classifiers whose scores will be merged must belong to a given path in the semantic hierarchy. The details of this method are as follows.

Starting from the leaf concept nodes and following the subsumption relationships, we compute the average confidence scores of all paths in the hierarchy. The decision function is then computed according to the sign of this average score, i.e. when the sign is positive for a given leaf concept then the image is annotated by it and all its subsuming concepts. A practical standpoint of this method is that the classification results of these SVM are independent. Therefore, in order to reduce the complexity of this method it is possible to run all SVM classifiers together and to compute the membership degree of an image to all classes ($c_j \in \mathcal{C}$) in one time. Subsequently, as illustrated in Figure 5.5, the decision function can be computed easily for all leaf concept-nodes according to the hierarchy structure. Thus, the complexity for labeling a given image is $\leq (2\mathcal{N} - 1)$.

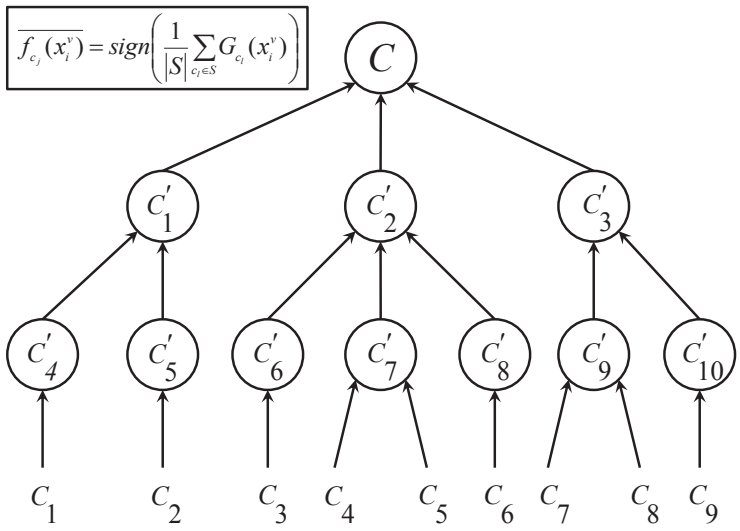


Figure 5.5: Decision function for the Bottom-Up Score Fusion method. For instance, given x_i^v the visual representation of an image i_i, $\overline{f_{c_5}(x_i^v)} = \frac{1}{3} * (\mathcal{G}_{c_5}(x_i^v) + \mathcal{G}_{c'_7}(x_i^v) + \mathcal{G}_{c'_2}(x_i^v))$, where $\mathcal{G}_{c_5}(x_i^v)$ is the decision function of the classifier associated with c_5, and so on. If $\overline{f_{c_5}(x_i^v)} > 0$, then the image i_i is annotated by $c_5,\ c'_7,\ c'_2$.

Concretely, let x_i^v be any visual representation of an image i_i (a visual feature vector), a classifier is trained for each concept class c_j in the hierarchy using our method One-Versus-Opposite-Nodes (OVON). $|\mathcal{C}| + |\mathcal{C}'|$ binary support vector machines are

then trained with a decision function:

$$\mathcal{G}_{c_l}(x_i^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x_i^v) + b \tag{5.1}$$

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

Radial Basis Function (RBF) kernels are used for training our SVM:

$$\mathbf{K}(x_k^v, x_i^v) = exp\Big(\frac{\|x_k^v - x_i^v\|^2}{\sigma^2}\Big) \tag{5.2}$$

Finally, the hierarchical decision function for calculating the membership degree of an image i_i to a concept class c_j is computed as follows:

$$\overline{f_{c_j}(x_i^v)} = sign\Big(\frac{1}{|\mathcal{S}|}\sum_{c_l \in \mathcal{S}} \mathcal{G}_{c_l}(x_i^v)\Big) \tag{5.3}$$

where $\mathcal{S}$ is the set of subsumers of c_j. $\mathcal{G}_{c_l}(x_i^v)$ is the decision function of the classifier associated with the concept c_l.

From a statistical standpoint, the final decision function $\overline{f_{c_j}(x_i^v)}$ is computed by achieving n measures of the same event ($n = |\mathcal{S}|$ is the hierarchy depth). Thus, the uncertainty about $\overline{f_{c_j}(x_i^v)}$ can be calculated as a function of the standard deviation [Vehkalahti, 2008]:

$$\sigma_{\overline{f_{c_j}(x_i^v)}} = \frac{\sigma}{\sqrt{n}} \tag{5.4}$$

where σ is the standard deviation of $\{f_{c_j}(x_i^v), c_j \in \mathcal{S}\}$. Therefore, the final decision function is $\sqrt{n}$ times more accurate than the one obtained from a single classifier. Consequently, one should expect better results using this method compared with flat classification.

5.5.2 Top-Down Classifiers Voting (TDCV)

The Top-Down Classifiers Voting (TDCV) method aims at decomposing the image classification problem into several complementary sub-tasks. It consists in building several classifiers that are able to discriminate one class from the others under a given parent node. Thus, to reach the final decision about the membership of an image to a given class, it is essential to descend the hierarchy according to the classifier decisions (votes). If a classifier in a given level of the hierarchy has responded negatively, then the subtree which it is the root is no longer explored. Otherwise, the hierarchy is explored until reaching one or more leaf concepts.

Algorithm 2 illustrates the *top-down* image classification process using this method. Starting from the root node, the decision functions of subsequent level nodes are evaluated. Concept nodes with a positive confidence value are recursively

explored until reaching leaf concepts. Several paths in the hierarchy can be simultaneously explored, and thus a test image can be associated to many classes as illustrated in Figure 5.6. For a given test image, if none of the leaf nodes of the explored paths is reached, then the image is annotated with the leaf concept having the higher confidence value.

Algorithm 2: Top-Down Classifiers Voting method

Input:
$\mathcal{SH}$: a semantic hierarchy,
i_i: a test image,
x_i^v: a visual representation of the image i_i,
$\{\mathcal{G}_{c_l}(x^v)\}$: the set of decision functions of the classifiers of concepts $c_l \in \{\mathcal{C} \cup \mathcal{C}'\}$.

Result:
$\mathscr{A}_i = \{c_j \mid c_j \in \{\mathcal{C} \cup \mathcal{C}'\}, j \geq 0\}$: predicted annotation for the image i_i.

begin
  $\mathscr{A}_i \leftarrow \emptyset$
  $\Omega \leftarrow \{\ c_j \mid \forall c_j \sqsubseteq \top, \nexists\ c_k\ s.t.\ c_j \sqsubseteq c_k,\ c_k \sqsubseteq \top\}$
  // (i.e., $\Omega \leftarrow$ immediate children concepts of $\top$)
  while $(|\Omega| \geq 1)$ **do**
    $\Upsilon \leftarrow \emptyset$
    foreach $(c_l \in \Omega)$ **do**
      if $(\mathcal{G}_{c_l}(x_i^v) > 0)$ **then**
        $\Upsilon \leftarrow \Upsilon + c_l$
      end
    end
    if $(|\Upsilon| = 0)$ **then**
      $\Upsilon \leftarrow \underset{c_l \in \Omega}{\operatorname{argmax}}\ \mathcal{G}_{c_l}(x_i^v)$
    end
    $\mathscr{A}_i \leftarrow \mathscr{A}_i \cup \Upsilon$
    $\Omega \leftarrow \{\ c_j \mid \forall c_l \in \Upsilon, c_j \sqsubseteq c_l, \nexists\ c_k\ s.t.\ c_j \sqsubseteq c_k,\ c_k \sqsubseteq c_l\}$
    // (i.e., $\Omega \leftarrow$ immediate children of concepts in Υ)
  end
  return $\mathscr{A}_i$
end
$*$ $\top$: stands for the root of the semantic hierarchy ($\mathcal{SH}$).

The TDCV method is efficient in terms of complexity since it requires to train less than $2\mathcal{N} - 1$ classifiers for hierarchical classification, and to evaluate less than $\log_2 \mathcal{N}$ decision nodes for labeling a test image - cf. Table 5.2. However, TDCV method is sensitive to the initial classification since classifiers at the subsequent levels cannot recover from the misclassification of a test image that may occur in a higher concept level. Thus, the classification error can be propagated to the leaf nodes. Nevertheless, as it will be demonstrated in Figure 5.9, the average precision is strongly high for the

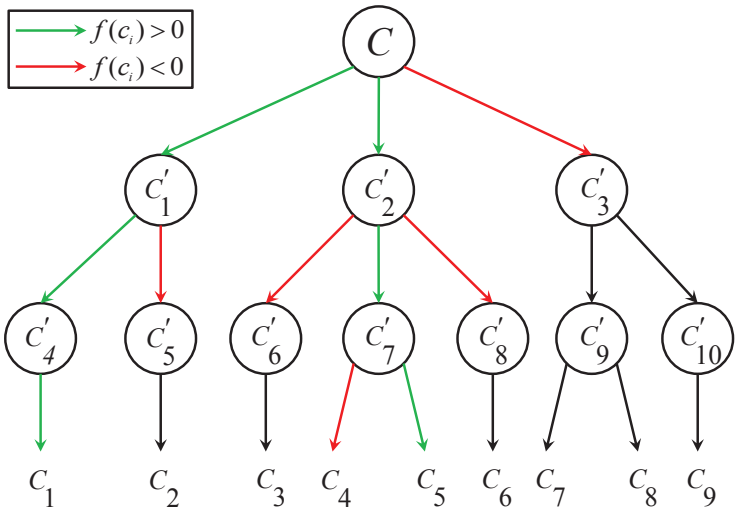


Figure 5.6: Decision function for the Top-Down Classifiers Voting method. Green arrows correspond to explored paths, red arrows correspond to rejected paths according to the classifier decision and black arrows correspond to not explored paths/nodes.

nodes in highest levels of the hierarchy, and therefore errors propagation is small. This average precision is high because the concept classes in higher levels of the hierarchy are sufficiently visually different, and therefore it is easier to find a boundary that separates these classes. Moreover, the more one moves up the hierarchy, classes become more balanced, i.e. more positive and negative examples for these classes.

5.6 Experimental Results

5.6.1 Visual Representations of Images

In order to describe image features, we used in this work the Bag-of-Features (BoF) representation (also known as Bag-of-Visual-Word (BoVW)) which is a widely known method [Li and Perona, 2005]. The BoF model has shown excellent performances and became one of the most widely used techniques for image classification and object recognition. In our approach, image features are described as follows: Lowe's DoG Detector [Lowe, 1999] is used for detecting a set of salient image regions. A signature of these regions is then computed using SIFT descriptor [Lowe, 1999]. Afterwards, given the collection of detected region from the training set of all categories, we generate a codebook of size $K = 1000$ by performing a 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.

5.6.2 Experimental Setup

Our experiments are performed on Pascal VOC'2010 dataset [Everingham et al., 2010]. This dataset contains 4998 training images and 5105 validation images which are used for our experiments. Each image may belong to one or more of the 20 existing classes. Since we do not disposed of the test set, we used the training set for training our concept classifiers and the validation set for evaluating the proposed methods.

The proposed methods for hierarchical image classification are compared to the following methods:

- Hierarchical classification method using the hierarchy illustrated in Figure 5.3 and OVA classifiers.
- Flat classification method described below.
- The H-SVM method proposed by [Marszalek and Schmid, 2007].
- A baseline method.

To perform a fair comparison, we used the same visual representation of images for all of these methods, i.e. Bag-of-Features representation. The flat classification is performed using $|\mathcal{C}|$ support vector machines One-Versus-All, where the inputs are the Bag-of-Features representation of images and the outputs are the desired SVM responses for each image (1 or -1). We used a *k-fold* cross-validation to overcome the unbalanced data problem, taking at each fold as many positive as negative images. Hierarchical image classification with One-Versus-All classifiers is performed by training a set of $(|\mathcal{C}| + |\mathcal{C}'|)$ hierarchical classifiers consistent with the structure of the hierarchy illustrated in Figure 5.3. The baseline method is built by taking the average submission results to VOC'2010 challenge. In the following, the evaluations are performed using the recall/precision curves and Average Precision scores (AP).

5.6.3 Experiments

In Figure 5.7, we compared our method *One-Versus-Opposite-Nodes* (OVON) for training hierarchical classifiers to the *One-Versus-All* (OVA) one. *OVON* performs a better result than the *OVA* classifiers, with an average precision of 63.25% versus 56.42% for the *OVA* hierarchical classification. We can also observe that our method *OVON* performs better on all concept classes, except for the two concepts "Train" and "Bus". As can be seen in Figure 5.3, these two concepts are at the lowest level of the hierarchy and do not share any further sibling concepts. Consequently, both concepts lack of sufficient positive/negative samples for training efficiently their classifiers (90 images for Bus and 113 images for Train). In the other hand, the *OVA* method has benefited from cross-validation (Repeated random sub-sampling validation). But despite this, we can observe that our method is better when considering all concept classes, since it decomposes the problem of classifiers training according to the hierarchy structure. Thus, the classification problem is simplified since we are

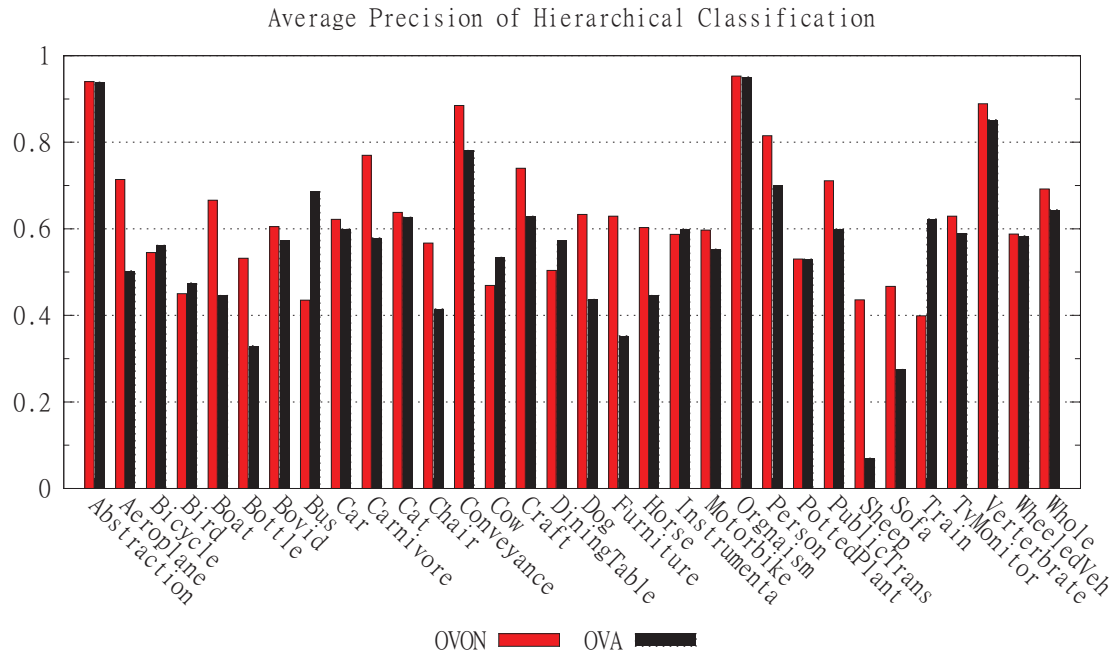


Figure 5.7: Comparison of the *One-Versus-Opposite-Nodes* (OVON) and the *One-Versus-All* (OVA) hierarchical classifiers on VOC'2010 dataset.

looking for a boundary that separates fewer concept classes, which is therefore easier to found.

In Figure 5.8, we compare our methods for hierarchical image classification: Bottom-Up Score Fusion (BUSF) and Top-Down Classifiers Voting (TDCV) to i) a baseline method introduced in the beginning of this section, ii) a flat classification method previously introduced, and iii) the *H-SVM* method of [Marszalek and Schmid, 2007]. Indeed, the authors proposed a method, called Hierarchy of SVM (*H-SVM*), based on WordNet for building a semantic hierarchy which is subsequently used for hierarchical image classification. In order to compare our methods to *H-SVM*, we built a hierarchy of concepts from WordNet using VOC'2010 dataset. The building of the hierarchy, as well as the training of hierarchical classifiers are performed in the same manner described in [Marszalek and Schmid, 2007].

As one can observe in Figure 5.8, our methods achieve a higher average precision than the flat classification method with a gain of +26.8% for the *BUSF* method and a gain of +16.04% for the *TDCV* method. Compared to the baseline method our approaches are slightly better. This can be explained by the efficient image features used in the submission of VOC challenge, and the basic one used in our approach. Moreover, we recall that we used only the half of the training set since we did not dispose of the test set used in the challenge. We also included in our evaluations the images marked as difficult, which are ignored in the challenge because they are considered as difficult to recognize. Thus, the obtained results are still promising and could be improved by incorporating more sophisticated image descriptors, as for instance SIFT + HOG, or SIFT + color/texture descriptors. A comparison of our methods for hierar-

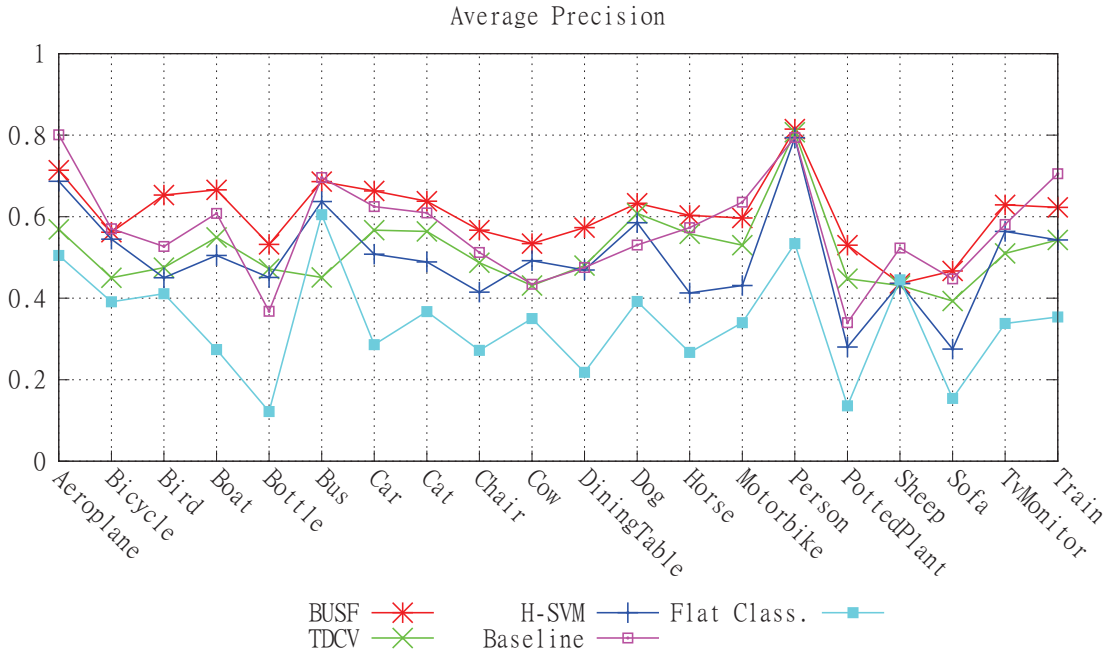
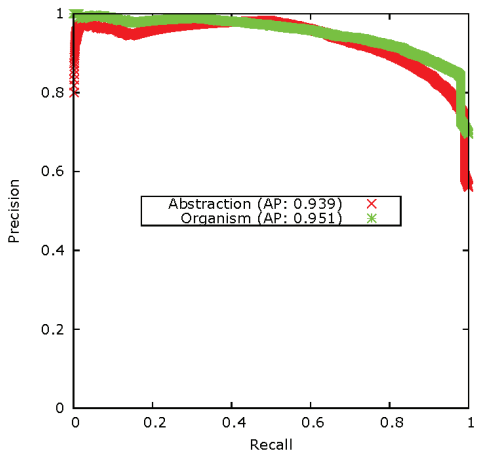


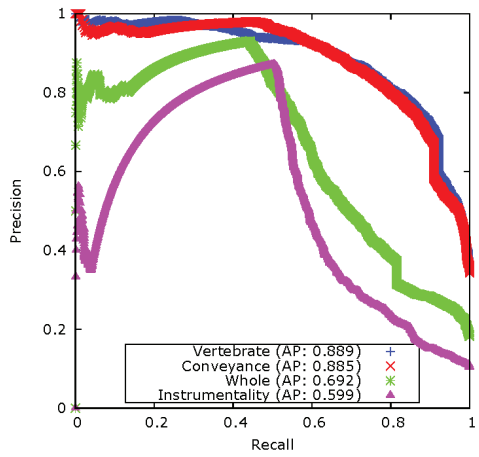
Figure 5.8: Comparison of our methods for hierarchical image classification: Bottom-Up Score fusion (BUSF) and Top-Down Classifiers Voting (TDCV), with the Baseline, H-SVM, and the flat classification method.

chical image classification to the *H-SVM* method of [Marszalek and Schmid, 2007] is also illustrated in Figure 5.8. Our method *BUSF* achieves a higher average precision than the others with a gain of +8.99% compared to the *TDCV* method and a gain of +10.76% compared to the method *H-SVM*. The average precision for these methods was as follows: 60.6% for the *BUSF* method, 51.61% for the *TDCV* method, and 49.84% for the method *H-SVM*. We can therefore conclude that there is a significant improvement in performance with the use of the proposed hierarchy and our methods for hierarchical image classification.

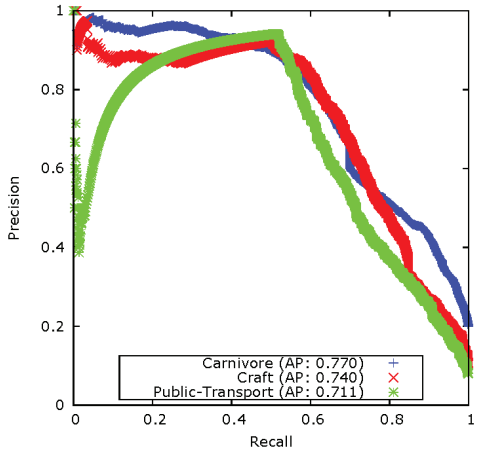
Figure 5.9 illustrates the average precision of the different concepts at the intermediate levels of the hierarchy. As shown in this figure, the classifiers accuracy decreases as we go deeper in the hierarchy. This is a logical result since the concept classes in higher levels of the hierarchy are sufficiently visually different, i.e. it is easier to find a boundary that separates these classes. The training data for these concepts is also more balanced. For instance, the ratio of positive/negative samples in VOC'2010 dataset is about 5%. *OVON* method allows overcoming this problem as it decomposes image classification into several sub-tasks. The ratio of positive/negative samples is 35.6% for OVON, i.e. the classes are more balanced and there is no need for techniques as over-sampling or under-sampling to recover the problem of unbalanced data. In addition, the results reported in Figure 5.9 demonstrate the effectiveness of our method for building semantic hierarchies (presented in Chapter 3) compared to the methods using textual knowledge for building hierarchies. Indeed, Figure 3.1 illustrates a hierarchy inferred from textual knowledge, i.e. built by extracting the

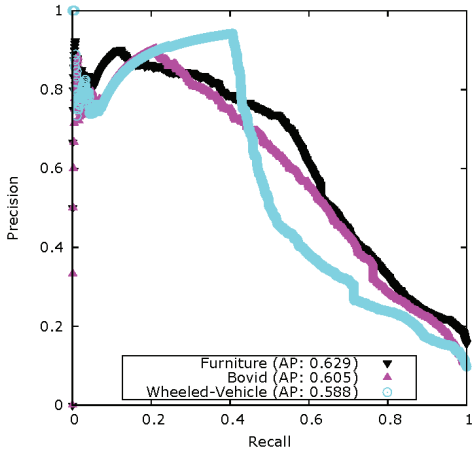


(a) AP for the 2^{nd} layer concepts.



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





(c) Average Precision for the fourth layer concepts.

Figure 5.9: Recall/Precision curves for the concepts of each level of the hierarchy.

relevant graph in WordNet linking all concepts of VOC'2010. The depth of this textual hierarchy is about 17 levels. So, as proved by the results shown in Figure 5.9, one can expect a significant decrease of the hierarchical classification results as we go deeper in the hierarchy. Therefore, the average precision for the leaf nodes (concepts) should be less relevant compared to our proposed method.

Finally, we illustrate in Table 5.2 the complexity of our methods for hierarchical image classification using the semantic hierarchy shown in Figure 5.3. Our methods show a better time-complexity for training classifiers compared to the approaches based on Decision Directed Acyclic Graphs, and the ones based on Binary Hierarchical Decision Trees. For image annotation (labeling new images), our method *BUSF* has showed a higher complexity compared to the others, but in the other hand offers a better accuracy (average precision) compared to these methods. In terms of complexity, our method *TDCV* is more efficient than all others for the labeling of new images, but does not perform as well as the *BUSF* method in terms of accuracy.

	Training	Labeling	VOC'10 Training	VOC'10 Labeling
DDAG	$(\mathcal{N}^2 - \mathcal{N})/2$	$\mathcal{N} - 1$	190 t	19 t'
BHDT	$2\mathcal{N} - 1$	$\log_2 \mathcal{N}$	39 t	5 t'
TDCV	$\leq 2\mathcal{N} - 1$	$\leq \log_2 \mathcal{N}$	32 t	4 t'
BUSF	$\leq 2\mathcal{N} - 1$	$\leq 2\mathcal{N} - 1$	32 t	32 t'

Table 5.2: Complexity (in terms of SVM runs) of our methods compared to the DDAG and BHDT approaches. t, t': stand for 1 unit of time. t: for training one classifier. t': classifier runtime.

5.7 Conclusion

As previously introduced in this chapter, hierarchical image classification was often considered as a binary classification problem, probably because of the lack of methods for automatically building semantic hierarchies for computer vision tasks. In Chapter 3, we have proposed a new approach for building semantic hierarchies dedicated to image annotation. Therefore, we have proposed in this chapter to investigate the contributions of such (*n-ary* like structure) hierarchies for hierarchical image classification.

In this chapter, we proposed an approach for hierarchical image classification using semantic hierarchies. Our approach is based on the structure of the semantic hierarchy to efficiently train hierarchical classifiers, i.e. draws benefits from the hierarchy structure to decompose the training process into several independent and complementary sub-tasks. Thus, it allows to gain in efficiency and in complexity as it was shown by the obtained results. We have also proposed two methods for computing a hierarchical decision function serving to annotate previously unseen images. The former is achieved by a top-down classifiers voting, while the second is based on a bottom-up score fusion. Compared to other approaches, our methods have achieved higher accuracy on Pascal VOC'2010 dataset. Specifically, the bottom-up score fusion has shown significant improvement in the classification accuracy, as it allows to reduce the inconsistency about classifier decisions by the fusion of scores of the classifiers belonging to a same path in the hierarchy.

Chapter 6

Multi-Stage Reasoning Framework for Image Annotation

Contents

6.1 Introduction

In Chapter 4 we have proposed an approach to automatically build a fuzzy multimedia ontology dedicated to image annotation. In our approach, visual and conceptual information were used to build a semantic hierarchy considered as a backbone of our ontology, and used to infer the subsumption relationships between the ontology concepts. Contextual and spatial information were thereafter added to our multimedia ontology in order to achieve a more expressive knowledge base of image semantics. Fuzzy description logics were used as formalism for representing our ontology. The choice of this formalism was motivated by its ability to represent the uncertainty and imprecision of these kinds of information, and also because it enables formal reasoning on concepts.

In this chapter, we propose a knowledge-based multi-stage reasoning framework for image annotation. Our framework uses the proposed multimedia ontology (in Chapter 4) in order to refine the image annotation, and to achieve a more consistent textual description of images. The reasoning tasks are achieved using fuzzy description logics reasoning. Finally, an empirical evaluation of our approach is performed on Pascal VOC'2009 and Pascal VOC'2010 datasets. The obtained results have shown a significant improvement on the average precision results.

The rest of this chapter is structured as follows. In Section 6.2 we introduce the motivations of our approach. Section 6.3 presents a global overview of the proposed framework. In Section 6.4, we introduce the proposed method for reasoning about image annotation in order to achieve a semantically consistent image annotation. Indeed, Subsection 6.4.1 introduces the proposed method for hierarchical image classification, Subsection 6.4.2 introduces the proposed method for reasoning about image annotation using the subsumption relationships, Subsection 6.4.3 introduces the reasoning process using contextual knowledge and Subsection 6.4.4 introduces the reasoning process using spatial information. Section 6.5 presents the experimental results obtained using the proposed framework for image annotation on the test datasets. This chapter is concluded in Section 6.6.

6.2 Context and Motivations

Despite significant progress shown by statistical approaches for images annotation, the semantic gap problem is still an open issue for image annotation and interpretation. In this context, several approaches have proposed recently to improve these tasks by the use of explicit knowledge.

A first category of approaches have proposed to use semantic hierarchies for image annotation and classification [Marszalek and Schmid, 2007, Fan et al., 2008a, Deng et al., 2009]. However, most of these approaches use the semantic hierarchies to reduce the complexity of the classification problem, or as a framework for hierarchical image classification. To our knowledge, none has used 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 have proposed to use multimedia ontologies in order to define a standard for the description of low-level multimedia content [Simou et al., 2005, Dasiopoulou et al., 2010], or to use them as a semantic repository for storing knowledge about image domain [Simou et al., 2008], or to use them as a framework for the semantic interpretation of multimedia content and reasoning over the extracted descriptions [Hollink et al., 2004, Hudelot et al., 2008, Dasiopoulou et al., 2008, Hudelot et al., 2010]. Indeed, ontologies allow modeling many valuable semantic relations between concepts which are missing in the semantic hierarchy models, as for instance the contextual and the spatial relationships. These relationships have been proved to be of prime importance for image annotation [Hollink et al., 2004, Hudelot et al., 2008, Hudelot et al., 2010, Straccia, 2010]. The reasoning power of ontological models has also been used for semantic image interpretation. In [Dasiopoulou et al., 2008, Hudelot et al., 2008, Hudelot et al., 2010], formal models of domain application knowledge are used through fuzzy description logics in order to help and to guide the semantic image analysis. [Dasiopoulou et al., 2008, Dasiopoulou et al., 2009] have proposed an approach for improving image annotation using ontological reasoning on the outputs of statistical concept classifiers. Nevertheless, their reasoning tasks were limited to the inconsistency checking with respect to the *TBox*, i.e. using only the subsumption relationships between concepts.

However, many problems were not raised by these approaches, or sometimes partially addressed. Firstly, it is widely accepted now that it is not possible to define a standard for the description of low-level multimedia content, since this domain is witnessing a fast evolution and it is in a constant improvement process. Secondly, many of these approaches are limited to provide a formalism allowing to use ontologies as a repository for storing knowledge about multimedia content. However, these approaches have not addressed the problem of reasoning about this knowledge, or when it is done, it is not performed on representative datasets of current real life applications. Consequently, the effectiveness of the stored knowledge and the proposed knowledge models have to be proved. The final category of approaches dealing with reasoning over the extracted descriptions can be qualified as "light". Indeed, these approaches have not addressed the problem of building ontologies dedicated to image

annotation/interpretation, and they are usually limited to use hand-built knowledge models. Moreover, they propose reasoning scenarios to deal with the problem of images annotation, but these scenarios are not tested on real data applications, and consequently their effectiveness is not assessed. However, this application domain, i.e. knowledge-driven approaches for image annotation, is still in its infancy and we believe that these approaches will be effectively improved in the near future.

In this dissertation, we proposed a tentative to reduce the above limitations of current approaches, i.e.:

1. Building multimedia ontologies dedicated to image annotation (proposed in Chapter 4). Our approach rely on mining image databases in order to achieve a representative knowledge base of image semantics (and respectively of a given application domain).

2. The use of the reasoning power of ontologies in order to produce a semantically consistent image annotation.

3. To our knowledge, this is the first attempt integrating such a complex reasoning process, i.e using subsumption relationships, and also contextual and spatial relationships.

4. Illustrate and test our approach on a public dataset assumed to be enough representative of real current applications.

To sum up, in Chapters 3 and 4, we have proposed new approaches to build explicit structured knowledge models about image semantics. In this chapter, we propose to use these built knowledge models in a framework for reasoning over the outputs of machine learning algorithms. Our objective is to provide a semantically consistent image annotation while handling the imprecision of machine learning approaches.

6.3 Overview of the Proposed Framework

Image classification is one of the most 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 a 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.

Our approach is motivated by the above assumption. Indeed, we propose in the following a multi-stage reasoning framework based on the multimedia ontology proposed in Chapter 4. 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 Figure 6.1.

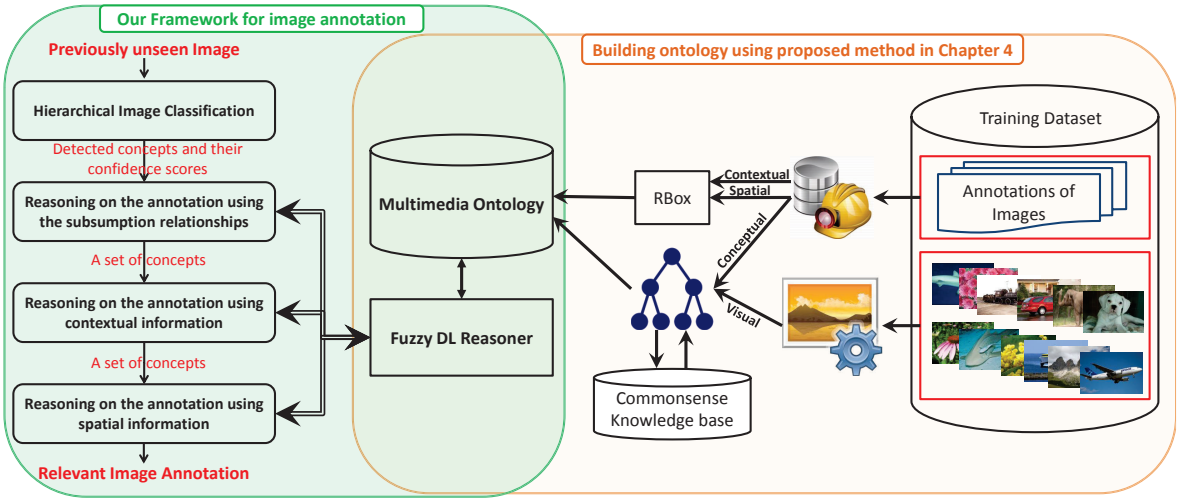


Figure 6.1: A global overview of our approach, consisting of two frameworks. The first framework allows building multimedia ontologies dedicated to image annotation, while the second allows reasoning on the image annotation in order to achieve a semantically relevant annotation.

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 the ontology concepts: i) subsumption relationships, ii) contextual relationships, and iii) spatial relationships (as for instance the proposed multimedia ontology in Chapter 4).

This multimedia ontology is then used within our framework for annotating previously unseen images. As illustrated in Figure 6.1, 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 subsumption relationships and our fuzzy DL reasoner. Inconsistent concepts are removed from the candidate annotation[1] of the input image.

[1] 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.

- 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.

6.4 Proposed Method: Multi-Stage Reasoning Framework for Image Annotation

In the following, we focus on the proposed knowledge-based framework for image annotation. We will detail the different components of our framework, as they are shown in Figure 6.2.

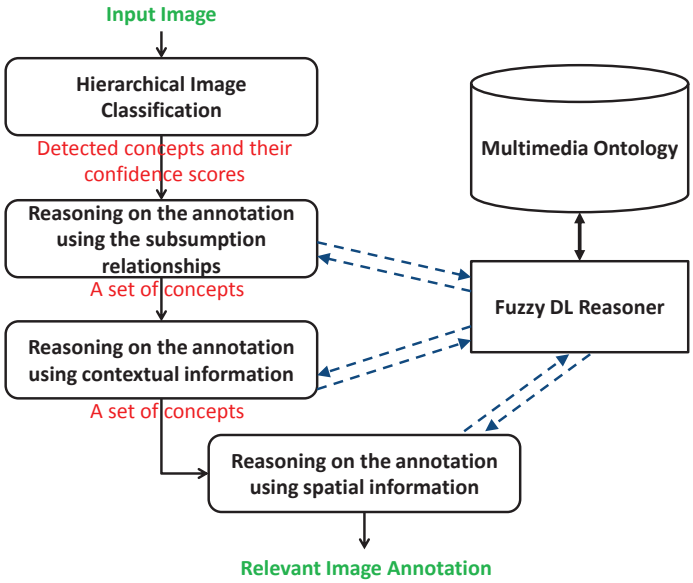


Figure 6.2: Proposed method: a knowledge-based multi-stage reasoning framework for image annotation.

Problem Formalization.

Given:

- $\mathcal{DB}$, an 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 a set of concepts $\{c_j \in \mathcal{C}, j = 1..n_{i_i}\}$ associated with a given image $i_i \in \mathcal{DB}$.

$\Rightarrow$ $\mathcal{DB}$ is only used for training the hierarchical classifiers.

- $\mathfrak{I}'$, a test set consisting of a set of previously unseen images, i.e. $\mathfrak{I}' = \{i'_1, i'_2, \cdots, i'_{\mathcal{L}'}\}$, with $\mathcal{L}'$ is the number of images in the test set.
- $\mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$, a fuzzy multimedia knowledge base consisting of:
  - A set of concepts $\mathcal{C} \cup \mathcal{C}'$, where $\forall c_i \in \mathcal{C}, \exists c_j \in \mathcal{C}' \; s.t \; c_i \sqsubseteq c_j$.
  - A set of roles representing the contextual and spatial relationships between all concept $(c_i, c_j \in \mathcal{C})$ (as introduced in Chapter 4).

Our objective is, given a previously unseen image $i'_i \in \mathfrak{I}'$, to provide a semantically consistent annotation for i'_i.

6.4.1 Hierarchical Image Classification

As stated above, our multi-stage reasoning framework for image annotation relies, in a first step, on hierarchical image classification. This chapter does not focus on the effectiveness of the hierarchical classification method. Thus, for the classification of previously unseen images, we may suppose that a classifier for each concept $(c_i \in \mathcal{C} \cup \mathcal{C}')$ of the semantic hierarchy is already trained by any given hierarchical classification method.

In Chapter 5, we have proposed an approach for performing hierarchical image classification. However, in the context of this chapter, we are interested in the use the subsumption relationships within a framework allowing explicit reasoning tasks on the semantic consistency of the classifier outputs (cf. Section 6.4.2). Therefore, for image classification, we simply used a hierarchical *One-Versus-All* SVM classifiers instead of using the method proposed in Chapter 5.

Consequently, in this work, the semantic hierarchy is only used to recover the set of positive and negative images for training the classifiers in the different layers.

Moreover, the decision function of each classifier is independent from its subsumed (child) and subsuming (parent) concept nodes. Concretely, given a semantic hierarchy, a classifier for each concept node of the hierarchy is trained by performing a One-Versus-All (OVA) Support Vector Machines. Therefore, for training the classifier of a target concept node, we took as positive samples all images associated with its children leaf nodes. The negative samples are all the other images of the database.

So, 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:

$$\mathcal{G}(x_i^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x_i^v) + b \tag{6.1}$$

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 again used for the training of our SVM:

$$\mathbf{K}(x_k^v, x_i^v) = exp\Big(\frac{\|x_k^v - x_i^v\|^2}{\sigma^2}\Big) \tag{6.2}$$

6.4.2 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 Figure 6.3. 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.

To sum up, 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 consequently the set of its hypernym concepts $\{c_k \in \mathcal{C}' \mid c_j : C > 0, c_k : D > 0, C \sqsubseteq D > 0\}$. Therefore, the reasoning process can be formulated using conjunctive queries as follows:

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

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 4.1, 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:

$$bed(\mathcal{KB}, a : ValidCC) = x_1 \otimes x_2 \otimes \cdots \otimes 1 = x_1 * x_2 * \cdots * 1 \tag{6.3}$$

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 image i'_i.

For instance, let us consider the first example in Figure 6.3. The image classification algorithm has detected "Motorbike" as a candidate concept (among others) for the considered image. However, according to the subsumption hierarchy (cf. Figure 4.5) 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, and thus this concept is removed from the list of candidates since $bed(\mathcal{KB}, Motorbike : ValidCC) = 0$.

Example 9 (Consistency checking of concept "Motorbike")**.**

$$\mathcal{KB} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \cup \{\langle a : Motorbike \geq 0.262 \rangle\} \cup \{\langle a : Weeled\_vehicule \geq 0 \rangle\} \cup$$
$$\{\langle a : Conveyance \geq 0 \rangle\} \cup \{\langle a : Abstraction \geq 0.109 \rangle\} \cup \{\langle a : Concept \geq 1 \rangle\}\rangle$$
$$bed(\mathcal{KB}, Motorbike : ValidCC) = 0.262 \otimes 0 \otimes 0 \otimes 0.109 \otimes 1 = 0$$

6.4.3 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 a given 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

		
Groundtruth:		
Sheep, Person.	Cat, Tv_monitor, Sofa.	Chair, Dining_Table: *Marked as Difficult*, Person.
Classifier Outputs for Concepts $\in \mathcal{C}$:		
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	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	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
Classifier Outputs for Concepts $\in \mathcal{C}'$:		
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	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	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
Reasoning on the annotations using the subsumption hierarchy:		
Cow: 0.310, Horse: 0.391, Person: 0.805, Sheep: 0.519 ~~Motorbike~~, ~~Dining_table~~	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~~	Bottle: 0.650, Chair: 0.825, Dining_table: 0.688, Person: 1.00, Potted_Plant: 0.470, Tv_monitor: 0.127 ~~Bicycle~~, ~~Dog~~, ~~Motorbike~~
Reasoning on the annotations using image context:		
Person: 0.805, Sheep: 0.519 ~~Horse~~, ~~Cow~~	Cat: 0.616, Chair: 0.348, Dining_table: 0.318, Sofa: 0.698, Tv_monitor: 0.342 ~~Person~~, ~~Potted_plant~~	Bottle: 0.650, Chair: 0.825, Dining_table: 0.688, Person: 1.00, Potted_Plant: 0.470 ~~Tv_monitor~~
Reasoning on the annotations using spatial information:		
Person: 0.805, Sheep: 0.519	Cat: 0.616, Chair: 0.348, Dining_table: 0.318, Sofa: 0.698, Tv_monitor: 0.342	Bottle: 0.650, Chair: 0.825, Dining_table: 0.688, Person: 1.00, Potted_Plant: 0.470

Figure 6.3: Illustrative examples of the proposed framework for image annotation.

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.

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 10 (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. Equation 4.8).

Therefore, according to Equation 4.1, the correlation likelihood between a candidate annotation CA and a given image i'_i can be computed as follows:

$$glb(\mathcal{KB}, a : \exists\, isAnnotatedBy.CA) = \alpha_b \otimes \alpha_r = (\alpha_1 \otimes \alpha_2 \otimes \alpha_3) \otimes \mu_{\text{isAnnotatedBy}(a,b)} \quad (6.4)$$

then,

$$ValidCA \equiv \exists\, isAnnotatedBy.CA \quad (6.5)$$

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:

$$bex(\mathcal{KB}, ValidCA) = \{\langle a, r \rangle | r = bed(\mathcal{KB}, a : ValidCA)\} \quad (6.6)$$

For instance, let us consider the first example in Figure 6.3. We show below some cases of DL reasoning using the contextual information:

Example 11 (DL Reasoning using image context)**.**

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

Consequently, with respect to the contextual information, the best explanation for the left image in Figure 6.3 is: $CA_1 \equiv Person \sqcap Sheep$.

Please note that, since most images of the Pascal VOC dataset contain only one or two concepts [Choi et al., 2010], and thus the distribution of multi-labeled images is not uniform, we computed the Equation 4.8 for this dataset as:

$$\mu_{\text{isAnnotatedBy}(Photo, Annotation_i)} = \frac{n_{Annotation_i}}{\mathcal{L}} * exp(\Lambda) \tag{6.7}$$

where $\Lambda = |Annotation_i|$.

6.4.4 Reasoning on Image Annotation 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. In Chapter 4 - Section 4.7, we have proposed a usage scenario illustrating the usefulness of spatial information and reasoning over this kind of knowledge in order to improve image annotation.

As part of this work, we 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 - cf. Algorithm 3.

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 3 details the different steps of this method.

Algorithm 3: Reasoning using spatial Information

Input: A valid candidate annotation: ValidCA
Result: Semantically consistent image annotation
begin
Find:
- $C, D \leftarrow \underset{x,y \in ValidCA}{\text{argmax}} \; x.hasAppearedwith(y)$
- Spatial arrangement $\leftarrow \underset{\chi \in \mathcal{DIR}}{\text{argmax}} \; C.\chi(D)$
- $E \leftarrow \underset{x \in ValidCA}{\text{argmax}} \; x.hasAppearedwith(C \sqcup D)$
- 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 6.3 - first example: "*This picture depicts a person standing on the left of a sheep. They are close to each other.*"

Figure 6.3 - 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 the information about detected concepts and their spatial location [Gupta and Mannem, 2012]. We will address the implementation of such a system in our future work.

6.5 Experiments

For evaluating our multi-stage reasoning framework for image annotation, we used two public datasets: Pascal VOC'2009 [Everingham et al., 2009], and Pascal VOC'2010 [Everingham et al., 2010] datasets. These datasets contain about 11 000 images annotated with 20 predefined concepts. Each image is annotated by one or more concepts (multi-labeled images). A fair comparison of our proposal to existing methods for image annotation is also performed on these datasets. in the following, we first introduce the used method for visual representation of images, then we present the obtained experimental results.

6.5.1 Visual Representation of Images

The Bag-of-Features (BoF) representation, also known as Bag-of-Visual-Word (BoVW), is used in this work 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 [Li and Perona, 2005]. In our approach, image features are described as follows: Lowe's DoG Detector [Lowe, 1999] is used for detecting a set of salient image regions. A signature of these regions is then computed using SIFT descriptor [Lowe, 1999]. Afterwards, given the collection of detected region from the training set of all categories, we generate a codebook of size $K = 1000$ 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.

6.5.2 Evaluation

Our experiments are performed on Pascal VOC'2009 and Pascal VOC'2010 datasets. Since we do not disposed of the test set, we used the training set for training our concept classifiers and the validation set for evaluating our approach.

In Figure 6.4, we compare our framework for image annotation to the following methods:

- A flat classification method.
- A hierarchical classification method using the hierarchy illustrated in Figure 4.5 and OVA classifiers.
- 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

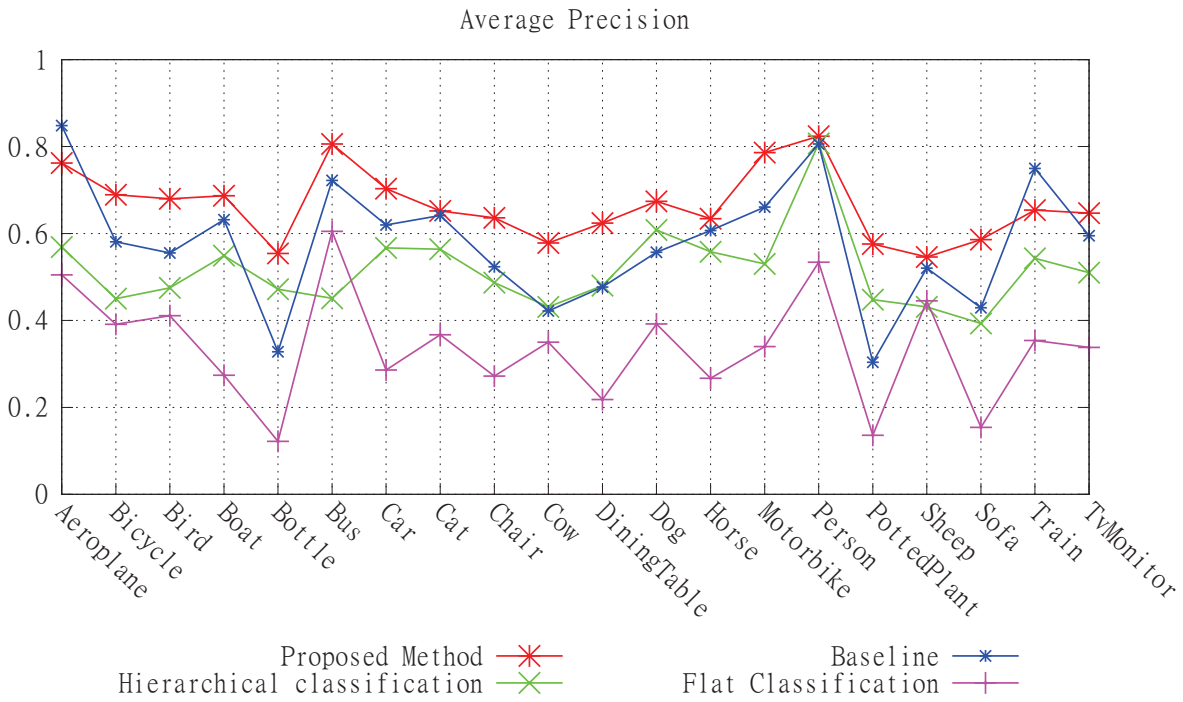


Figure 6.4: Comparison of our framework for image annotation with: a flat classification method, a hierarchical classification one, and a baseline method. The comparison is performed on VOC'2010 dataset.

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 Figure 4.5. For more details about the hierarchical classification method see Section 6.4.1. The results are evaluated in terms of Average Precision (AP) scores.

As illustrated in Figure 6.4, our method for image annotation performs better results than the other ones, 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 much more multi-labeled images. Indeed, in the Pascal VOC dataset the proportion of images labeled with more than 2 concepts is small comparing to the total number of images [Choi et al., 2010].

In table 6.1, we compare our multi-stage reasoning framework for image annotation to the method of [Zhou et al., 2010] on Pascal VOC'2009 dataset. [Zhou et al., 2010] proposed a method for image classification using local visual descriptors 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 de-

	Proposed Method (AP)	[Zhou et al., 2010] (AP)
Aeroplane	82.2	87.1
Bicycle	74.1	67.4
Bird	69.2	65.8
Boat	64.5	72.3
Bottle	52.1	40.9
Bus	80.4	78.3
Car	70.1	69.7
Cat	61.7	69.7
Chair	63.8	58.5
Cow	62.7	50.1
Dining_Table	68.9	55.1
Dog	63.2	56.3
Horse	62.7	71.8
Motorbike	76.1	70.8
Person	83.2	84.1
Potted_Plant	57.1	31.4
Sheep	64.4	51.5
Sofa	58.1	55.1
Train	72.8	84.7
Tv_Monitor	66.7	65.2
AP on all concepts	67.7	64,29

Table 6.1: Comparison of our method for image annotation with the one of [Zhou et al., 2010] on Pascal VOC'2009 dataset.

scriptors are then aggregated to form image-level feature vector. The image-level feature vector is finally fed into a classifier to perform image classification. As illustrated in table 6.1, our method performs better than the one proposed by [Zhou et al., 2010], and achieves a gain of +3.41%. 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 Figure 6.3, 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'2010 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 more efficient image descriptor since this is not

the aim of our paper. Accordingly, the obtained results can be further improved by incorporating other image features for example.

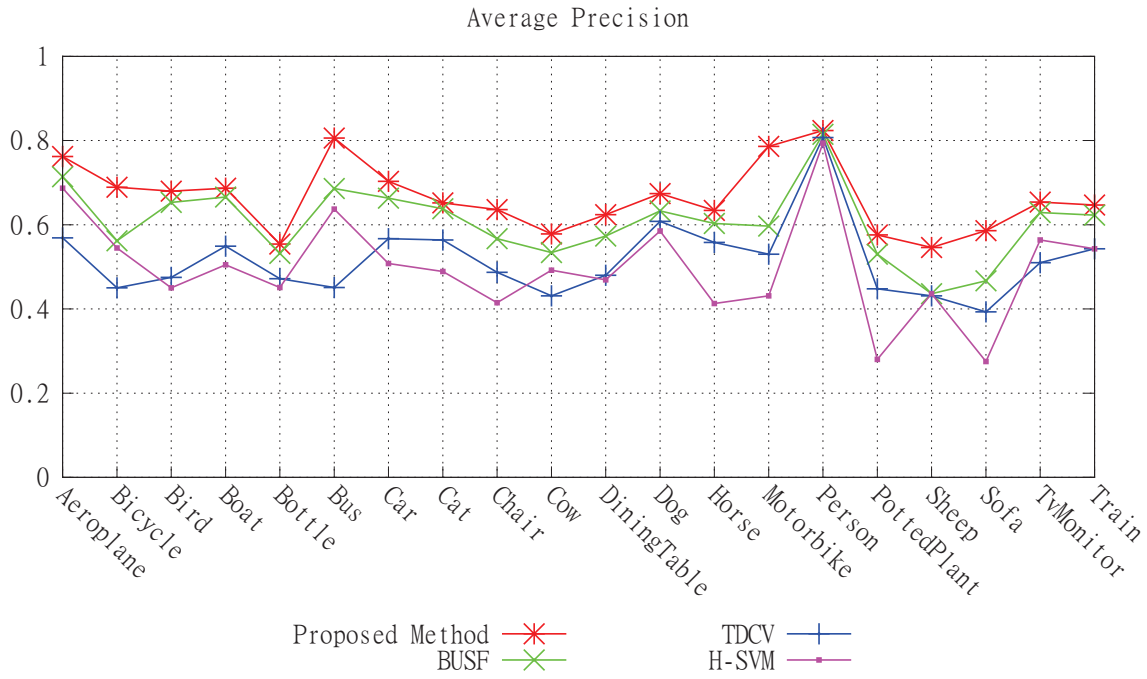


Figure 6.5: Comparison of our framework for image annotation to previous work on Pascal VOC'2010 dataset. Our approach outperforms on all classes comparing to the other ones.

In Figure 6.5, we compare our framework for image annotation to the following methods: Bottom-Up Score Fusion (*BUSF*) and Top-Down Classifiers Voting (*TDCV*) which were proposed in Chapter 5, and Hierarchy of SVM (*H-SVM*) method of [Marszalek and Schmid, 2007]. As it can be seen on this figure, our multi-stage reasoning framework for image annotation outperforms on all classes comparing to the other ones. Please note that this comparison was performed using the same dataset, i.e. Pascal VOC'2010, the same training/validation sets from the VOC'2010 dataset, and the same visual representation of images - cf. Section 6.5.1. Therefore, it is clear that the proposed approach allows achieving a significant improvement for image annotation, and that the contextual and spatial information are of great importance to improve the accuracy of image annotation.

Finally, we want to highlight that some images in the VOC dataset are badly annotated. For instance in the third example of Figure 6.3, 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 a false positive detection in the evaluation of our method since they are missing in the ground-truth. For the second example of

Figure 6.6: An example of a badly annotated image in the VOC'2010 dataset. Ground-truth: Person. Annotation provided by our method: $\langle Bottle : 0.982, Chair : 0.281, Dining\_table : 0.493, Person : 1.00, Tv\_monitor : 0.333 \rangle$.

Figure 6.3, 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" and "Coffee_Table". In Figure 6.6, 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 these concepts, but according to the ground-truth these detections counted as false positives.

6.6 Conclusion

In Chapter 4 we have proposed an approach to automatically build a fuzzy multimedia ontology suitable for image annotation and interpretation. Our multimedia ontology incorporates visual, conceptual, contextual and spatial knowledge in order to effectively model image semantics. Based on our multimedia ontology, we proposed in this chapter a new method for image annotation based on hierarchical image classification and a multi-stage reasoning framework for checking the consistency of the produced annotation. The proposed framework uses different knowledge about image concepts in order to remove the uncertainty about image annotation introduced during the image classification process. These knowledge include: subsumption rela-

tionships, contextual relationships and spatial relationships. The reasoning tasks are achieved in Fuzzy Description Logics (Fuzzy-DLs). An empirical evaluation of our approach is performed on Pascal VOC'2009 and Pascal VOC'2010 datasets. The obtained results have shown a significant improvement on the average precision results, thus demonstrating the effectiveness of the proposed method.

As an extension to this work, we propose in the near future to test our approach using more robust image descriptors, and using a larger image datasets. Furthermore, a good direction to enhance our proposal is to implement a system that provides a human-like descriptions of images, similar to the work proposed by [Gupta and Mannem, 2012]. Indeed, the authors have shown in their paper that it is possible to automatically generate a description of an image based on its annotation. We also plan to implement and test the usage scenarios described in Chapter 4 - Section 4.7.

Conclusions and Future Research Directions

Chapter 7

Conclusions and Future Research Directions

Image annotation represents a major economic issue and involves several applications, including but not limited to: social networks, the Web, medical imaging, biocomputing, remote sensing, news, image and video monitoring and information retrieval. Recently, many advances have been made to reduce the semantic gap problem, and to improve automatic image annotation systems. These advances include computer vision techniques (such as image features, image descriptors, sparse representation, etc.) and artificial intelligence (such as machine learning, data clustering, knowledge engineering, etc.). However, automatic image annotation is still an open issue and does not seem to be resolved in the next few years. Indeed, current systems for image annotation face significant problems, including the semantic gap, the scalability problem, the subjectivity of image semantics, the lack of robustness of current computer vision techniques, and more.

In this dissertation, we address the problem of automatic image annotation from the perspective of building and using explicit semantic models, such as semantic hierarchies and ontologies. Our overall goal is to narrow the semantic gap and to achieve a semantically consistent image annotation. Indeed, recent efforts on knowledge modeling and knowledge-driven approaches for image annotation have shown that it is possible to reduce the semantic gap. Our concern in this thesis was therefore to demonstrate this assumption, and to show how it would be possible to use these explicit and structured knowledge models to improve image annotation.

The remaining of this chapter is structured as follows. The first part reviews the different contributions proposed in this PhD work and their significance. Specifically, we focus on how these contributions can improve existing systems for image annotation and what issues could be answered/narrowed using the proposed knowledge models. In the second part, we present our perspectives to improve the proposed approaches and our future research directions.

7.1 Contributions and their Significance

As aforementioned, the aims of this thesis are twofold. As a first step, we have focused our contributions on providing effective tools for the automatic building of structured and explicit knowledge models dedicated to image annotation. This is intended to propose knowledge models that are representative of image semantics, and allow modeling image contexts in an accurate way. Subsequently, we have concentrated our contributions on the effective use of this knowledge in order to achieve a more accurate image annotation. Our final purpose is to narrow the semantic gap, i.e. decrease the gap between the visual content of images and their meaning (image semantics). Therefore, the achieved contributions in this dissertation can be categorized in two main topics: 1) the automatic building of explicit and structured multimedia knowledge models, 2) the use of structured knowledge models to improve image annotation, and specifically, the use of formal fuzzy reasoning in order to reduce the uncertainty about image annotation. The detailed set of accomplishments are introduced in the following.

7.1.1 A Thorough Study of State-of-the-art

We proposed in this dissertation a thorough survey of the state-of-the-art regarding image annotation approaches. A special attention was given to knowledge-driven approaches for semantic image annotation. Firstly, we have summarized the proposals of current approaches, and then we pointed out what is missing in these approaches and which directions could be borrowed to narrow the semantic gap. Because of the huge number of existing methods for image annotation, the proposed state of the art is not exhaustive, and was intended only to show the limits of current approaches. Therefore, we have focused on the set of methods that we have considered as relevant with respect to the topic of our work. We finally have proposed a tentative to classify the different knowledge-based approaches for image annotation according to the level of expressiveness of the knowledge models [Bannour and Hudelot, 2011].

Significance:

⇒ The proposed survey of the state-of-the-art on semantic image annotation is focused on knowledge-based approaches for image annotation. This survey was the starting point of the directions taken in this PhD work, and the proposed contributions. It has also served to explain our standpoints about several issues related to image annotation, image retrieval, and terminologies that follow from these topics.

7.1.2 Building Semantic Hierarchies Faithful to Image Semantics

The first main contribution of this dissertation deals with the automatic building of semantic hierarchies for the purpose of image annotation. Indeed, according to our

assumptions in the Chapter 3 - Section 3.2.4, a suitable semantic hierarchy for image annotation should take account of the visual information and also the conceptual and the contextual ones. Consequently, we have proposed a new image-semantic measure, named '*Semantico-Visual Relatedness of Concepts*' (SVRC), which allows to estimate the semantic similarity between concepts with respect to these three kinds of information. This proposed measure incorporates visual, conceptual and contextual information in order to provide a measure which is more meaningful and more representative of image semantics. Subsequently, we have proposed a new methodology to automatically build semantic hierarchies suitable for image annotation. The building is based on the previously proposed measure (*SVRC*) and on a set of simple but effective rules, named *TRUST-ME*, in order to connect the concepts with higher relatedness till the building of the final hierarchy [Bannour and Hudelot, 2012a, Bannour and Hudelot, 2012b, Bannour and Hudelot, 2013b].

Significance:

⇒ The proposed measure, *SVRC*, is representative of image semantics as it incorporates the different modalities of images. Indeed, our experiments have shown that image modalities have different distributions in the semantic space, and their combination can lead to a more meaningful correlation between concepts in the image domain.

⇒ The built semantic hierarchy is faithful to image semantics, since it is a based on the previously proposed measure (i.e. *SVRC*) and a set of rules allowing to link the concepts sharing a higher "semantico-visual" relatedness.

⇒ The semantico-visual relatedness measure, and consequently the built hierarchy, can adapt to an application domain and even to the user background since it incorporates a contextual similarity which is domain dependant.

7.1.3 Building Fuzzy Multimedia Ontologies for Image Annotation

A major issue with semantic hierarchies is that they are limited to describe the subsumption relationships between concepts, i.e. only the '*is-a*' relationship is used in these semantic structures. In this dissertation, we proposed to go further in the modeling of image-concepts relationships, and therefore the modeling of image semantics. Consequently, we proposed a new approach for building an ontology of spatial and contextual information, suitable for reasoning about the coherence of image annotation. In our approach, visual and conceptual information were used to build a semantic hierarchy that will serve as a backbone of our ontology. Contextual and spatial information about image concepts were thereafter incorporated in the ontology to model other semantic relationships between these concepts. Fuzzy description logics were used as a formalism to represent our ontology and the inherent uncertainty and imprecision of this kind of information. The proposed ontology was automatically built by mining image databases and the provided annotations, which

make it suitable for the image domain [Bannour and Hudelot, 2013a].

Significance:

- The building of our ontology is fully automatic and allows therefore reducing the scalability problem of (formal) ontologies building.
- The specification of our ontology and the represented knowledge are not extracted from an existing commonsense ontology, but gathered by the mining of image databases. Consequently, the built ontology is representative of image semantics and is domain-independent.
- Our multimedia ontology is designed as a knowledge base and allows *terminological* and *assertional* reasoning. Specifically, it allows performing fuzzy reasoning in order to reduce the uncertainty about image annotations.

7.1.4 Improving Image Annotation using Semantic Hierarchies

Image annotation has been considered in the last decade as a multi-class classification problem, i.e. finding the belonging of a considered image to a given set of concepts (or classes). Recently, many approaches have proposed to use semantic hierarchies in order to improve image annotation. These hierarchies was used as a framework for hierarchical image classification, and thus to improve classifiers accuracy and to reduce the complexity of managing large scale data. In this dissertation, we investigated the contribution of semantic hierarchies for hierarchical image classification. We proposed therefore a new method based on the hierarchy structure to train efficiently hierarchical classifiers. Our method, named One-Versus-Opposite-Nodes, allows decomposing the problem into several independent tasks and therefore it scales well with large databases. We also proposed two methods for computing a hierarchical decision function that served to annotate new image samples. The former was performed using a top-down classifiers voting, while the second was based on a bottom-up score fusion [Bannour and Hudelot, 2012c, Bannour and Hudelot, 2013b].

Significance:

⇒ Previous approaches for hierarchical image classification have proposed to build hierarchies by the recursive partitioning of the set of classes, or by the agglomerative clustering of the classes, resulting in binary trees like structures. In this dissertation, we showed that the hierarchy depth has a serious impact on the classification accuracy, and therefore we proposed to not restrict the image classification to binary structures.

⇒ The proposed methods draw profits from the semantic structure of the hierarchy to decompose the image classification problem into several independent and complementary sub-tasks, resulting in a lower complexity and a higher accuracy for the image classification results.

⇒ Performing image classification using our framework allows a multi-level of abstraction image annotation, i.e. the target image is annotated by a set of concepts (from specific to generic) of a given path in the hierarchy.

7.1.5 Multi-Stage Reasoning Framework for Image Annotation

Finally, we proposed a framework 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 relevant decision on the belonging of a given image to the set of concept classes. 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. Our approach is of interest since it illustrates the reasoning power of multimedia ontologies and their effectiveness for achieving good decisions on image annotation [Bannour and Hudelot, 2013a].

Significance:

⇒ Previous methods dealing with multimedia ontologies are only providing some formalism to use ontologies as a repository for storing knowledge about image content. Few attempts have considered the problem of reasoning about this knowledge, but on limited data and without considering real applications. Thus, the effectiveness of the stored knowledge has to be proved. In this dissertation, we address deeply the reasoning process to show the usefulness of multimedia ontologies and the effectiveness of the stored knowledge about images.

⇒ Our framework has allowed solving many issues of image annotation. Indeed, image annotation is a difficult task because of the uncertainty introduced by the statistical learning algorithms, the scalability problem, and the dependency on the accuracy of the ground truth of the training dataset. Our approach uses explicit knowledge in order to reduce this uncertainty by supplying a formal framework to reason about the consistency of extracted information from images.

7.2 Future Research Directions

Many contributions were proposed in this dissertation to deal with the image annotation problem. These contributions are in no way complete solutions, and could be improved in several manners. Indeed, as aforementioned in Chapter 2, the semantic gap problem is an open-ended problem and seems to not be solved in the near future. Moreover, automatic image annotation is a difficult task because of the uncertainty introduced by machine learning algorithms. Thus, the topic of image annotation is witnessing an incremental improvement, and many efforts are still needed before

achieving efficient systems dealing with this problem. In the following, we propose potential directions that can be explored further.

- Our major concern in this dissertation was to provide knowledge-driven models for improving image annotation accuracy. Thus, in this work we have not sought to implement more robust/sophisticated image descriptors when performing our evaluation. It is therefore of interest to see how far the proposed approaches can be used to push the state-of-the-art performance by using one of the most sophisticated classification (or detection) method.

- In order to evaluate the robustness of our approaches with large scale data, it is important to test our proposals on larger image datasets, i.e. including thousands of concepts and millions of images. This will allow for instance, to see if the proposed methods for building semantic hierarchies and hierarchical image classification are still efficient with such voluminous data. It is also widely known that DL reasoners suffer from the scalability problem, and it is important to survey this limit on real data applications.

- As an extension to our proposal in Chapter 4 dealing with the building of multimedia ontologies, we propose in the near future to implement and to test the usage scenarios described in Section 4.7. Indeed, two methods were proposed: the first allowing to learn very complex concepts by mining multimedia datasets, while the second allows reducing significantly the complexity of object detection process (specifically when dealing with a large annotation vocabulary).

- Implementing a system that provides a human-like descriptions of images. Similar works were proposed by [Kulkarni et al., 2011, Gupta and Mannem, 2012]. Indeed, the authors have shown in their papers that it is possible to automatically generate a textual description of images based on their annotations. Our framework for image annotation, proposed in Chapter 6, allows to dispose of meaningful attributes about concepts relationships. Therefore, performing such an approach for image description on the outputs of our framework will allow achieving an efficient description of images, i.e. a detailed description of image content through the description of image concepts and concept relationships.

Appendices

Appendix 1: Publications of the Author

International Journals with Peer-review:

(1) Building and Using Fuzzy Multimedia Ontologies for Semantic Image Annotation, *MTAP 2013, Multimedia Tools and Applications*, to appear, doi={10.1007/s11042-013-1491-z}.

(2) Imiol: a System for Indexing Images by their Semantic Content Based on Possibilistic Fuzzy Clustering and Adaptive Resonance Theory Neural Networks Learning, *AAI 2010, Applied Artificial Intelligence*, Volume 24, Issue 9, pages 821-846, October 2010.

National Journals with Peer-review:

(1) Construction de Hiérarchies Sémantiques pour l'Annotation d'Images, *RIA 2013, Revue d'Intelligence Artificielle*, Volume 27, Number 1, pages 11-37, 2013.

International Conferences with Peer-review:

(1) Hierarchical Image Annotation Using Semantic Hierarchies, *CIKM 2012, 21st ACM International Conference on Information and Knowledge Management*, Maui, Hawaii, USA, October 2012.

(2) Building Semantic Hierarchies Faithful to Image Semantics, *MMM 2012, 18th International Multimedia Modeling Conference*, Klagenfurt, Austria, January 2012.

(3) Towards Ontologies for Image Interpretation and Annotation, *CBMI 2011, Content-Based Multimedia Indexing, 2011 9th International Workshop on*, Madrid, Espagne, pages 211-216, Juin 2011.

National Conferences with Peer-review:

(1) Combinaison d'information visuelle, conceptuelle, et contextuelle pour la construction automatique de hiérarchies sémantiques adaptées à l'annotation d'images, *RFIA 2012, Actes de la conférence Reconnaissance des Formes et Intelligence Artificielle*, Lyon, France, January 2012. (Awarded Best Paper of the conference)

Bibliography

[Ah-Pine et al., 2009] Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., and Renders, J.-M. (2009). Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications*, 42:31–56.

[Alm, 2006] Alm, C. O. (2006). Challenges for annotating images for sense disambiguation. In *ACL workshop on Frontiers in Linguistically Annotated Corpora*.

[Armitage and Enser, 1997] Armitage, L. H. and Enser, P. G. (1997). Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299.

[Arndt et al., 2007] Arndt, R., Troncy, R., Staab, S., Hardman, L., and Vacura, M. (2007). COMM: designing a well-founded multimedia ontology for the web. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 30–43, Berlin, Heidelberg. Springer-Verlag.

[Atif et al., 2011] Atif, J., Hudelot, C., and Bloch, I. (2011). Abduction in description logics using formal concept analysis and mathematical morphology: application to image interpretation. In *In Concept Lattices and Applications (CLA'11)*.

[Atif et al., 2013] Atif, J., Hudelot, C., and Bloch, I. (2013). Explanatory reasoning for image understanding using formal concept analysis and description logics. *IEEE Transactions on Systems, Man and Cybernetics*, page To appear.

[Baader, 2011] Baader, F. (2011). What's new in description logics. *Informatik-Spektrum*, 34:434–442.

[Baader et al., 2003] Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

[Bagdanov et al., 2007] Bagdanov, A. D., Bertini, M., Bimbo, A. D., Serra, G., and Torniai, C. (2007). Semantic annotation and retrieval of video events using multimedia ontologies. In *International Conference on Semantic Computing (ICSC'07)*, pages 713 –720.

[Banerjee and Pedersen, 2003] Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial intelligence (IJCAI'03)*, pages 805–810.

[Bannour, 2009] Bannour, H. (2009). Une approche sémantique basée sur l'apprentissage pour la recherche d'image par contenu. In *COnférence en Recherche d'Infomations et Applications (CORIA'09)*, pages 471–478.

[Bannour et al., 2009] Bannour, H., Hlaoua, L., and el Ayeb, B. (2009). Survey of the adequate descriptor for content-based image retrieval on the web: Global versus local features. In *COnférence en Recherche d'Infomations et Applications (CORIA'09)*, pages 445–456.

[Bannour and Hudelot, 2011] Bannour, H. and Hudelot, C. (2011). Towards ontologies for image interpretation and annotation. In *Content-Based Multimedia Indexing (CBMI'11)*.

[Bannour and Hudelot, 2012a] Bannour, H. and Hudelot, C. (2012a). Building Semantic Hierarchies Faithful to Image Semantics. In *Proceedings of the 18th international conference on Advances in Multimedia Modeling (MMM'12)*, pages 4–15.

[Bannour and Hudelot, 2012b] Bannour, H. and Hudelot, C. (2012b). Combinaison d'information visuelle, conceptuelle, et contextuelle pour la construction automatique de hiérarchies sémantiques adaptées à l'annotation d'images. In *Actes de la conférence Reconnaissance des Formes et Intelligence Artificielle (RFIA 2012)*, pages 462–469, Lyon, France.

[Bannour and Hudelot, 2012c] Bannour, H. and Hudelot, C. (2012c). Hierarchical image annotation using semantic hierarchies. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*, pages 2431–2434.

[Bannour and Hudelot, 2013a] Bannour, H. and Hudelot, C. (2013a). Building and using fuzzy multimedia ontologies for semantic image annotation. *Multimedia Tools and Applications (MTAP'13)*, pages 1–35. http://dx.doi.org/10.1007/s11042-013-1491-z.

[Bannour and Hudelot, 2013b] Bannour, H. and Hudelot, C. (2013b). Construction de hiérarchies sémantiques pour l'annotation d'images. *Revue d'Intelligence Artificielle (RIA'13)*, 27(1):11–37.

[Barnard et al., 2003] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

[Barnard and Forsyth, 2001] Barnard, K. and Forsyth, D. A. (2001). Learning the semantics of words and pictures. In *Proceedings of the International Conference on Computer Vision (ICCV'11)*, pages 408–415.

[Bart et al., 2008] Bart, E., Porteous, I., Perona, P., and Welling, M. (2008). Unsupervised learning of visual taxonomies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

[Belkhatir et al., 2004] Belkhatir, M., Mulhem, P., and Chiaramella, Y. (2004). Integrating perceptual signal features within a multi-facetted conceptual model for automatic image retrieval. In *Advances in Information Retrieval, 26th European Conference on IR Research (ECIR'04)*, pages 267–282.

[Bengio et al., 2010] Bengio, S., Weston, J., and Grangier, D. (2010). Label embedding trees for large multi-class tasks. In *Advances in Neural Information Processing Systems (NIPS'10)*, pages 163–171.

[Biederman, 1972] Biederman, I. (1972). Perceiving Real-World Scenes. *Science*, 177:77–80.

[Biederman, 1987] Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147.

[Blei et al., 2004] Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Neural Information Processing Systems (NIPS'04)*.

[Blei and Jordan, 2003] Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR'03)*, pages 127–134, New York, NY, USA. ACM.

[Bloch, 1999] Bloch, I. (1999). Fuzzy relative position between objects in image processing: A morphological approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'99)*, 21:657–664.

[Bloch, 2005] Bloch, I. (2005). Fuzzy spatial relationships for image processing and interpretation: a review. *Image and Vision Computing*, 23(2):89 – 110.

[Bobillo and Straccia, 2009] Bobillo, F. and Straccia, U. (2009). Fuzzy description logics with general t-norms and datatypes. *Fuzzy Sets and Systems*, 160(23):3382–3402.

[Bobillo and Straccia, 2011] Bobillo, F. and Straccia, U. (2011). Reasoning with the finitely many-valued lukasiewicz fuzzy description logic SROIQ. *Information Sciences*, 181(4):758–778.

[Boucher and Le, 2005] Boucher, A. and Le, T. (2005). Comment extraire la semantique d'une image? In *Conference Internationale Sciences Electroniques, Technologies de l'Information et des Telecommunications (SETIT'05)*, pages 295–306, Tunisie.

[Boutell et al., 2004] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.

[Bruno et al., 2008] Bruno, E., Moenne-Loccoz, N., and Marchand-Maillet, S. (2008). Design of multimodal dissimilarity spaces for retrieval of video documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'08)*, 30(9):1520–1533.

[Budanitsky and Hirst, 2006] Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32:13–47.

[Carbonetto et al., 2004] Carbonetto, P., de Freitas, N., and Barnard, K. (2004). A statistical model for general contextual object recognition. In *Proceedings of the European Conference on Computer Vision (ECCV'04)*, Lecture Notes in Computer Science, pages 350–362. Springer.

[Carneiro et al., 2007] Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3).

[Carneiro and Vasconcelos, 2005] Carneiro, G. and Vasconcelos, N. (2005). Formulating semantic image annotation as a supervised learning problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 163–168 vol. 2.

[Carson et al., 2002] Carson, C., Belongie, S., Greenspan, H., and Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'02)*, 24(8):1026–1038.

[Cevikalp, 2010] Cevikalp, H. (2010). New clustering algorithms for the support vector machine based hierarchical classification. *Pattern Recognition Letters*, 31(11):1285 – 1291.

[Chen and Wang, 2004] Chen, Y. and Wang, J. Z. (2004). Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939.

[Choi et al., 2010] Choi, M. J., Lim, J., Torralba, A., and Willsky, A. (2010). Exploiting hierarchical context on a large database of object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 129 –136.

[Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

[Clinchant et al., 2011] Clinchant, S., Ah-Pine, J., and Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 44:1–44:8, New York, NY, USA. ACM.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 886 –893.

[Dasiopoulou and Kompatsiaris, 2010] Dasiopoulou, S. and Kompatsiaris, I. (2010). Trends and issues in description logics frameworks for image interpretation. *Artificial Intelligence: Theories, Models and Applications*, 6040:61–70.

[Dasiopoulou et al., 2009] Dasiopoulou, S., Kompatsiaris, I., and Strintzis, M. (2009). Applying fuzzy DLs in the extraction of image semantics. In Spaccapietra, S. and Delcambre, L., editors, *Journal on Data Semantics XIV*, volume 5880 of *Lecture Notes in Computer Science*, pages 105–132. Springer Berlin / Heidelberg.

[Dasiopoulou et al., 2008] Dasiopoulou, S., Kompatsiaris, I., and Strintzis, M. G. (2008). Using Fuzzy DLs to Enhance Semantic Image Analysis. In *International Conference on Semantic and Digital Media Technologies (SAMT'08)*, pages 31–46.

[Dasiopoulou et al., 2010] Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., and Strintzis, M. G. (2010). Enquiring MPEG-7 based multimedia ontologies. *Multimedia Tools and Applications (MTAP'10)*, 46:331–370.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[Deng et al., 2011a] Deng, J., Berg, A., and Fei-Fei, L. (2011a). Hierarchical semantic indexing for large scale image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, pages 785 –792.

[Deng et al., 2010] Deng, J., Berg, A. C., Li, K., and Fei-Fei, L. (2010). What does classifying more than 10,000 image categories tell us? In *Proceedings of the European Conference on Computer Vision (ECCV'10)*.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*.

[Deng et al., 2011b] Deng, J., Satheesh, S., Berg, A. C., and Li, F.-F. (2011b). Fast and balanced: Efficient label tree learning for large scale object recognition. In *Advances in Neural Information Processing Systems (NIPS'11)*, pages 567–575.

[Deselaers and Ferrari, 2011] Deselaers, T. and Ferrari, V. (2011). Visual and semantic similarity in ImageNet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, pages 1777 –1784.

[Divvala et al., 2009] Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., and Hebert, M. (2009). An empirical study of context in object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 1271–1278.

[Djeraba, 2003] Djeraba, C. (2003). Association and content-based retrieval. *IEEE Transaction on Knowledge and Data Engineering (TDDE'03)*, 15:118–135.

[Dong and Li, 2006] Dong, A. and Li, H. (2006). Multi-ontology based multimedia annotation for domain-specific information retrieval. In *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy (SUTC'06)*, pages 158–165.

[Dong et al., 2012] Dong, P., Mei, K., Zheng, N., Lei, H., and Fan, J. (2012). Training inter-related classifiers for automatic image classification and annotation. *Pattern Recognition*, 0(0):to appear.

[Duda et al., 2001] Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*. Wiley.

[Duygulu et al., 2002] Duygulu, P., Barnard, K., Freitas, J. F. G. d., and Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision (ECCV'02)*, pages 97–112. Springer-Verlag.

[Eakins, 2002] Eakins, J. P. (2002). Towards intelligent image retrieval. *Pattern Recognition*, 35(1):3 – 14.

[Enser and Sandom, 2003] Enser, P. and Sandom, C. (2003). Towards a comprehensive survey of the semantic gap in visual image retrieval. In *Proceedings of the 2nd international conference on Image and video retrieval*, CIVR'03, pages 291–299.

[Enser, 1993] Enser, P. G. B. (1993). Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52.

[Escalante et al., 2008] Escalante, H. J., Hérnadez, C. A., Sucar, L. E., and Montes, M. (2008). Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval (MIR'08)*, pages 172–179. ACM.

[Everingham et al., 2009] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2009). The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[Fan et al., 2007] Fan, J., Gao, Y., and Luo, H. (2007). Hierarchical classification for automatic image annotation. In *international ACM SIGIR conference on Research and development in information retrieval (SIGIR'07)*, pages 111–118.

[Fan et al., 2008a] Fan, J., Gao, Y., and Luo, H. (2008a). Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Transactions on Image Processing (TIP'08)*, 17(3):407–426.

[Fan et al., 2008b] Fan, J., Gao, Y., Luo, H., and Jain, R. (2008b). Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia (TMM'08)*, 10(2):167–187.

[Fan et al., 2009] Fan, J., Luo, H., Shen, Y., and Yang, C. (2009). Integrating visual and semantic contexts for topic network generation and word sense disambiguation. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'09)*.

[Fei-Fei and Perona, 2005] Fei-Fei, L. and 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'2005)*, volume 2, pages 524 – 531.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

[Felzenszwalb et al., 2010] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'10)*, 32(9):1627 –1645.

[Feng et al., 2004] Feng, S., Manmatha, R., and Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pages 1002–1009.

[Feng and Lapata, 2010] Feng, Y. and Lapata, M. (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT'10, pages 831–839.

[Fergus et al., 2003] Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, volume 2, pages II–264 – II–271.

[Fergus et al., 2009] Fergus, R., Weiss, Y., and Torralba, A. (2009). Semi-supervised learning in gigantic image collections. In *Neural Information Processing Systems (NIPS'09)*, pages 522–530.

[Forsyth and Fleck, 1997] Forsyth, D. A. and Fleck, M. M. (1997). Body plans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, page 678. IEEE Computer Society.

[Galleguillos and Belongie, 2010] Galleguillos, C. and Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722.

[Galleguillos et al., 2010] Galleguillos, C., McFee, B., Belongie, S., and Lanckriet, G. (2010). Multi-class object localization by combining local contextual interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 113 –120.

[Galleguillos et al., 2008] Galleguillos, C., Rabinovich, A., and Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8.

[Gangemi et al., 2002] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with DOLCE. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management - Ontologies and the Semantic Web (EKAW 02)*, pages 166–181, London, UK, UK. Springer-Verlag.

[Gao and Koller, 2011] Gao, T. and Koller, D. (2011). Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *Proceedings of the International Conference on Computer Vision (ICCV'11)*.

[Goh et al., 2001] Goh, K.-S., Chang, E., and Cheng, K.-T. (2001). SVM binary classifier ensembles for image classification. In *Proceedings of the tenth international conference on Information and knowledge management (CIKM'01)*, pages 395–402.

[Gould et al., 2009] Gould, S., Fulton, R., and Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In *Proceedings of the International Conference on Computer Vision (ICCV'09)*, pages 1 –8.

[Griffin and Perona, 2008] Griffin, G. and Perona, P. (2008). Learning and using taxonomies for fast visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

[Gronau et al., 2008] Gronau, N., Neta, M., and Bar, M. (2008). Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, 20(3):371–388.

[Gruber, 1995] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928.

[Guillaumin et al., 2010] Guillaumin, M., Verbeek, J., and Schmid, C. (2010). Multimodal semi-supervised learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 902 –909.

[Gupta and Mannem, 2012] Gupta, A. and Mannem, P. (2012). From image annotation to image description. *Neural Information Processing*, 7667:196–204.

[Hamadi et al., 2012] Hamadi, A., Quenot, G., and Mulhem, P. (2012). Two-layers re-ranking approach based on contextual information for visual concepts detection in videos. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6.

[Hammer and Villmann, 2007] Hammer, B. and Villmann, T. (2007). How to process uncertainty in machine learning? In *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN'07)*, pages 79 – 90.

[Hare et al., 2006] Hare, J. S., Sinclair, P. A. S., Lewis, P. H., Martinez, K., Enser, P. G., and Sandom, C. J. (2006). Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. In *European Semantic Web Conference (ESWC'06)*.

[Hartz and Neumann, 2007] Hartz, J. and Neumann, B. (2007). Learning a knowledge base of ontological concepts for high-level scene interpretation. In *International Conference on Machine Learning and Applications (ICMLA'07)*, pages 436–443.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

[Hauptmann et al., 2007] Hauptmann, A., Yan, R., and Lin, W.-H. (2007). How many high-level concepts will fill the semantic gap in news video retrieval? In *International Conference on Image and Video Retrieval (CIVR'07)*, pages 627–634.

[Heitz and Koller, 2008] Heitz, G. and Koller, D. (2008). Learning spatial context: Using stuff to find things. In *Proceedings of the European Conference on Computer Vision (ECCV'08)*, pages 30–43.

[Hollink et al., 2004] Hollink, L., Nguyen, G., Schreiber, G., Wielemaker, J., Wielinga, B., and Worring, M. (2004). Adding spatial semantics to image annotations. In *International Workshop on Knowledge Markup and Semantic Annotation*.

[Hoogs et al., 2003] Hoogs, A., Rittscher, J., Stein, G., and Schmiederer, J. (2003). Video content annotation using visual analysis and a large semantic knowledgebase. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2003)*, volume 2, pages II–327 – II–334.

[Horridge and Bechhofer, 2011] Horridge, M. and Bechhofer, S. (2011). The owl api: A java api for owl ontologies. *Semant. web*, 2(1):11–21.

[Horrocks et al., 2003] Horrocks, I., Patel-Schneider, P. F., and van Harmelen, F. (2003). From SHIQ and RDF to OWL: the making of a Web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1):7–26.

[Hotz and Neumann, 2005] Hotz, L. and Neumann, B. (2005). Scene interpretation as a configuration task. *Künstliche Intelligenz, 3/2005, BöttcherIT Verlag, Bremen*, 3:59–65.

[Hudelot, 2005] Hudelot, C. (2005). *Towards a Cognitive Vision Platform for Semantic Image Interpretation; Application to the Recognition of Biological Organisms*. PhD thesis, Université de Nice Sophia-Antipolis.

[Hudelot et al., 2008] Hudelot, C., Atif, J., and Bloch, I. (2008). Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159:1929–1951.

[Hudelot et al., 2010] Hudelot, C., Atif, J., and Bloch, I. (2010). Integrating bipolar fuzzy mathematical morphology in description logics for spatial reasoning. In *European Conference on Artificial Intelligence (ECAI'10)*, pages 497–502.

[Hudelot et al., 2005] Hudelot, C., Maillot, N., and Thonnat, M. (2005). Symbol grounding for semantic image interpretation: From image data to semantics. In *Proceedings of the Tenth IEEE International Conference on Computer Vision Workshops (ICCVW'05)*, pages 1875–1883, Washington, DC, USA. IEEE Computer Society.

[Hunter, 2001] Hunter, J. (2001). Adding multimedia to the semantic web - building an mpeg-7 ontology. In *Semantic Web Working Symposium (SWWS'01)*, pages 261–281.

[Jaimes and fu Chang, 2000] Jaimes, A. and fu Chang, S. (2000). A conceptual framework for indexing visual information at multiple levels. In *Storage and Retrieval for Image and Video Databases (SPIE'00)*, pages 2–15.

[Jaimes and Smith, 2003] Jaimes, A. and Smith, J. (2003). Semi-automatic, data-driven construction of multimedia ontologies. In *Multimedia and Expo (ICME'03)*, volume 1, pages I–781–4.

[Jain et al., 2005] Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270 – 2285.

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

[Jörgensen, 1996] Jörgensen, C. (1996). Indexing images: Testing an image description template. In *ASIS Annual Conference Proceedings*.

[Jörgensen, 1998] Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management*, 34(2-3):161 – 174.

[Kück et al., 2004] Kück, H., Carbonetto, P., and de Freitas, N. (2004). A constrained semi-supervised learning approach to data association. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*, pages 1–12. Springer.

[Kompatsiaris and Hobson, 2008] Kompatsiaris, Y. and Hobson, P. (2008). *Semantic Multimedia and Ontologies: Theory and Applications*. Springer.

[Kulkarni et al., 2011] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., and Berg, T. (2011). Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1601 –1608.

[Lampert et al., 2008] Lampert, C. H., Blaschko, M. B., and Hofmann, T. (2008). Beyond sliding windows: Object localization by efficient subwindow search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

[Lavrenko et al., 2003] Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A model for learning the semantics of pictures. In *Neural Information Processing Systems (NIPS'03)*. MIT Press.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2169–2178.

[Li and Perona, 2005] Li, F.-F. and 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)*, volume 2, pages 524–531, Washington, DC, USA.

[Li and Wang, 2003] Li, J. and Wang, J. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1075 – 1088.

[Li et al., 2010] Li, L.-J., Wang, C., Lim, Y., Blei, D. M., and Li, F.-F. (2010). Building and using a semantivisual image hierarchy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 3336 –3343.

[Li and Shapiro, 2002] Li, Y. and Shapiro, L. G. (2002). Consistent line clusters for building recognition in cbir. In *International Conference on Pattern Recognition (ICPR'02)*, page 30952, Washington, DC, USA. IEEE Computer Society.

[Liu and Singh, 2004] Liu, H. and Singh, P. (2004). ConceptNet - a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

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

[Logozzo and Cortesi, 2006] Logozzo, F. and Cortesi, A. (2006). Semantic hierarchy refactoring by abstract interpretation. In *international conference on Verification, Model Checking, and Abstract Interpretation (VMCAI'06)*, pages 313–331.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision (ICCV'99)*.

[Maillot et al., 2007] Maillot, N., Chevallet, J.-P., and Lim, J. (2007). Inter-media pseudo-relevance feedback application to imageclef 2006 photo retrieval. In Peters, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D., Rijke, M., and Stempfhuber, M., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 735–738. Springer Berlin Heidelberg.

[Maillot and Thonnat, 2008] Maillot, N. and Thonnat, M. (2008). Ontology based complex object recognition. *Image and Vision Computing*, 26(1):102 – 113.

[Maillot et al., 2004] Maillot, N., Thonnat, M., and Hudelot, C. (2004). Ontology based object learning and recognition: application to image retrieval. In *International Conference on Tools with Artificial Intelligence (ICTAI'2004)*, pages 620 – 625.

[Marr, 1982] Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.

[Marszalek and Schmid, 2007] Marszalek, M. and Schmid, C. (2007). Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1 –7.

[Marszalek and Schmid, 2008] Marszalek, M. and Schmid, C. (2008). Constructing category hierarchies for visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 479–491.

[Martinet et al., 2011] Martinet, J., Chiaramella, Y., and Mulhem, P. (2011). A relational vector space model using an advanced weighting scheme for image retrieval. *Information Processing & Management*, 47(3):391 – 414.

[Mezaris et al., 2003] Mezaris, V., Kompatsiaris, I., and Strintzis, M. G. (2003). An ontology approach to object-based image retrieval. In *International Conference on Image Processing (ICIP'03)*, pages 511–514.

[Mikolajczyk et al., 2004] Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the European Conference on Computer Vision (ECCV'04)*, Lecture Notes in Computer Science, pages 69–82. Springer.

[Müller et al., 2010] Müller, H., Clough, P., Deselaers, T., and Caputo, B., editors (2010). *ImageCLEF experimental evaluation in visual information retrieval*, volume INRE. Springer.

[Müller et al., 2004] Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23.

[Möller and Neumann, 2008] Möller, R. and Neumann, B. (2008). Ontology-based reasoning techniques for multimedia interpretation and retrieval. In Kompatsiaris, Y. and Hobson, P., editors, *Semantic Multimedia and Ontologies*, pages 55–98. Springer London.

[Mohan et al., 2001] Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'01)*, 23(4):349 –361.

[Monay and Gatica-Perez, 2003] Monay, F. and Gatica-Perez, D. (2003). On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, ACM MM'03, pages 275–278.

[Monay and Gatica-Perez, 2004] Monay, F. and Gatica-Perez, D. (2004). PLSA-based image auto-annotation: constraining the latent space. In *Proceedings of the 12th annual ACM international conference on Multimedia (ACM MM'04)*, pages 348–351.

[Monay and Gatica-Perez, 2007] Monay, F. and Gatica-Perez, D. (2007). Modeling semantic aspects for cross-media image indexing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1802 –1817.

[Mori et al., 1999] Mori, Y., Takahashi, H., and Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*.

[Mylonas et al., 2009] Mylonas, P., Spyrou, E., Avrithis, Y., and Kollias, S. (2009). Using visual context and region semantics for high-level concept detection. *IEEE Transaction on MultiMedia*, 11(2):229–243.

[Naphade et al., 2006] Naphade, M., Smith, J. R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., and Curtis, J. (2006). Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13:86–91.

[Neumann and Möller, 2008] Neumann, B. and Möller, R. (2008). On scene interpretation with description logics. *Image Vision Computing*, 26(1):82–101.

[Oliva and Torralba, 2007] Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527.

[Osuna et al., 1997] Osuna, E., Freund, R., and Girosit, F. (1997). Training support vector machines: an application to face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 130 –136.

[Panofsky, 1972] Panofsky, E. (1972). *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. An Icon Edition Series. Icon.

[Papageorgiou and Poggio, 1999] Papageorgiou, C. and Poggio, T. (1999). Trainable pedestrian detection. In *(ICIP 99) Image Processing, Proceedings of the InternationalConference on*, volume 4, pages 35 –39.

[Papageorgiou and Poggio, 2000] Papageorgiou, C. and Poggio, T. (2000). A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33.

[Patwardhan and Pedersen, 2006] Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together (EACL'06)*.

[Peraldi et al., 2007] Peraldi, S. E., Kaya, A., Melzer, S., Möller, R., and Wessel, M. (2007). Multimedia interpretation as abduction. In *International Workshop on Description Logics (DL'07)*.

[Petridis et al., 2006] Petridis, K., Anastasopoulos, D., Saathoff, C., Kompatsiaris, Y., and Staab, S. (2006). MOntoMat-Annotizer: image annotation linking ontologies and multimedia low-level features. In *International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'06)*.

[Pixable Blog, 2011] Pixable Blog (2011). How much do you know about Facebook Photos? http://blog.pixable.com/2011/02/14/facebook-photo-trends-infographic/. [Online; accessed 10-December-2012].

[Platt et al., 2000] Platt, J. C., Cristianini, N., and Shawe-taylor, J. (2000). Large margin DAG for multiclass classification. In *Advances in Neural Information Processing Systems (NIPS'10)*.

[Popescu and Grefenstette, 2011] Popescu, A. and Grefenstette, G. (2011). Social media driven image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR'11)*, pages 33:1–33:8. ACM.

[Popescu et al., 2007] Popescu, A., Moëllic, P.-A., and Millet, C. (2007). Semretriev: an ontology driven image retrieval system. In *Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR'07)*, pages 113–116.

[Pujol et al., 2006] Pujol, O., Radeva, P., and Vitria, J. (2006). Discriminant ECOC: a heuristic method for application dependent design of error correcting output codes. *Pattern Analysis and Machine Intelligence (TPAMI'06), IEEE Transactions on*, 28(6):1007 –1012.

[Rabinovich et al., 2007] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2007). Objects in context. In *Proceedings of the International Conference on Computer Vision (ICCV'07)*.

[Rasiwasia et al., 2010] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia (MM'10)*, pages 251–260. ACM.

[Reimer et al., 2011] Reimer, U., Maier, E., Streit, S., Diggelmann, T., and Hoffleisch, M. (2011). Learning a lightweight ontology for semantic retrieval in patient-centered information systems. *IJKM*, 7(3):11–26.

[Ren and Cheng, 2008] Ren, Y. and Cheng, X. (2008). Semantic-based image retrieval using fuzzy domain ontology. *Intelligent Information Technology Applications, 2007 Workshop on*, 2:141–145.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial intelligence (IJCAI'95)*, pages 448–453.

[Rohrbach et al., 2010] Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., and Schiele, B. (2010). What helps? where? and why? semantic relatedness for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 910 –917.

[Romdhane et al., 2010] Romdhane, L. B., Bannour, H., and el Ayeb, B. (2010). IMIOL: a system for indexing images by their semantic content based on possibilistic fuzzy clustering and adaptive resonance theory neural networks learning. *Applied Artificial Intelligence*, 24(9):821–846.

[Rui et al., 1998] Rui, Y., Huang, T. S., and Mehrotra, S. (1998). Relevance feedback techniques in interactive content-based image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36.

[Russell et al., 2007] Russell, B. C., Torralba, A., Liu, C., Fergus, R., and Freeman, W. T. (2007). Object recognition by scene alignment. In *Advances in Neural Information Processing Systems (NIPS'07)*.

[Russell et al., 2008] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173.

[Santini et al., 2001] Santini, S., Gupta, A., and Jain, R. (2001). Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering*, 13(3):337–351.

[Schapire et al., 1998] Schapire, R., Freund, Y., Bartlett, P., and Lee, W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

[Sethi and Coman, 2001] Sethi, I. K. and Coman, I. L. (2001). Mining Association Rules Between Low-Level Image Features and High-Level Concepts. *Proceedings of the SPIE Data Mining and Knowledge Discovery*, page 279Ű290.

[Shanahan, 2005] Shanahan, M. (2005). Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science*, 29:103–134.

[Shatford, 1986] Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, 6(3):39–62.

[Shatford, 1994] Shatford, S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science*, 45(8):583–588.

[Shen and Fan, 2010] Shen, Y. and Fan, J. (2010). Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *Proceedings of the international conference on Multimedia (MM'10)*, pages 5–14.

[Simou et al., 2008] Simou, N., Athanasiadis, T., Stoilos, G., and Kollias, S. (2008). Image indexing and retrieval using expressive fuzzy description logics. *Signal, Image and Video Processing*, 2(4):321–335.

[Simou et al., 2005] Simou, N., Tzouvaras, V., Avrithis, Y., Stamou, G., and Kollias, S. (2005). A visual descriptor ontology for multimedia reasoning. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05)*.

[Sivic et al., 2008] Sivic, J., Russell, B. C., Zisserman, A., Freeman, W. T., and Efros, A. A. (2008). Unsupervised discovery of visual object class hierarchies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

[Smeulders et al., 2000] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380.

[Smith and Chang, 1997] Smith, J. and Chang, S.-F. (1997). Visually searching the web for content. *MultiMedia, IEEE*, 4(3):12 –20.

[Smith et al., 2003] Smith, J., Naphade, M., and Natsev, A. (2003). Multimedia semantic indexing using model vectors. In *International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages II – 445–8 vol.2.

[Snoek et al., 2007] Snoek, C. G., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., and Worring, M. (2007). Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia (TMM'07)*, 9(5):975–986.

[Snoek et al., 2006] Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, ACM MM'06, pages 421–430.

[Solomatine and Shrestha, 2009] Solomatine, D. P. and Shrestha, D. L. (2009). A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research*, 45.

[Spaccapietra et al., 2004] Spaccapietra, S., Cullot, N., Parent, C., and Vangenot, C. (2004). On spatial ontologies. In *Brazilian Symposium On Geoinformatics*.

[Stoilos and Stamou, 2007] Stoilos, G. and Stamou, G. B. (2007). Extending fuzzy description logics for the semantic web. In *Workshop on OWL: Experiences and Directions (OWLED'07)*.

[Straccia, 2001] Straccia, U. (2001). Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research*, 14:137–166.

[Straccia, 2006] Straccia, U. (2006). A fuzzy description logic for the semantic web. In Sanchez, E., editor, *Fuzzy Logic and the Semantic Web*, volume 1 of *Capturing Intelligence*, pages 73 – 90. Elsevier.

[Straccia, 2009] Straccia, U. (2009). Towards spatial reasoning in fuzzy description logics. In *international conference on Fuzzy Systems (FUZZ-IEEE'09)*, pages 512–517.

[Straccia, 2010] Straccia, U. (2010). An ontology mediated multimedia information retrieval system. In *Multiple-Valued Logic (ISMVL'10)*, pages 319–324.

[Straccia, 2012] Straccia, U. (2012). Description logics with fuzzy concrete domains. *Computing Research Repository (CoRR)*, abs/1207.1410.

[Szummer and Picard, 1998] Szummer, M. and Picard, R. W. (1998). Indoor-Outdoor image classification. In *International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, page 42, Washington, DC, USA. IEEE Computer Society.

[Tang et al., 2011] Tang, J., Hong, R., Yan, S., Chua, T.-S., Qi, G.-J., and Jain, R. (2011). Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology*, 2(2):14:1–14:15.

[Tollari, 2006] Tollari, S. (2006). *Indexation et recherche d'images par fusion d'informations textuelles et visuelles (Image indexing and retrieval by combining textual and visual informations)*. PhD thesis, Université du Sud Toulon-Var.

[Tommasi et al., 2008] Tommasi, T., Orabona, F., and Caputo, B. (2008). Discriminative cue integration for medical image annotation. *Pattern Recognition Letters*, 29(15):1996 – 2002.

[Tong and Chang, 2001] Tong, S. and Chang, E. (2001). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia (MM'01)*, pages 107–118. ACM.

[Torralba, 2003] Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191.

[Torralba et al., 2008] Torralba, A., Fergus, R., and Freeman, W. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958 –1970.

[Torralba et al., 2004] Torralba, A., Murphy, K., and Freeman, W. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages II–762 – II–769.

[Torralba et al., 2007] Torralba, A., Murphy, K., and Freeman, W. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'07)*, 29(5):854 –869.

[Torralba et al., 2010] Torralba, A., Murphy, K. P., and Freeman, W. T. (2010). Using the forest to see the trees: exploiting context for visual object detection and localization. *Communications of the ACM*, 53(3):107–114.

[Tousch et al., 2008] Tousch, A.-M., Herbin, S., and Audibert, J.-Y. (2008). Semantic lattices for multiple annotation of images. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval (MIR'08)*, pages 342–349. ACM.

[Tousch et al., 2012] Tousch, A.-M., Herbin, S., and Audibert, J.-Y. (2012). Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345.

[Town, 2006] Town, C. (2006). Ontological inference for image and video analysis. *Machine Vision and Applications*, 17(2):94–115.

[Treisman and Gelade, 1980] Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136.

[Vailaya et al., 2001] Vailaya, A., Figueiredo, M. A. T., Jain, A. K., and Zhang, H.-J. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130.

[Vailaya et al., 1998] Vailaya, A., Jain, A., and Zhang, H. J. (1998). On image classification: City vs. landscape. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '98)*, page 3.

[Vehkalahti, 2008] Vehkalahti, K. (2008). The concise encyclopedia of statistics by yadolah dodge. *International Statistical Review*, 76(3):460–461.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages I–511 – I–518.

[Wang et al., 2006] Wang, M., Hua, X.-S., Song, Y., Yuan, X., Li, S., and Zhang, H.-J. (2006). Automatic video annotation by semi-supervised learning with kernel density estimation. In *Proceedings of the 14th annual ACM international conference on Multimedia (MM'06)*, pages 967–976.

[Wei and Ngo, 2007] Wei, X.-Y. and Ngo, C.-W. (2007). Ontology-enriched semantic space for video search. In *International Conference on Multimedia (MM'07)*, pages 981–990.

[Wei and Tao, 2010] Wei, Y. and Tao, L. (2010). Efficient histogram-based sliding window. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 3003–3010. IEEE.

[Wikipedia-Iconography, 2012] Wikipedia-Iconography (2012). Iconography in art history. http://en.wikipedia.org/wiki/Iconography. [Online; accessed 15-December-2012].

[Wikipedia-Opposite_Semantics, 2012] Wikipedia-Opposite_Semantics (2012). Opposite semantics. http://en.wikipedia.org/wiki/Opposite_(semantics). [Online; accessed 15-December-2012].

[Wu et al., 2008] Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., and Li, S. (2008). Flickr distance. In *International Conference on Multimedia (MM'08)*.

[Wu et al., 2012] Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., and Li, S. (2012). Flickr distance: A relationship measure for visual concepts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):863 –875.

[Xiao et al., 2010] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 3485–3492. IEEE.

[Yao et al., 2010] Yao, B., Yang, X., Lin, L., Lee, M. W., and Zhu, S.-C. (2010). I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485 –1508.

[Zhang et al., 2012] Zhang, D., Islam, M. M., and Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346 – 362.

[Zhang et al., 2002] Zhang, Q., Goldman, S. A., Yu, W., and Fritts, J. (2002). Content-based image retrieval using multiple-instance learning. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML'02)*, pages 682–689.

[Zhou et al., 2010] Zhou, X., Yu, K., Zhang, T., and Huang, T. (2010). Image classification using super-vector coding of local image descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*.

[Zhou et al., 2007] Zhou, Z.-H., Zhan, D.-C., and Yang, Q. (2007). Semi-supervised learning with very few labeled training examples. In *Proceedings of the 22nd national conference on Artificial intelligence (AAAI'07)*, pages 675–680.

[Zhu, 2006] Zhu, X. (2006). Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, Department of Computer Science.

[Znaidia et al., 2012] Znaidia, A., Shabou, A., Popescu, A., le Borgne, H., and Hudelot, C. (2012). Multimodal feature generation framework for semantic image classification. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 38:1–38:8, New York, NY, USA. ACM.

[Zweig and Weinshall, 2007] Zweig, A. and Weinshall, D. (2007). Exploiting object hierarchy: Combining models from different category levels. In *Proceedings of the International Conference on Computer Vision (ICCV'07)*, pages 1 –8.