Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Semantic hierarchies for image annotation: A survey

Anne-Marie Tousch [a,b], Stéphane Herbin [a,*], Jean-Yves Audibert [b,c]

[a] ONERA - The French Aerospace Lab, Palaiseau, France
[b] Université Paris-Est/Ecole des Ponts ParisTech/IMAGINE, Marne-La-Vallée, France
[c] SIERRA - ENS/INRIA, Paris, France

ARTICLE INFO

Article history:
Received 25 April 2010
Received in revised form
15 April 2011
Accepted 28 May 2011

Keywords:
Image annotation
Semantic description
Structured vocabulary
Image retrieval

ABSTRACT

In this survey, we argue that using structured vocabularies is capital to the success of image annotation. We analyze literature on image annotation uses and user needs, and we stress the need for automatic annotation. We briefly expose the difficulties posed to machines for this task and how it relates to controlled vocabularies. We survey contributions in the field showing how structures are introduced. First we present studies that use unstructured vocabulary, focusing on those introducing links between categories or between features. Then we review work using structured vocabularies as an input and analyze how the structure is exploited.

1. Introduction

Digital images are ubiquitous in modern life. Professional databases are used by journalists or in the advertising industry; video surveillance generates tera bytes of data every day; remote sensing images are integrated in user friendly environments. The evolution of the Internet and communication facilities has given access to huge mass of data to a general public eager to share experience and information on so-called *Web 2.0* applications.

To become manageable and to meet scalability requirements, images are usually complemented by extra formal representations called *metadata* which serves as an informative index or even as a substitute for the data itself. Metadata may contain various types of information: date, location, symbolic description, physical properties, ... It can be expressed as a free-text, or in a more constrained format.

Fig. 1 gives an example of a textual metadata coming with a journalistic picture from Reuters Press Agency. This metadata contains various types of information expected to be useful to a journalist: the sports event, its date and location, the people involved, the circumstances, etc. One can already discriminate two categories of information in this short text: one that can be deduced from the image itself using background interpretation knowledge, and one that cannot. In this survey, we are interested in the first category, i.e. information that can be inferred from the sole image, from its *content*, independently from all other contextual clues or sources.

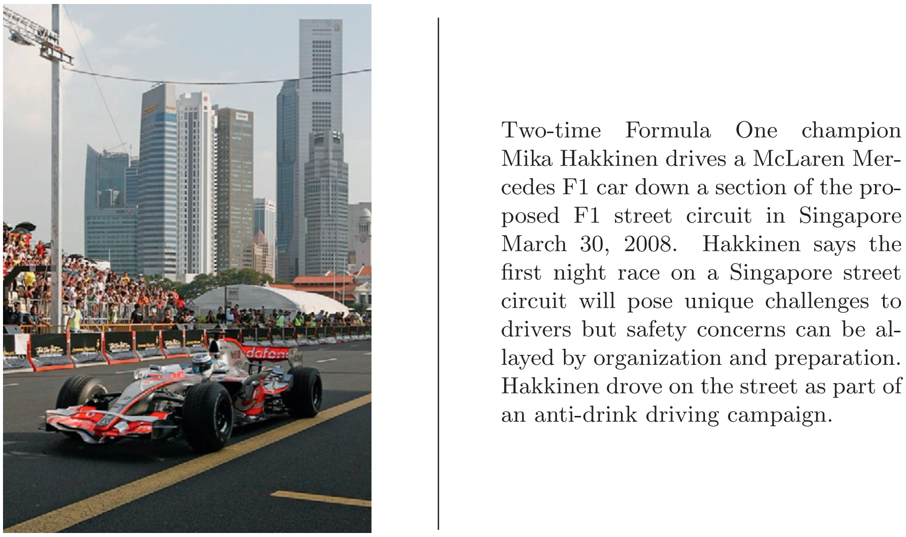
Designing algorithmic procedures to annotate images has been the subject of much research. The original, and still very active, trend of studies has addressed the annotation as an arrangement of processing modules such as feature extraction, region segmentation, saliency detection, pattern recognition, etc. It is concerned mainly by object or people detection and recognition functions and assesses the reliability of the processing chains on limited data samples. In the last decades, the availability of large multimedia databases has brought new issues such as image indexing and retrieval, shifting the performance objectives to the mastering of large amount of data but sometimes with lower requirements.

More recently, the development and use of data sharing applications has motivated an increasing interest in semantic representations. Indeed, since effective communication relies on shared languages and practices, a logical evolution is to embed representations in a semantic structure in order to make them understandable by humans or processed by computers. If we go back to Fig. 1, the attached annotation, although concise, has already a rather complex structure. It mixes various levels of description according to different points of view or facets. Moreover, in certain circumstances, the higher precision level might be superfluous, and it might be sufficient to know that the image represents a Formula One event, or even more simply, a kind of sport.

Fig. 2 is an attempt to capture the trend of research with respect to semantic image annotations. While there is no unified vocabulary to express the idea of hierarchical and multi-faceted

* Corresponding author. Tel.: +33 1 80 38 65 69.
E-mail addresses: amtousch@cvdm-solutions.com (A.-M. Tousch), stephane.herbin@onera.fr (S. Herbin), audibert@imagine.enpc.fr (J.-Y. Audibert).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.05.017



Two-time Formula One champion Mika Hakkinen drives a McLaren Mercedes F1 car down a section of the proposed F1 street circuit in Singapore March 30, 2008. Hakkinen says the first night race on a Singapore street circuit will pose unique challenges to drivers but safety concerns can be allayed by organization and preparation. Hakkinen drove on the street as part of an anti-drink driving campaign.

Fig. 1. Annotation example from Reuters pictures (credits: Vivek Prakash/Reuters).

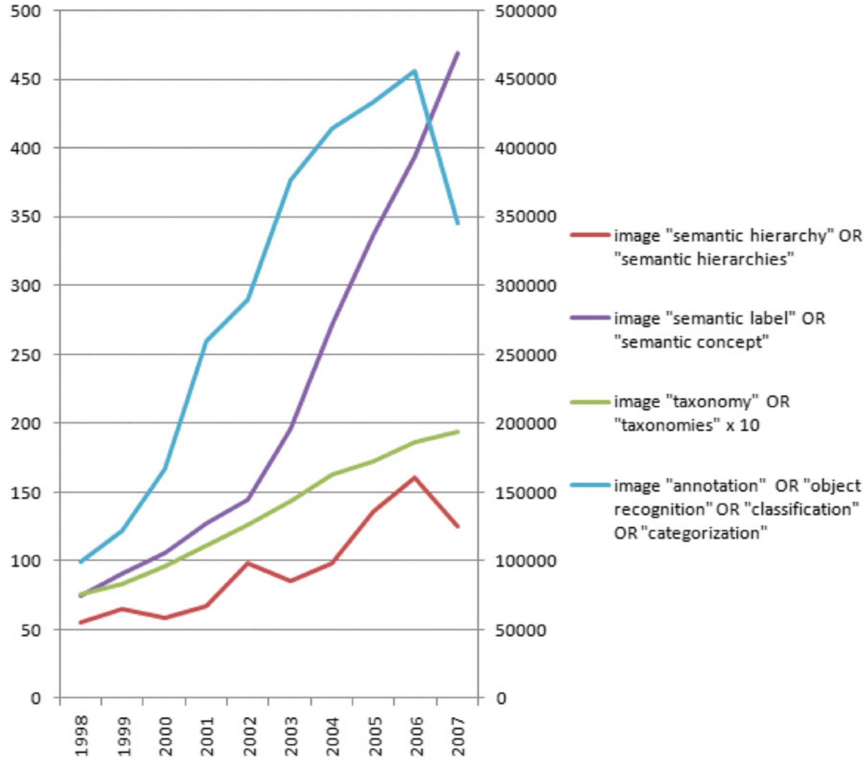


Fig. 2. Research trends with respect to semantic image annotation, captured by Google Scholar results. The curves give the number of articles returned by Google Scholar when searching for the mentioned requests, scaled to fit in a same figure. The first two (red and violet) correspond to the left axis, and the other (green and blue) correspond to the right axis. The green curve, corresponding to image "taxonomies", has been multiplied by a factor 10, again for readability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

annotation or description, results in a Google Scholar search show an increase in articles mentioning taxonomies or semantic hierarchies in the context of images.

This article intends to present the state of the art of the techniques aiming at producing rich image content description using shared vocabularies. A specific effort will be given to show how multiple levels of precision and richer semantics have been touched upon. Although clearly connected, object recognition and image retrieval technical aspects are not the subject of this survey. Interested reader should refer to surveys by Datta et al. [21] and Liu et al. [64].

Automatic annotation is a complex task. As shown in the examples, the choice of the words being used is highly dependent on the application, user needs, and user context. These different possible descriptions for an image are the subject of Section 2. We will study the different annotations that can be applied to an image, and the role of the user in the selection of keywords. Then we will explain the technical difficulties, epitomized by the

Please cite this article as: A.-M. Tousch, et al., Semantic hierarchies for image annotation: A survey, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.05.017

so-called *semantic gap*, in Section 3.1. We will also make a short presentation of the different attempts being made toward bridging this gap through object recognition techniques, and of the vocabulary structures that could be used with it—i.e. the kinds of existing semantic hierarchies.

Eventually, we will review significant works in both object/scene recognition and image retrieval: first, those that do not use semantic structures as an input in Section 4, and second, those that do in Section 5.

2. The nature and use of semantics in image description

The metadata given in Fig. 1 is probably the highest level of annotation one can imagine. The image is placed in its initial context. Given only the description, one can guess quite easily the composition of the picture: a Formula-1 car in a street, buildings in the background, barriers delineating the circuit, and people. Some information can be extracted from the data, but there are also non-visual information contained in the description. For instance, it informs us that (i) Mika Hakkinen has been F1 world champion twice, (ii) the race he takes part in is part of an anti-drink driving campaign, (iii) the photo has been taken on March 30, 2008. This information cannot be inferred from the picture data alone.

We will follow the analysis of Shatford Layne [86] about metadata typology which divides it into four categories:

1. Biographical attributes, i.e. concerning the creation of the image (location, time, …),
2. Subject attributes, that describe the content of the image,
3. Exemplified attributes, i.e. the kind of illustration that the image is (photograph, cartoon, sketch, …),
4. Relationship attributes, which are links to related images (e.g. a painting could be linked to a corresponding sketch).

The present survey only addresses the problem of content description, i.e. subject attributes if we take the above terminology.

2.1. Semantic levels

As the example of Fig. 1 demonstrates, metadata as such is made of different kinds of information. Our focus is on image *content*, for which the image is the only source of information. In the theory of Shatford Layne, there are several aspects of content description:

1. The *Of-ness* vs. *About-ness*, i.e. the objective and concrete description vs. the subjective and abstract one. For instance, if "someone crying" is an objective description, "pain" is a subjective one.
2. The description can be *generic*, as in "a bridge", or *specific*, as in "Brooklyn bridge".
3. It can have four facets: time, localization, event, or object.

Jörgensen [55] extends this idea, and shows that an image can be described from different facets and that in image search, an image should be reachable from a number of entry points, rather than a unique one. Formally, we can say that images and their concepts are not linked together by a single hierarchy, but multiple. Moreover, following the distinction made by Shatford Layne [86], she separates between *perceptual* attributes, i.e. objective ones, and *interpretive* attributes that depend on (necessarily subjective) interpretation. She adds *reactive* attributes that describe the personal reactions of a person seeing the image.

Enser and Sandom [28] adapt from Jörgensen [55] using *perceptual*, *generic-interpretive*, *specific-interpretive* and *abstract* levels of description.

Jaimes and Chang [51] offer to structure the content of the image using 10 levels between visual image features and abstract interpretation, shown in Table 1. The first four levels refer to the perceptual aspects or syntax. The last six levels correspond to semantics or visual concepts. The authors point out that the distinction between the levels need not be strict; it is rather an aid to understanding the issue. The higher the level, the more knowledge is involved in the interpretation. The four syntactic levels are fully objective and purely numerical description of the image. The semantic levels can be compared with Shatford Layne's and Jörgensen's distinctions between *generic*, *specific* and *abstract* levels, together with a distinction between the description of objects (or local components, e.g. a F1, skyscrapers, a street, …) and the description of the scene (in Fig. 1, e.g. a F1 race). An example is given in Fig. 3. To the best of our knowledge, today's algorithms do not exceed level 7 in automatic recognition.

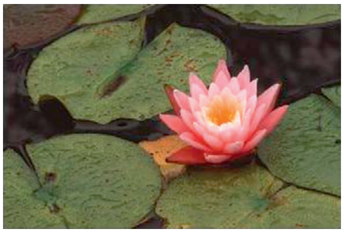
Hollink et al. [48] take the same levels as Jaimes and Chang [51] and use Unified Modeling Language (UML) in an attempt to transcribe the annotation levels formally. This way they can easily describe a scene as a composition of objects, and even an object can be a description of objects, enabling recursion. Any of these (scene or object) can be described with three levels of genericness/specificity/abstraction, and a description may stem from any of the facets: space, time, event or object, as in Shatford Layne [86].

Hare et al. [46] propose a gradation similar to Jaimes and Chang [51] from the raw image to high-level semantics: raw image; visual descriptors; objects (i.e. segmentation); object names; semantics (i.e. meaning at the scene level).

Eakins et al. [26] show different aspects of abstraction, which somehow can be related to the scene's generic, specific, and abstract levels of Jaimes and Chang [51]. *Contextual* abstraction depends on the knowledge of the environment, which can be quite basic (or generic). *Cultural* abstraction refers to interpretation using a specific cultural knowledge (e.g. to understand the meaning of a religious ceremonial). *Emotional* abstraction refers to interpretation strongly influenced by the personal background of the viewer (related to Jörgensen's *reactive* attribute). Finally,

Table 1
The description levels suggested by Jaimes and Chang [51].

1	Type, technique
2	Global distribution
3	Local structure
4	Global composition
5	Generic objects
6	Generic scene
7	Specific objects
8	Specific scene
9	Abstract objects
10	Abstract scene



1.	photo
2.	(histogram)
4.	(segmentation)
5.	flower, leaf, water
6.	nature
7.	water lily and its leaves
8.	pond
10.	stillness, coldness

Fig. 3. Example of annotations at different levels.

to understand *technical* abstraction one needs expert knowledge of the domain, e.g. to interpret X-rays.

As a further example, the content of the image in Fig. 1 can be described at several levels depending on the viewer's background knowledge: an outdoor scene; a car, a street, buildings, people, sky; a town; a car race, a business district; a Formula One, ….

Several hierarchical relations can be introduced in these descriptions. We introduce the following notations to describe them. Let $\prec$ be the symbol for an Is-A relationship between two categories. $A \prec B$ means that A is a B, i.e. any object in category A is also in category B. Let $\sqsubset$ represent the Part-Of relation: $A \sqsubset B$ if A is a part of B, and $\wedge$ be the co-occurrence relation: $A \wedge B$ when A and B are found in a same image. For the example of Fig. 1, we then have

- McLaren Mercedes Formula One $\prec$ F1 $\prec$ sports car $\prec$ car,
- Modern town with skyscrapers $\prec$ town,
- Sport car $\wedge$ spectators $\sqsubset$ car race,
- Buildings $\wedge$ street $\sqsubset$ town.

Describing the content of an image has been the subject of studies from archivists, among others. Even when annotated by humans, what to annotate, and how, is not straightforward. We have seen that an image can be described at several levels of genericness, or semantic precision, and following several description paths, or facets. It depends on the user's objective what description level will prevail. In other words, what is the description level most likely to be interesting? We will now review several user studies and try to give insight into this question in the following section.

2.2. The user, his goal, his context

So far we have seen that images can be interpreted at several levels of semantic description (genericness/specificity), and of abstraction (About-ness/Of-ness or subjective/objective). Which level is meaningful depends on the user's context.

As Hollink et al. [48] point it out, a first context feature is the application domain. What is the database to be searched? Does it cover a broad domain? Does it handle a large vocabulary?

The user level of expertize is also a key factor. Enser and Sandom [28] make a distinction between generalists and specialists. For Hollink et al. [48], there are various degrees between the two, and it depends on the domain. Jaimes [50] underlines the connection between the aim of the user and the form of the query: a user searching for something specific will choose specific keywords, whereas if he has only a vague idea, he will prefer browsing (following the organization of the database, i.e. choosing categories from a menu, and viewing thumbnails). This is the conclusion of several studies, such as Markkula and Sormunen [67] and Frost et al. [38]. Markkula and Sormunen [67] observed that journalists searching for generic concepts in journal archives had difficulties with keyword search and preferred browsing. The keyword annotations given by archivists were better fitted for specific queries. Frost et al. [38] noted that users that were not well acquainted with the database content also preferred browsing.

The next aspect that depends on the user is the search mode. Indeed, the way one searches for images is highly dependent on its context. Hollink et al. [48] lists several modes of image search: browsing, keywords-based, using keywords with logical operators, free-text, example-based, or sketch-based. Eakins [25] gives different types of queries, for text only:

1. Those using image *primitives* (such as "find images containing yellow stars in a circle"),
2. Those using *logical* attributes, requiring a minimum of image interpretation (as in "find images of a train passing on a bridge"),
3. Those using *abstract* attributes, demanding complex reasoning on the content of images (e.g. "find images about freedom").

Most of the available software is searching for content at level 1. Yet studies show that most of the users queries are at level 2, and a lot of them are at level 3 [5,25].

Jaimes [50] further describes user behaviors. Given a search method, the user can adopt different search behaviors, such as: using exploration to make the query more and more specific; using intuition; being more purposive, having something specific in mind; etc.

Finally, the user will be guided by its task. Hollink et al. [48] describe it as a spectrum between a *data pole* and an *object pole*, where the data pole is for informative images and the object pole for the decorative aspects of images. Eakins et al. [26] describe seven types of tasks. We show them in a different order compared to their article, putting them in an "informative" order between the data pole and the object pole:

information processing where the data itself is of primary importance (e.g. X-rays),
information dissemination where the information has to be transmitted to someone else (e.g. mug-shots to police),
illustration where images are accompanied by another media (e.g. text for news images),
generation of ideas where images are the starting point of a creative process (e.g. architecture),
learning where the image is used to acquire knowledge (as in art or history),
emotive where the reaction to the image is prevalent (e.g. advertising),
aesthetic value is found at the object pole where images are used for decoration.

For all the different uses of images, we can see that only a few of them are related to low-level image features (even aesthetic value might not always be fully described by those and is partly subjective). Thus, having the possibility to infer semantic concepts from the image seems crucial to organize, index and search image databases. This is where annotation is essential. Moreover, it stems from these studies that annotation should be multi-faceted (i.e. from different perspectives/views) and multi-level, so as to address as much user needs as possible.

3. About semantic analysis

3.1. The semantic gap

In most cases, automatic annotation follows two steps: (1) extraction of informative low-level visual features; (2) interpretation of these features into high-level concepts. In the literature, the problem arising from trying to associate low-level features (i.e. numerical data) to high-level concepts (i.e. semantic metadata) is called the *semantic gap*. While bridging this semantic gap is natural for a human being, it is far from being obvious to the machine. Smeulders et al. [88] give the following definition: *the semantic gap is 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.*

Visual recognition in the human brain consists in linking the image printed on the retina with a representation stored in memory [97]. The first stages of visual processing have been

Please cite this article as: A.-M. Tousch, et al., Semantic hierarchies for image annotation: A survey, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.05.017

studied extensively and are now claimed to be well understood. How recognition occurs is less clear, as well as how objects and concepts as expressed by language are stored in memory. Therefore, imitation is unattainable. The human brain is able to recognize almost instantly a huge number of objects, and to put them into language requires little effort for a human being. To date, such incredible possibilities could not be reached by machines.

The notion of *semantics* has been widely used by researchers in the image processing community to designate automatic processes manipulating natural language at some stage, as opposed to meaningless numerical data. Most of them are not interested in the philosophical debate about the nature of *concepts*, or the relation between *knowledge* and *meaning*. Neither will we get into such a debate. Rather we will follow the customary and shallow use of the word, i.e. we will refer to *semantics* when words of the natural language are used to describe an image. *Semantic analysis* will refer to any kind of transcription of an image into a linguistic expression.

The problem of semantic analysis has been addressed by several communities, with different approaches and influences. People from *artificial intelligence* will rely on the use of knowledge bases, with sophisticated vocabularies structured by ontologies. In the *computer vision* community, statistical methods are widely used, using quite simple vocabularies (they can be large, but are generally not structured). In *indexation* or *image retrieval* issues, user input is taken into account more, and studies mix inspirations from linguistics and statistical methods. In the next subsection, we explain more about the different structures in vocabularies that can be met in image semantic analysis together with the basic problems tackled by object recognition.

3.2. Object recognition and controlled vocabularies

Apart from the fact that deciding the level of semantic interpretation might not be obvious, object recognition in itself is a complex problem, for several reasons:

- The world is complex: it contains a lot of different objects. For instance, one can easily recognize thousands of categories of objects, tens of thousands if counting types of objects [12].
- Visually, inter-category variations can be very small (e.g. a bird and a plane, seen from a distance), whereas intra-category variations may be high (especially for man-made objects, as for instance different kinds of chairs, but also for natural objects, as for instance different butterflies).
- The conditions under which images are taken are unpredictable and strongly affect appearance. Indeed, the same object seen from different points of view can change a lot. Lighting conditions, background, occlusions all affect appearance without necessarily influencing image interpretation.
- A single object can often change appearance. For instance, a human face can change expression, and an animal can change position.

Thus, even without considering complex semantics, machine recognition of simple objects is difficult. The problem of object recognition could be solved in special cases, for specific industrial applications for instance. Recognition in more general cases is still an issue.

Studies in cognitive psychology show that category recognition occurs first at a fundamental category level called *basic level* (see [82,73,54] for more details). Recognition of super-categories and sub-categories happens next, using more detailed analysis of visual appearance, especially looking details dependent of the category recognized first.

It is therefore logical to test recognition at this basic level before making more specific assumptions. In object categorization, an image is associated to a unique term describing the object contained in the image, using a vocabulary with fixed size K, $\mathcal{V}=\{w_1,\dots,w_K\}$, that is, a list of K categories that can be recognized.

A *controlled vocabulary* is a vocabulary of reference, with fixed size, that is used for indexing. It can have a structure, such as a thesaurus, a taxonomy or an ontology. Gilchrist [42] gives an insight into these three kinds of structured vocabularies, starting from their definition and explaining their use among scientists of different communities. Garshol [40] gives more detailed definitions that we reuse in the following.

The term *taxonomy* originally referred to the tree structure used for the classification of species in biological science. In computer science, the term is used for about any kind of hierarchy between objects. Typically, it is used for Is-A hierarchies, i.e. subtype/supertype relationships, also called hyponymy/hypernymy. For instance, *car* is an hyponym of *vehicle*, and *vehicle* is the hypernym of *car*.

There are contradictions between authors concerning the difference between taxonomy and thesaurus, which shows the vague character of the usage of these terms. As a matter of fact, the two structures come from two separate domains, namely biology and document indexing, and were used quite independently.

We will define a *thesaurus* as an extension of a taxonomy. A thesaurus, apart from describing hyponymy/hypernymy relations, also links together words that are synonyms, giving explanation about word usage, word equivalence and preferences. Also, it will give *related* terms, i.e. words that belong to the same domain, and often there will be a short definition of the meaning of the term as used in the vocabulary.

An *ontology* is a model for the formal description of concepts [40]. It is defined by a set of types (of concepts), of properties and of relationship types between concepts (or objects). The formal model should be machine-readable. The relationship types are more diverse than for a thesaurus and could in theory be labeled by any type.

We will also use *semantic networks* as an intermediate between thesauri and ontologies, describing more relationships than thesauri but less formal than ontologies. Fig. 4 shows a semantic network representation with several types of relations that are commonly used in the vision community. Two kinds of nodes are possible: *concept nodes* represent categories of objects (such as "car"), and *instance nodes* represent instances of objects, i.e. occurrences (such as "Jimmy's car"). In Fig. 4, we represent concept nodes differently for categories and parts, for better clarity. In the following, our interest lies in concepts and not in instances, i.e. relations of type Is-A-Kind-Of are not part of our study.

We will use the term "semantic hierarchy" with the more general meaning of "semantic structure", assuming there are hierarchical relations between concepts (but not restricted to trees).

In the context of semantic analysis, so far, object recognition has been focused on using simple flat controlled vocabularies, i.e. unstructured, and often of small size, as we will see later. Different problems were addressed:

- Object detection consists in deciding whether a particular object category is present or not in an image,
- Localization aims at giving the position and scale of a particular object, and is often associated with detection,
- Categorization consists in assigning one global label to the image, chosen in a fixed-size list of labels,
- Identification is used either for the categorization of objects in sub-categories, either for the recognition of a particular instance of an object (e.g. "Jane's car"),
- Annotation assigns a number of labels to the image, selected from a fixed vocabulary,

Please cite this article as: A.-M. Tousch, et al., Semantic hierarchies for image annotation: A survey, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.05.017

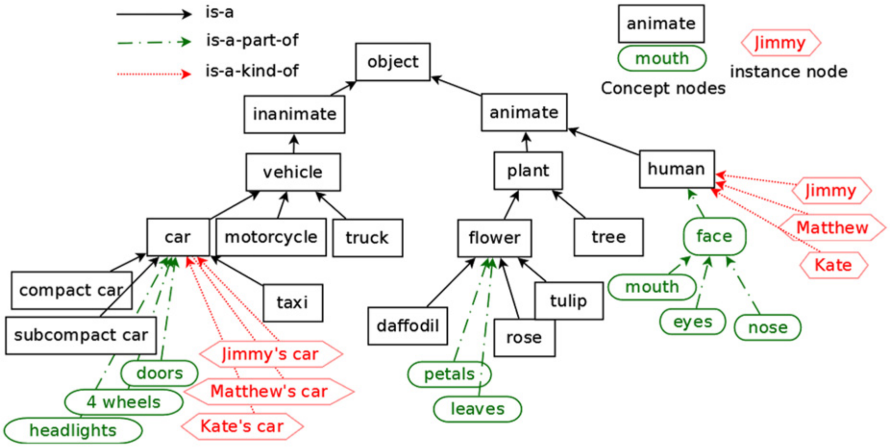


Fig. 4. Example of a semantic network with relations between concepts usually met in object recognition.

- Correspondence, also called region annotation, associates terms to image regions (similar to annotation of regions).

All these problems are related: for instance, correspondence is used for categorization (especially for scenes), and categorization for annotation. Typically, identification is done after categorization. Solving these problems relies heavily on the use of image processing techniques and machine learning.

The problem of semantic analysis, and of the bridging of the semantic gap, has been extensively studied in the past decades. Indeed, the whole field of computer vision can be seen as trying to give an image understanding ability to machines. Object recognition has been a central objective since the beginning: starting from single instance recognition, to object categorization and global scene description. Reviews of the state of the art in object recognition can be found in Mundy [72], for a historical perspective, Pinz [77] for a lengthy review of techniques used in object categorization, and Bosch et al. [14] for scene recognition.

The newer image retrieval and multimedia database management issues have fostered the use of a semantic level of analysis. Several reference papers at the end of the last decade review works and techniques in content-based image retrieval [84,88]. Eakins [25] speaks in favor of more artificial intelligence in the domain. Recent reviews are given by Liu et al. [64], which is of particular interest since it advocates the introduction of semantic aspects in image retrieval, and Datta et al. [21], a very extensive review presenting the different issues and trends in this area. Understanding the image content, or bridging the semantic gap, is now a problem commonly raised in the image retrieval community.

In the following, we distinguish between two kinds of approaches used for the semantic analysis of images: (a) classical approaches looking for a mapping between image numerical data and a flat vocabulary; (b) approaches exploiting a structured vocabulary known beforehand.

In the next section, we will focus on methods used for semantic analysis relying on flat vocabularies (Section 4). We will see that hierarchies are then naturally introduced, and we will study methods relying on structured vocabularies (Section 5).

4. Semantic image analysis using unstructured vocabularies

Among the classical approaches, several types of methods are being used:

1. *Direct methods* using a plain representation of data and plain statistical methods,
2. *Linguistic methods* based on the use of an intermediate *visual vocabulary* between raw numerical data and high-level semantics,
3. *Compositional methods* where parts of the image are identified (typically using segmentation) before the whole image or its parts are annotated,
4. *Structural methods* where a geometry of parts is used,
5. *Hierarchical compositional methods* where a hierarchy of parts is constructed for recognition,
6. *Communicating methods* when information is shared between categories,
7. *Hierarchical methods* that search for hierarchical relationships between categories,
8. *Multilabel methods* assigning several global labels simultaneously to an image.

Methods of types 1 and 2 – i.e. direct methods and linguistic methods – generally do not introduce rich semantics. In the following sections, we will focus on the other types of methods, trying to underline how multi-level or multi-faceted annotations are introduced.

4.1. Compositional methods

Compositional methods introduce richer semantics by tagging parts rather than the whole image. Barnard and Forsyth [10] begin by segmenting the image. They use a hierarchical generative model to map labels to regions, and assign a global label to the image. Duygulu et al. [24] extend this model to a "translation model", mapping regions to words, and grouping words corresponding to similar regions. Jeon et al. [52] start from the translation model and suppress the bijection constraint, thus proposing a cross-media relevance model. Fan et al. [33] also use a similar method based on segmentation to detect salient objects before annotating images.

Vogel and Schiele [99] suggest quite a different method: the image is divided in small squares representing a "local concept". Histograms of occurrences of local concepts are used to describe the image globally and to classify it into a scene category. They show that using occurrences of semantic concepts is much more efficient than using occurrence of visual words. They also use this representation to measure the typicality of scenes, which is interesting regarding the theories of Rosch et al. [82], which state that some object instances are more typical of their categories than others.

4.2. Structural methods

Several works underline the importance of using geometrical constraints between parts [36,56,59]. While not tagging these directly, they often notice that the visual "words" they get correspond to semantic concepts (e.g. when describing cars they get wheels, etc.).

Some works do use relations between tagged regions. In scene recognition, Aksoy et al. [2] try to reduce the semantic gap by taking into account the spatial relations between image regions (rather than object parts), using a visual grammar. In a first step, they use classifiers to assign labels to regions. Then, the grammar is used to classify the image in a scene category. Datta et al. [19] also use spatial links between regions to annotate images. Gupta and Davis [45] solve the correspondence problem by exploiting object labels together with prepositions (e.g. "the car is *on* the road") and comparative adjectives (smaller, ...) to resolve ambiguities. Parikh and Chen [75] present hSOs, or *hierarchical semantics of objects*. Working on scene images, they locate salient objects and learn the contextual links between them. For instance, they learn that a computer monitor is often found in the same region as the keyboard, and that finding a telephone next to them would not be very surprising.

4.3. Hierarchical compositional methods

Mojsilovic et al. [71] propose to narrow the semantic gap by using semantic indicators as an intermediate between low-level features and image categories. These indicators (skin, sky, water, ...) were determined by experiments with human subjects. First they extract low-level features from the image, both local and global, quantize and name them, and use these to discover the semantic indicators. For instance, the visual features are used to find a blue area on top of the image, interpreted as sky. Semantic indicators are then used for categorization. Sudderth et al. [91] hierarchically learn parts, objects and scenes. Their image representation is similar to Fergus et al. [36], but with the ability to share parts between objects, and objects between scenes. Li et al. [62] suggest a global method to segment the image, annotate regions and the whole image and categorize the scene, based on the use of a hierarchical generative model.

Epshtein and Ullman [29] build a hierarchy of visual features. Starting with an informative fragment, they search recursively for smaller informative fragments in it. In [30], they extend this method to find so-called *semantic* fragments automatically. Fidler and Leonardis [37] also build a hierarchy of parts using unsupervised statistics. Each layer is built by composition of features of the previous layers. Lower layers are built from simple features, and are category-independent. Higher levels describe more complex features, specific to each category, and sometimes correspond to semantic features (e.g. wheels appear to be among the features for cars). Another method to recognize simultaneously several objects in an image is proposed by Ahuja and Todorovic [1]. They build a hierarchy of object parts based on their co-occurrence in a same object. For instance, "roof" and "chimney" often appear together and can be grouped into a parent node named "roof with chimney". "Windows-panels" can appear both on "windows" and on "doors" and thus has these two categories as parents.

4.4. Communicating methods

A continual problem in learning for recognition is the number of categories that can be learnt by a system. Generalization to new models is not straightforward. And the more categories, the more complex the model, the more memory is needed, the more time for learning. Communicating approaches are meant to follow one or both of these two aims: facilitate integration of new categories in the system, and reduce the complexity of the classifier.

Perronnin [76] learns a vocabulary composed of two parts: the first is universal, shared by all categories, the other is category-specific. Fei-Fei et al. [35] propose a model aimed at learning new categories from a few examples only. They use a Bayesian approach, where a model *a posteriori* is learnt for each category based on the update of a model *a priori* of the world, using a few observations of the new category. Wang et al. [100] model latent themes as an intermediate layer between the visual vocabulary and categories. These themes group words together and are shared among categories. Todorovic and Ahuja [93] learn to find occurrences of objects and object parts shared by several categories. This was also the case with the article by the same authors cited as a hierarchical compositional methods [1]. They outperform SPM on Caltech-256.

Amit et al. [3] introduce a method for multiclass classification that allows sharing both category visual features and classifier parameters. To do so, they reformulate the SVM convex optimization problem using the trace-norm instead of the Frobenius norm used in the original multiclass formulation of SVMs by Crammer and Singer [18]. In this model, information sharing is implicit.

Torralba et al. [95] use a multi-task approach to multiclass classification, using boosting with binary classifiers. They note that this approach considerably reduces the number of features necessary and allows fast classification. In classical methods, the number of features used is linear with the number of categories, whereas in theirs it is logarithmic.

More recently, Thomas et al. [92] propose *cognitive feedback*. They exploit what they have learnt in recognizing objects in a number of images to find new low-level metadata on objects in unknown categories. For instance, they can recognize object parts.

4.5. Hierarchical methods

Hierarchical methods are used in a number of domains as an efficient tool to simplify complex problems. "Divide and Conquer" is a famous principle, and it would also be interesting for object recognition. However, it is not clear how to group categories. When building hierarchies automatically, people generally use low-level visual data rather than conceptual organization of the vocabulary. We call these *visual* hierarchies.

Vasconcelos [98] proposes a method for image indexing based on mixture models built in a hierarchy, allowing better performance and faster classification. Fan and Geman [34] build a hierarchy of classifiers where the terminal nodes give the categorization, with the possibility of having several leaf nodes corresponding to the same category. Wang et al. [100] use the way latent themes are shared among categories and images to build a hierarchy with the initial words as the leaf nodes, and the latent themes as intermediate nodes. The built taxonomy is thus a by-product of the classification algorithm and seems not to be exploited.

Many methods were proposed recently to build visual hierarchies, i.e. hierarchies where the words of the vocabulary are leaf nodes and that describe visual similarities between categories. Some use it to improve classification performance, other to accelerate it. It shows that building hierarchies between concepts is a subject of growing interest.

Bart et al. [11] propose a Bayesian method to find a taxonomy such that an image is generated from a path in the tree. Similar images have a lot of common nodes on their associated paths and therefore a short distance to each other. The method presented is unsupervised and the leaves correspond to visually similar images independent of the categories. Yet they also propose a supervised version in which the leaves would correspond to the categories. With a bag-of-features representation, the supervised hierarchy performed better (68%) than an LDA model (64%) on an image database of 13 scene categories. The LDA model (*latent Dirichlet allocation*) is a generative model that assumes that visual words are generated from a finite set of latent themes.

Sivic et al. [87] suggest an improvement to this model by further supposing that the visual words that are generated have a tree structure. This model, called *hierarchical-LDA*, was originally developed for text analysis. They use it together with segmentation in an unsupervised framework and show that categories correspond to leaf nodes. Griffin and Perona [44] build a hierarchy for faster classification. Instead of using a multiclass classifier over all the categories, they go top-down in the hierarchy. To build it, they first classify images to estimate a confusion matrix. They group together confusing categories in a bottom-up manner. They also build a top-down hierarchy for comparison, by successively dividing categories. Both hierarchies show similar results for speed and accuracy. Marszałek and Schmid [69], compared with the previous methods, remove the separability constraint at each level. When a category cannot be assigned to a node without ambiguity, it is assigned to both, and decision is put forth on the next level. Their hierarchies have a structure more complex than trees.

Another approach is to use human feedback to automatically build more intuitive hierarchies. Such structures are no longer *visual* hierarchies. Rege et al. [80] use humans in the loop through relevance feedback and build a semantic hierarchy based on user experience. Categories are not explicit but stem from the information extracted through the feedback. By combining feedback from all the users, they are able to estimate the hierarchical organization of images, allowing more intuitive search and browsing.

4.6. Multilabel methods

Most of the approaches studied in the previous sections are tackling the problem of categorization. The term *multilabel* is often used for the correspondence problem. Here we will focus on approaches that allow several labels to describe the same image *globally*, or the same object.

This problem of assigning simultaneously a number of labels to an image has been studied formally by Boutell et al. [16]. The main issue is, if a training image is labeled with two labels, it can be used for training none or both categories, or for training a new category grouping the two. The authors study the possible scenarios and show that the most interesting is to use such an image as a positive example for the two categories corresponding to the labels. Another is proposed by Carneiro et al. [17] based on *multiple instance learning*. In this model, a label is assigned to a group of examples when at least one of them is a positive. They use this scheme to estimate the probability of all labels to be associated with the image.

4.7. Conclusion

The methods seen so far address the problem of object recognition using only a basic vocabulary. Yet it is clear that using hierarchical representations attracts attention, as shown by recent interest both for so-called *feature hierarchies* and *visual hierarchies*. The use of context calls for compositional relations between objects of a scene, and part-based approaches make it clear that "conceptual" object parts are useful for recognition. Indeed, several authors are able to name the parts found automatically [1,30,37].

Finding hierarchical links between the categories is a fast expanding research subject. Structure seems to be a good solution for increasing the speed, the accuracy, and reducing the complexity of the systems. In the next section, we will review some of the methods that have been using the vocabulary structure as a supplementary input.

5. Semantic image analysis using structured vocabularies

The methods relying on semantic structures generally use the following two types of relations:

1. Is-A-Part-Of relations are found in compositional methods, such as part-based models, especially for detection and identification.
2. Is-A relations or inheritance relations can be used for sharing descriptors or for better classification.

However, the barrier is not always clear: for instance, using a common parent category might be useful for determining common parts.

We suggest a finer organization of methods, based on how semantic relations are introduced in the system. Possible groups of methods are

1. *Linguistic methods* where the semantic structure is used at the level of the vocabulary, independently from the image, e.g. to expand the vocabulary,
2. *Compositional methods* that use meronyms (i.e. object components),
3. *Communicating methods* use semantic relations to share information between concepts, be it for feature extraction or for classification,
4. *Hierarchical methods* use Is-A relations to improve categorization and/or to allow classification at different semantic levels.

In the literature, the expression "semantic hierarchy" is equally used for both composition and inheritance relations. This is correct but is ambiguous about the true difference of nature that exists between the two.

Most of the systems built with an input structured vocabulary use WordNet. WordNet is a very rich lexical database in English. For each word, some information is given such as a definition, polysemy, synonyms, hypernyms/hyponyms (inheritance relations) and meronyms/holonyms (composition relations).

5.1. Linguistic methods

Approaches in this section exploit a knowledge representation for a richer annotation of images at the level of the vocabulary only. They are generally used in a content-based image retrieval context. Aslandogan et al. [6] use WordNet hierarchy both for query expansion and database vocabulary expansion. Instead of simply matching keywords, they are able to look for similar words (according to WordNet structure). Using Is-A and Member-Of (the people equivalent of Part-Of, e.g. a "musician" is part of a "musical group"), they limit the expansion to words within a given distance to the initial keyword. Yang et al. [103] also perform a vocabulary expansion using WordNet and then reduce the number of words keeping only the most significant statistically. Barnard et al. [8] exploit WordNet for disambiguation: if an annotation word is polysemic, they select the most relevant meaning by comparing with neighbors in WordNet (using hypernyms, holonyms, meronyms, etc.) and in the image. Liu et al. [63]

also use WordNet for vocabulary expansion, especially to access more specific words.

Wang et al. [101] extract keywords from text associated with the images, build a thesaurus automatically using WordNet and image features, and associate words to image regions. Datta et al. [20] use the hierarchy (WordNet) to compute a relevance measure of each term selected to annotate an image. They also use a semantic distance on the hierarchy at query time. Jin et al. [53] similarly use semantic relations and correlation measures to remove irrelevant keywords. Lam and Singh [57] use WordNet to define a similarity measure combined with a visual similarity. Li et al. [62] also use WordNet to remove incoherences between labels and to group synonyms.

Ontologies are used by Soo et al. [89] to standardize annotations and facilitate matching between a query and images in the database. Hollink et al. [47] propose a method to combine annotations formatted with different ontologies in a common structure. Yang et al. [104] use *semantic feedback*: user feedback is processed at the semantic level, based on WordNet, rather than on visual similarities.

Popescu et al. [79] search for similar images both conceptually and visually. Conceptual similarity is computed using text associated with images and a WordNet-based distance. They show that it is more efficient to search for visually similar images among images associated with more specific nodes (hyponyms) than among those associated with words at the generic level.

5.2. *Compositional methods*

The methods presented here use semantic hierarchies based on meronymy/holonymy relations essentially. Though not always explicitly presented as such by the authors, they fall into this category as ‘‘semantic parts’’ (meronyms) are recognized in order to recognize an object (the holonym). Two main methods are used to exploit these relations: those based on artificial intelligence, using logical inference, and those based on statistics.

An early work in this category is that of Rosenthal and Bajcsy [83]. Using a knowledge base, and links between objects as knowledge rules, they use inferences from parts to recognize an object. Part-based models using semantic parts often focus their interest on a single class of objects, such as people in Mohan et al. [70]. In this article, the authors first identify legs, arms and a head, together with their spatial relations, and use it to find people. A similar approach applied to face detection is presented by Arandjelovic and Zisserman [4], who begin by localizing the eyes, the mouth and other manually selected interest points.

In the more general case of scene recognition, Srikanth et al. [90] extend the translational model of Duygulu et al. [24]. They use WordNet to build the vocabulary and for classification.

5.3. *Communicating methods*

As one can get the intuition, if categorization is made among labels that have a common super-category, (a) potentially, there are more common parts between these neighboring categories than with more distant ones; (b) it might be possible to know what are the details (or parts) to look at to differentiate between classes. For instance, when categorizing vehicles, it might be possible to look at wheels and rear mirrors to distinguish between cars and motorcycles. Studies in cognitive psychology confirm this intuition that a more detailed analysis is necessary to recognize sub-categories (e.g. [54]). Experiments also show that different mechanisms are used for each *basic-level* categories [41]. Levi et al. [60] remark that humans are able to learn to recognize new categories from a small number of examples. They assume that features useful for recognition of a sub-category are likely to be relevant for a new sibling sub-category. They build an algorithm accordingly that is able to integrate the most relevant features depending on the category.

Bar-Hillel and Weinshall [7] develop this idea even more explicitly. They recognize *basic-level* categories by using a part-based generative model. To distinguish between sub-categories, they then use discriminative classifiers on the ‘‘specific’’ parts.

5.4. *Hierarchical methods*

Articles presented here are exploiting Is-A relations (i.e. hypernymy/hyponymy) to help the learning of categories.

Maillot and Thonnat [65] combine domain representation with machine learning techniques to categorize specific-domain objects. Objects are classified top-down, going down the hierarchy while a category has sub-categories in which it could be classified. Feature extraction can be adapted depending on the candidate category.

Torralba et al. [94] use millions of tiny images with nearest neighbors. As images are labeled with WordNet nouns, they can categorize images at different levels using a hierarchical vote, where a label node also votes for its parents.

Zweig and Weinshall [105] present a detailed study of classification performances for the binary case (i.e. class of interest vs. background). Categories at the leaf level are tested against the background. They test training a leaf-level category by using other categories in the positive training set, such as siblings, or super-categories (parents, grand-parents). They show that using more generic categories can help recognition.

Marszałek and Schmid [68] exploit WordNet for categorization of objects. They extract the relevant sub-graph of WordNet and simplify it according to the categories of interest. They propose to classify images top-down, starting from the root node and selecting the best hyponym at each stage, down to a leaf node. This hierarchical algorithm shows performance similar to the classical one-vs-rest ‘‘flat’’ classifier and a visual hierarchy, but allows better performance at generic levels. They also suggest using meronyms. The detection of object components leads unfortunately to a large increase of the number of classifiers used by the method. On the positive side, it allows a small increase in performance, and more importantly, has the benefit to provide information reusable when new objects with similar components enter the database.

Fan et al. [31], Gao and Fan [39] and Fan et al. [32,33] suggest using a conceptual ontology that they build using both conceptual similarity (based on WordNet) and visual similarity. They propose a hierarchical boosting algorithm based on this ontology (i.e. hierarchy) allowing image annotation at different levels of genericness. The classification task is made top-down and features are added to avoid error propagation.

Tousch et al. [96] use a hand-made hierarchy describing the ‘‘car’’ domain to classify images according to a trade-off between semantic precision and accuracy. The hierarchy is not limited to a tree. Basically, they classify images for all possible labels and compute the probability that each consistent combination of labels is associated with the image. They also use meronyms as they describe cars using specific parts detectors.

Binder et al. [13] show that using a taxonomy structure, one can outperform multiclass classification approaches, with respect to a taxonomy loss. They incorporate the structure in the learning process by using a modified version of SVMs.

5.5. *Conclusion*

Hierarchies have not been used extensively so far, except in image retrieval where WordNet has been used to improve textual

Table 2
Performance as reported by the authors of a number of methods described in the review. A: accuracy, R: recall, P: precision.

Paper	# Categories	# Keywords	# Training images	Type of images	Performances
Classification					
Yang et al. [103]	10	530	2500	Landscape	A: 80%
Lazebnik et al. [58]	101	N/A	30 per category	Caltech-101	A: 64.6%
Griffin et al. [43]	101	N/A	30 per category	Caltech-101	A: 67.6%
Griffin et al. [43]	256	N/A	30 per category	Caltech-256	A: 34.1%
Bosch et al. [15]	256	N/A	30 per category	Caltech-256	A: 45.3%
Marszałek and Schmid [68]	10	42/563	1277	Pascal VOC 2006	A: 80–90%
Bosch et al. [14]	6	N/A	600	Corel, nature scenes	A: 76.92%
Vogel and Schiele [99]	6	9	600	Scenes	A: 74.1%
Fan et al. [32]	120		350+ per category	Corel+LabelMe	P: 40–95%
Li et al. [62]	8 scenes, 30 objects	1256	600	Sport scenes	R: 73% (objects), 54% (scenes)
ImageNet Large Scale Visual Recognition Challenge [49]	1000	N/A	1.2 million	Imagenet	A: 53%
Image retrieval					
Srikanth et al. [90]	N/A	371 (42 predicted)	4500	Corel	P: 26.34%, R: 27.24%
Papadopoulos et al. [74]	N/A	4	40	Beach vacation	A: 83.20%
ALIPR/Li and Wang [61]	599	332	80 per category	Corel	P: ≈ 40%, R: ≈ 12% (1 mot)
Makadia et al. [66]	N/A	260	4500	Corel	P: 27%, R: 32%
		291	17,825	IAPR-TC12	P: 28%, R: 29%
		269	19,659	ESP-Game	P: 22%, R: 25%

queries and annotation. A major hindrance has certainly been the difficulty to find appropriate ontologies, though WordNet seems to be a good starting point for a lot of applications. Hierarchies, and especially their meronymy relations, are often used for specific applications (e.g. [70,4,7,65,96]).

6. Evaluation

The methods presented in this review were tested on several databases. Performance is evaluated using accuracy rate for classification and/or precision and recall values for image retrieval. Categories (or words) are evaluated independently, i.e. no structure is used in the process. For a given category $\mathcal{C}$, let N_t be the number of test images labeled as $\mathcal{C}$. Let N^+ be the number of images classified in $\mathcal{C}$, and $\mathcal{N}_t^+$ be the number of images correctly classified in $\mathcal{C}$. Precision and recall values for class $\mathcal{C}$ are given by

$$P = \frac{N_t^+}{N^+}; \quad R = \frac{N_t^+}{N_t}. \tag{1}$$

In classification, accuracy A is often used and corresponds to the mean of the confusion matrix diagonal.

Databases are a crucial issue in image recognition [78], and until a couple of years ago, no database would use annotations related to a semantic hierarchy. As a rule, people would use simple vocabularies and project them into a hierarchy such as WordNet. In 2007, Barnard et al. [9] released annotations for the Corel database, where annotations of image segments would correspond to WordNet nodes, together with specific evaluation methods. Griffin et al. [43] released Caltech-256 and proposed an associated hierarchy. In 2009, Deng et al. [22] proposed to populate the WordNet hierarchy with images and created ImageNet. This interesting database combines a rich vocabulary (more than a 1000 WordNet concepts) with a fairly high number of images per concept. The *Visipedia* project [102] proposes a database of images of 200 birds with a corresponding ontology. To our knowledge, it is the only one that addresses data classification at a subordinate or fine grained level.

Performance evaluation is shown in Table 2 for a number of reviewed articles. The first report that can be done is that there is a big difference between categorization and image retrieval (or annotation). Clearly, results depend heavily on the number of categories or terms used. The techniques used cannot be the only reason for this difference, since they follow a common scheme (i) extract image features and (ii) learn correspondence with keywords. The first conclusion is therefore that it is difficult to keep a high level of performance while increasing the size of the domain (i.e. the number of terms). However, this first impression is not the only explanation. Some databases are "easier" than others. Indeed, the example of Griffin et al. [43] speaks for itself: results drop when getting from Caltech-101 to Caltech-256. The number of categories is not the sole explanation: tests by the same authors using 100 categories from Caltech-256 also show poor results comparing to Caltech-101 (about 40–45% accuracy on 100 categories from Caltech-256 vs. 67.6% for Caltech-101).

7. Discussion

The idea of this survey was to establish a state of the art of the emerging field of structured annotations for image data. The use of structured data both as input and output of the annotation process has a long history in research areas with a symbolic flavour such as data mining, logical artificial intelligence and computational linguistics. In fields such as computer vision, pattern recognition and statistical learning, which mostly rely on a numerical formalism, the volume of studies, although increasing, is more circumscribed.

Progress has been made undoubtedly in the last decade toward better image understanding. More and more efficient image descriptions have been developed. New learning tools have made unstructured classification easy and accurate as long as the number of classes is small (see the performances in Table 2). However, when it comes to large number of classes, performances are still far from being satisfying. Public reactions to systems such as ALIPR[1] [61] show that common people are far from adopting it.[2] The "Semantic Gap" question is still waiting for a clear and satisfying formulation and solution.

In the late 90s, it has been explained that there was a gap between the user needs and the purely visual-based image retrieval systems [85,88]. Concept-based image retrieval [84,27]

[1] http://www.alipr.com/

[2] Comments to an article: *ALIPR Helps People Decide*: *Hot Or Not*? http://gizmodo.com/213698/alipr-helps-people-decide-hot-or-not—accessed June 2009.

Please cite this article as: A.-M. Tousch, et al., Semantic hierarchies for image annotation: A survey, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.05.017

advocated for incorporating semantic features. The issue has been largely worked on since, yet our review shows that efficiency is still restricted to small vocabularies. Scalability in vocabulary size is indeed the new issue for future image data annotation. Our study shows that using hierarchies such as WordNet helped to increase the size of the annotation vocabulary. Up to now, only a few attempts have been made to use it in the recognition process itself, and most of the studies use it as annotations' *post-processing* to enhance the richness of the vocabulary, for instance.

The current generation of image interpretation algorithms fully exploits the power of empirical learning techniques. They have demonstrated their usefulness on several problems compared with "old fashioned" geometric approaches, for instance. However, they often suffer from a Black Box syndrome, i.e. they are usually not able to justify the quality of their results, limiting the deployment of such solutions in real or operational contexts. Structured outputs, with the adequate representation space, may be able to take advantage of the expressive capacity of pure symbolic approaches, as in Markov logic networks [81].

In the same spirit, annotation noise is a problem that could be handled more easily with structured vocabularies. In unstructured vocabulary spaces, annotation metrics or topology are discrete and rely mainly on 0–1 loss or their smoother versions. A richer structure, allowing hierarchical relations, makes possible the management of intermediate outputs and the definition of several trade-offs, for instance between precision or complexity of description and confidence [96]. Controlling such type of trade-off is a way to bound the effect of annotation noise in complex processing chains.

Standardization of structured vocabulary is an open issue. Most of the available studies rely either on WordNet or on specifically tailored annotation spaces. WordNet is clearly limited in its structure and expressive capacity. As an example, let us take the "car" node of WordNet. Hyponym labels "Compact" and "Ambulance" are compatible. These correspond to two conceptual aspects or facets of a car: the first describes its size, the second its function. Multi-faceted or multiple non-exclusive annotations have not been fully addressed [96]. One of the reason is the lack of available data: in ImageNet, for instance, images are univocally annotated. Following WordNet, they make no distinction between siblings nodes. It would desirable to have access to richer hierarchies and corresponding data bases that would encode explicitly different ways of characterizing the same piece of data and allow multi-level semantic analysis.

WordNet content is also limited in its scope. Professional or expert applications exploit specific vocabularies, usually at a subordinate level. If we stay in the same vehicle context, WordNet does not contain any vocabulary describing car brands or models. However, recently available databases [102] provide fine grained expert annotations, giving rise to new types of problems and annotation schemes.

Empirical performance evaluation is now compulsory in any image interpretation study. However, clear metrics able to evaluate how annotation process fails to bridge the semantic gap is still waited. Most of the metrics measure the gap between an ideal groundtruth and the actual annotation and count errors between categories whatever their semantic relations. For instance, a confusion between a dog and a horse has the same importance as between a dog and a car. The recent ImageNet challenge [49] presents alternative ways of evaluating annotation results, some of them taking into account the taxonomy provided with the database. Tousch et al. [96] propose to define an error along a genericness/specificity axis based on the fact that errors at specific levels are less problematic than at generic levels. However, quantitative measures of relevance or consistency of annotations, which are an important dimension of the semantic gap issue, seem to have been disregarded.

Table 3
Comparison of unstructured and structured vocabulary approaches.

Unstructured vocabulary
Pros
• Generic design: all labels at the same level
• Clear but limited evaluation metrics
• Availability of evaluation databases
• Rather efficient with a limited vocabulary
• Large volume of studies in computer vision, pattern recognition and statistical learning
Cons
• Scalability in vocabulary size not demonstrated
• No flexibility in semantic level of description
• Algorithms are rather "Black box" (i.e. decisions are hard to justify)
• Noise or uncertainty in annotation not easily handled
• Mainly general category level vocabulary, not fine grained
Structured vocabulary
Pros
• Handle large vocabularies
• Can manage various levels of description or categorization, contextual phenomena, user point of view
• Can exploit knowledge based representations
• Can produce explanation of decision
• Large volume of studies in artificial intelligence, data mining and computational linguistics
Cons
• No satisfying standard semantic hierarchy
• Evaluation metrics not finalized
• Few learning or evaluation databases available

Table 3 summarizes the above discussion, and compares the pros and cons of the unstructured and structured vocabulary approaches.

8. Conclusion

This review has presented some key aspects of the tremendous research work that has been carried out toward bridging the semantic gap. A number of observations were made that can be summarized as follows:

- The use of semantic hierarchies is of growing interest, whether as an input or as an automatic construction; they appear to be helpful when the number of categories grows;
- Semantic hierarchies allow better performance than no structure when working on vocabularies only, independently of the images;
- Recent studies prove an interest in the use of semantic hierarchies for semantic image analysis.

Future avenues of research should develop the following topics:

- *Full exploitation of semantic hierarchies*: People have often been using only some of the hierarchical relations, i.e. compositions (Part-Of) *or* inheritance (Is-A), rarely both (with the exception of [68]). Efforts toward a "total understanding" as in Li et al. [62] should also incorporate "total" hierarchies.
- *Multi-level and multi-faceted image interpretation*: Hierarchies can be used to create richer forms of interpretation output. Words describing different aspects of the image should be used and weighted according to their relevance/importance in the image, and/or their reliability. This is a first step toward annotations that can adapt to context.
- *Evaluation protocols adapted to semantic hierarchy handling*: A few image databases with a semantic flavoured ground truth

Please cite this article as: A.-M. Tousch, et al., Semantic hierarchies for image annotation: A survey, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.05.017

are available. Specific metrics addressing various aspects of the semantic annotation process still need to be settled.

References

[1] N. Ahuja, S. Todorovic, Learning the taxonomy and models of categories present in arbitrary images, in: ICCV07, 2007, pp. 1–8.

[2] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, J. Tilton, Learning Bayesian classifiers for scene classification with a visual grammar, in: GeoRS, vol. 43, no. 3, 2005, pp. 581–589.

[3] Y. Amit, M. Fink, N. Srebro, S. Ullman, Uncovering shared structures in multiclass classification, in: ICML'07, 2007, pp. 17–24.

[4] O. Arandjelovic, A. Zisserman, Automatic face recognition for film character retrieval in feature-length films, in: CVPR05, 2005, pp. 860–867.

[5] L.H. Armitage, P.G. Enser, Analysis of user need in image archives, Journal of Information Science 24 (4) (1997) 287–299.

[6] Y.A. Aslandogan, C. Thier, C.T. Yu, J. Zou, N. Rishe, Using semantic contents and wordnet in image retrieval, in: SIGIR Forum, vol. 31, no. SI, 1997, pp. 286–295.

[7] A. Bar-Hillel, D. Weinshall, Subordinate class recognition using relational object models, in: NIPS'06, 2006, pp. 73–80.

[8] K. Barnard, P. Duygulu, D. Forsyth, Clustering art, in: CVPR'01, 2001, pp. II 434–II 439.

[9] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, J. Kaufhold, Evaluation of localized semantics: data methodology, and experiments, International Journal of Computer Vision 77 (1–3) (2008) 199–217.

[10] K. Barnard, D. Forsyth, Learning the semantics of words and pictures, ICCV'01, vol. 2, IEEE, Vancouver, 2001, pp. 408–415.

[11] E. Bart, I. Porteous, P. Perona, M. Welling, Unsupervised learning of visual taxonomies, in: CVPR08, 2008, pp. 1–8.

[12] I. Biederman, Recognition-by-components: a theory of human image understanding, Psychological Review 94 (2) (1987) 115–147.

[13] A. Binder, M. Kawanabe, U. Brefeld, Efficient classification of images with taxonomies, in: ACCV'09, 2009.

[14] A. Bosch, X. Muñoz, R. Martí, Which is the best way to organize/classify images by content? IVC 25 (6) (2007) 778–791.

[15] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: ICCV'07, 2007, pp. 1–8.

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

[17] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 394–410.

[18] K. Crammer, Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, Journal of Machine Learning Research 2 (2002) 265–292.

[19] R. Datta, W. Ge, J. Li, J.Z. Wang, Toward bridging the annotation-retrieval gap in image search by a generative modeling approach, in: MULTIMEDIA'06, 2006, pp. 977–986.

[20] R. Datta, W. Ge, J. Li, J.Z. Wang, Toward bridging the annotation-retrieval gap in image search, International Transactions on Multimedia 14 (3) (2007) 24–35.

[21] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 1–60.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: CVPR'09, 2009, pp. 248–255.

[24] P. Duygulu, K. Barnard, J.F.G.D. Freitas, D.A. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: ECCV'02, 2002, pp. 97–112.

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

[26] J.P. Eakins, P. Briggs, B. Burford, Image retrieval interfaces: a user perspective, in: CIVR'04, 2004, pp. 628–637.

[27] P. Enser, Visual image retrieval: seeking the alliance of concept-based and content-based paradigms, Journal of Information Science 26 (4) (2000) 199–210.

[28] P.G.B. Enser, C.J. Sandom, Towards a comprehensive survey of the semantic gap in visual image retrieval, in: CIVR'03, 2003, pp. 291–299.

[29] B. Epshtein, S. Ullman, Feature hierarchies for object classification, in: ICCV'05, 2005, pp. 220–227.

[30] B. Epshtein, S. Ullman, Semantic hierarchies for recognizing objects and parts, in: CVPR'07, 2007, pp. 1–8.

[31] J. Fan, Y. Gao, H. Luo, Hierarchical classification for automatic image annotation, in: SIGIR'07, 2007, pp. 111–118.

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

[33] J. Fan, Y. Gao, H. Luo, R. Jain, Mining multilevel image semantics via hierarchical classification, IEEE Transactions on Multimedia 10 (2) (2008) 167–187.

[34] X. Fan, D. Geman, Hierarchical object indexing and sequential learning, in: ICPR'04, 2004, pp. 65–68.

[35] L. Fei-Fei, R. Fergus, P. Perona, A Bayesian approach to unsupervised one-shot learning of object categories, in: ICCV'03, 2003, p. 1134.

[36] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: CVPR'02, vol. 264, 2003.

[37] S. Fidler, A. Leonardis, Towards scalable representations of object categories: learning a hierarchy of parts, in: CVPR07, 2007, pp. 1–8.

[38] C.O. Frost, B. Taylor, A. Noakes, S. Markel, D. Torres, K.M. Drabenstott, Browse and search patterns in a digital image database, Information Retrieval 1 (4) (2000) 287–313.

[39] Y. Gao, J. Fan, Incorporating concept ontology to enable probabilistic concept reasoning for multi-level image annotation, in: MIR'06, 2006, pp. 79–88.

[40] L.M. Garshol, Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all, Journal of Information Science 30 (4) (2004) 378–391.

[41] C. Gerlach, Category-specificity in visual object recognition, Cognition 111 (3) (2009) 281–301.

[42] A. Gilchrist, Thesauri, taxonomies and ontologies—an etymological note, Journal of Documentation 59 (1) (2003) 7–18.

[43] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, Technical Report no. 7694, California Institute of Technology, 2007.

[44] G. Griffin, P. Perona, Learning and using taxonomies for fast visual categorization, in: CVPR'08, 2008, pp. 1–8.

[45] A. Gupta, L.S. Davis, Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifiers, in: ECCV'08, 2008, pp. 16–29.

[46] J.S. Hare, P.H. Lewis, P.G.B. Enser, C.J. Sandom, Mind the gap: another look at the problem of the semantic gap in image retrieval, in: Multimedia Content Analysis, Management and Retrieval, SPIE, vol. 6073, 2006, pp. 607309. 1–607309.12.

[47] L. Hollink, G. Schreiber, B. Wielemaker, B. Wielinga, Semantic annotation of image collections, in: KCAP'03, 2003.

[48] L. Hollink, G. Schreiber, B. Wielinga, M. Worring, Classification of user image descriptions, International Journal of Human Computer Studies 61 (5) (2004) 601–626.

[49] ImageNet Large Scale Visual Recognition Challenge 2010, url 〈http://www.image-net.org/challenges/LSVRC/2010/index〉.

[50] A. Jaimes, Human factors in automatic image retrieval system design and evaluation, SPIE, vol. 6061, 2006, pp. 606103.1–606103.9.

[51] A. Jaimes, S.-F. Chang, Conceptual framework for indexing visual information at multiple levels, SPIE, vol. 3964, 2000, pp. 2–15.

[52] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: SIGIR'03, 2003, pp. 119–126.

[53] Y. Jin, L. Khan, L. Wang, M. Awad, Image annotations by combining multiple evidence wordnet, in: MULTIMEDIA'05, 2005, pp. 706–715.

[54] P. Jolicoeur, M.A. Gluck, S.M. Kosslyn, Pictures and names: making the connection, Cognitive Psychology 16 (2) (1984) 243–275.

[55] C. Jörgensen, Attributes of images in describing tasks, Information Processing & Management 34 (2–3) (1998) 161–174.

[56] A. Kushal, C. Schmid, J. Ponce, Flexible object models for category-level 3D object recognition, in: CVPR'07, 2007, pp. 1–8.

[57] T. Lam, R. Singh, Semantically relevant image retrieval by combining image and linguistic analysis, in: ISV'C06, 2006, pp. 770–779.

[58] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: CVPR'06, 2006, pp. 2169–2178.

[59] B. Leibe, A. Ettlin, B. Schiele, Learning semantic object parts for object categorization, Image and Vision Computing 26 (1) (2008) 15–26.

[60] K. Levi, M. Fink, Y. Weiss, Learning from a small number of training examples by exploiting object categories, in: LCV'04, 2004, p. 96.

[61] J. Li, J. Wang, Real-time computerized annotation of pictures, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6) (2008) 985–1002.

[62] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: classification annotation and segmentation in an automatic framework, in: CVPR'09, 2009, pp. 2036–2043.

[63] S. Liu, L.-T. Chia, S. Chan, On the move to meaningful Internet systems 2004: CoopIS, DOA, and ODBASE, in: Lecture Notes in Computer Sciences, vol. 3291, 2004, pp. 1050–1061.

[64] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognition 40 (1) (2007) 262–282.

[65] N.E. Maillot, M. Thonnat, Ontology based complex object recognition, Image and Vision Computing 26 (1) (2008) 102–113.

[66] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: ECCV '08, 2008, pp. 316–329.

[67] M. Markkula, E. Sormunen, End-user searching challenges indexing practices in the digital newspaper photo archive, Information Retrieval 1 (4) (2000) 259–285.

[68] M. Marszałek, C. Schmid, Semantic hierarchies for visual object recognition, in: CVPR'07, October 2007, pp. 1–7.

[69] M. Marszałek, C. Schmid, Constructing category hierarchies for visual recognition, in: ECCV'08, 2008, pp. 479–491.

[70] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (4) (2001) 349–361.

[71] A. Mojsilovic, J. Gomes, B. Rogowitz, Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues, International Journal of Computer Vision 56 (1–2) (2004) 79–107.

[72] J. Mundy, Object recognition in the geometric era: a retrospective, in: CLOR'06, 2006, pp. 3–28.
[73] G.L. Murphy, E.E. Smith, Basic-level superiority in picture categorization, Journal of Verbal Learning and Verbal Behavior 21 (1) (1982) 1–20.
[74] G.T. Papadopoulos, V. Mezaris, S. Dasiopoulou, I. Kompatsiaris, Semantic image analysis using a learning approach and spatial context, in: SAMT'06, 2006.
[75] D. Parikh, T. Chen, Hierarchical semantics of objects (hsos), in: ICCV'07, 2007, pp. 1–8.
[76] F. Perronnin, Universal and adapted vocabularies for generic visual categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2008) 1243–1256.
[77] A. Pinz, Object categorization, Foundations and Trends in Computer Graphics and Vision 1 (4) (2005) 255–353.
[78] J. Ponce, M. Hebert, C. Schmid, A. Zisserman, Toward category-level object recognition, in: Lecture Notes in Computer Sciences, vol. 4170, 2006.
[79] A. Popescu, C. Millet, P.-A. Moëllic, Ontology driven content based image retrieval, in: CIVR'07, 2007, pp. 387–394.
[80] M. Rege, M. Dong, F. Fotouhi, Building a user-centered semantic hierarchy in image databases, Multimedia Systems 12 (4–5) (2007) 325–338.
[81] M. Richardson, P. Domingos, Markov logic networks, Machine Learning 62 (1–2) (2006) 107–136.
[82] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cognitive Psychology 8 (3) (1976) 382–439.
[83] D.A. Rosenthal, R. Bajcsy, Visual and conceptual hierarchy: a paradigm for studies of automated generation of recognition strategies, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (3) (1984) 319–325.
[84] Y. Rui, T. Huang, S. Chang, Image retrieval: current, techniques promising directions, and open issues, Journal of Visual Communication and Image Representation 10 (1) (1999) 39–62.
[85] S. Santini, R. Jain, Beyond query by example, in: WMSP'98, 1998, pp. 3–8.
[86] S. Shatford Layne, Some issues in the indexing of images, Journal of the American Society for Information Science 45 (8) (1994) 583–588.
[87] J. Sivic, B. Russell, A. Zisserman, W. Freeman, A. Efros, Unsupervised discovery of visual object class hierarchies, in: CVPR'08, 2008, pp. 1–8.
[88] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.
[89] V.-W. Soo, C.-Y. Lee, C.-C. Li, S.L. Chen, C.-C. Chen, Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques, in: JCDL'03, 2003, pp. 61–72.
[90] M. Srikanth, J. Varner, M. Bowden, D. Moldovan, Exploiting ontologies for automatic image annotation, in: SIGIR'05, 2005, pp. 552–558.
[91] E.B. Sudderth, A. Torralba, W.T. Freeman, A.S. Willsky, Learning hierarchical models of scenes, objects, and parts, ICCV'05, vol. II, 2005, pp. 1331–1338.
[92] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, L.V. Gool, Shape-from-recognition: recognition enables meta-data transfer, CVIU (2009).
[93] S. Todorovic, N. Ahuja, Learning subcategory relevances to category recognition, in: CVPR'08, 2008.
[94] A. Torralba, R. Fergus, W. Freeman, 80 million tiny images: a large dataset for non-parametric object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (11) (2008) 1958–1970.
[95] A. Torralba, K.P. Murphy, W.T. Freeman, Sharing visual features for multiclass and multiview object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 854–869.
[96] A.-M. Tousch, S. Herbin, J.-Y. Audibert, Semantic lattices for multiple annotation of images, in: MIR'08, 2008, pp. 342–349.
[97] M.J. Tovée, An Introduction to the Visual System, Cambridge University Press, 1996.
[98] N. Vasconcelos, Image indexing with mixture hierarchies, in: CVPR'01, 2001, pp. 3–10.
[99] J. Vogel, B. Schiele, Semantic modeling of natural scenes for content-based image retrieval, International Journal of Computer Vision 72 (2) (2007) 133–157.
[100] G. Wang, Y. Zhang, L. Fei-Fei, Using dependent regions for object categorization in a generative framework, in: CVPR'06, 2006, pp. 1597–1604.
[101] X. Wang, W. Ma, X. Li, Data-driven approach for bridging the cognitive gap in image retrieval, ICME'04, vol. 3, 2004.
[102] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD Birds 200, California Institute of Technology, CNS-TR-2010-001, 2010.
[103] C. Yang, M. Dong, F. Fotouhi, Learning the semantics in image retrieval—a natural language processing approach, in: CVPRW'04, 2004, p. 137.
[104] C. Yang, M. Dong, F. Fotouhi, Semantic feedback for interactive image retrieval, in: MULTIMEDIA'05, 2005, pp. 415–418.
[105] A. Zweig, D. Weinshall, Exploiting object hierarchy: combining models from different category levels, in: ICCV'07, 2007, pp. 1–8.

Anne-Marie Tousch received an engineering degree from ENSEA in 2006, the M.Sc. degree from the University of Cergy-Pontoise the same year and the Ph.D. degree in Computer Science from Ecole des Ponts ParisTech in 2010. She is now a research engineer at CVDM-Solutions, a company specialising in computer vision technologies. Her research interest includes information retrieval, machine learning and computer vision.

Stéphane Herbin received an engineering degree from the Ecole Supérieure d'Electricité (Supélec), the M.Sc. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign, and the Ph.D. degree in applied mathematics from the Ecole Normale Supérieure de Cachan. Employed by ONERA since 2000, he works in the Information Processing and Modelling Department. His main research interests are stochastic modeling and analysis for object recognition and scene interpretation in images and video.

Jean-Yves Audibert received the Ph.D. degree in Mathematics from the University of Paris 6 in 2004. Since then, he is a researcher in the Computer Science department at Ecole des Ponts ParisTech. In parallel, he has worked in a joint INRIA/ENS/CNRS project: Willow from 2007 to 2010 and Sierra since 2011. His research interest and publications range from Statistics to Computer Vision, including theoretical properties of machine learning procedures, sequential prediction, bandit policies, aggregation of estimators, boosting algorithms, kernel machines, object recognition and image segmentation.

Please cite this article as: A.-M. Tousch, et al., Semantic hierarchies for image annotation: A survey, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.05.017