# Building Semantic Hierarchies Faithful to Image Semantics

Hichem Bannour, Céline Hudelot

# Building Semantic Hierarchies Faithful to Image Semantics

Hichem Bannour and Céline Hudelot

Applied Mathematics and Systems Department, Ecole Centrale Paris
92 295 CHÂTENAY-MALABRY, France
{Hichem.bannour, Celine.hudelot}@ecp.fr

**Abstract.** This paper proposes a new image-semantic measure, named "Semantico-Visual Relatedness of Concepts" (*SVRC*), to estimate the semantic similarity between concepts. The proposed measure incorporates visual, conceptual and contextual information to provide a measure which is more meaningful and more representative of image semantics. We also propose a new methodology to automatically build a semantic hierarchy suitable for the purpose of image annotation and/or classification. The building is based on the previously proposed measure *SVRC* and on a new heuristic, named *TRUST-ME*, to connect concepts with higher relatedness till the building of the final hierarchy. The built hierarchy explicitly encodes a general to specific concepts relationship and therefore provides a semantic structure to concepts which facilitates the semantic interpretation of images. Our experiments showed that the use of the constructed semantic hierarchies as a hierarchical classification framework provides a better image annotation.

## 1 Introduction

Achieving high level semantic interpretation of images is necessary to match user expectations in image retrieval systems. Effective tools are then required to allow a precise semantic description of images and allow at the same time a good interpretation of them. A wide number of approaches have been proposed for automatic image annotation, i.e. the textual description of images, to address the well-known *semantic gap* [23] problem. However in most of the proposed approaches the semantics is often limited to its perceptual manifestation, i.e. by the learning of high-level concepts from low-level features [3, 14]. These approaches adequately describe the visual content of images but are unable to extract image semantics as humans can do. They are also faced with the scalability problem when dealing with broad content image databases [16]. The obtained performance varies significantly according to the concept number and the targeted data sets as well [13]. This variability may be explained by the huge intra-concept variability and wide inter-concept similarities on their visual properties that often lead to uncertain annotations and even contradictory. Thus, it is clear there is a lack of coincidence between the high-level semantic concepts and the low-level features, and that semantics is not always correlated

with visual appearance. Therefore, the only use of machine learning seems to be insufficient to solve the problem of image annotation.

A new trend to overcome the aforementioned problems is to use semantic hierarchies [2]. Indeed, the use of explicit knowledge such as semantic hierarchies can help reduce, or even remove this uncertainty by supplying formal frameworks to argue about the coherence of extracted information from images. Semantic hierarchies have shown to be very useful to narrow the semantic gap [7]. Three types of hierarchies have been recently explored for image annotation and classification: 1) language-based hierarchies: based on textual information (ex. tags, surrounding context, WordNet, Wikipedia, etc.) [18, 24, 8], 2) visual hierarchies: based on low-level image features [22, 4, 26], 3) semantic hierarchies: based on both textual and visual features [15, 9, 25]. Although the two first approaches have received more attention, they showed a limited success in their general usage. Indeed, conceptual semantics is often not correlated with perceptual semantics, and is then insufficient to build a good hierarchy for image annotation. Whereas perceptual semantics cannot lead by itself to have a meaningful semantic hierarchy, as it is hard to interpret in higher levels of abstraction. Therefore, it seems mandatory to combine the both component of image semantics in order to build a semantic hierarchy faithful to image application purposes. The use of semantic hierarchies is then more convenient as they consider both, perceptual and conceptual semantics.

The rest of this paper is structured as follows: Section 2 reviews some related work. Section 3 introduces our proposal to build suitable semantic hierarchies for image annotation. Section 4 reports our experimental results on Pascal VOC dataset. The paper is concluded in Section 5.

## 2 Related Work

Several methods [15, 9, 18, 24, 22, 4] have been proposed to build semantic hierarchies dedicated to image annotation. A semantic hierarchy classifier based on WordNet is proposed in [18]. Their hierarchy is built by extracting the relevant subgraph of WordNet that may link all concepts. ImageNet is proposed in [8], which is a large-scale ontology of images built upon the backbone of WordNet. LSCOM [19] aims to design a taxonomy with a coverage of around 1000 concepts for broadcast news video retrieval. An Ontology-enriched Semantic Space (OSS) was built in [24] to ensure globally consistent comparison of semantic similarities. The above approaches can be qualified as language-based hierarchies, as those hierarchies are built upon textual information. While these hierarchies are useful to provide a meaningful structure (organization) for concepts, they ignore visual information which is an important part of image semantics.

Other approaches are based on visual information [22, 4, 26]. An image parsing to text description (I2T) framework is proposed in [26], which generates text descriptions for images and videos. I2T is mainly based on an And-or Graph for visual knowledge representation. Sivic & al. propose to group visual objects using a multi-layer hierarchy tree that is based on common visual elements [22]. Bart

& al. proposed a Bayesian method to organize a collection of images into a tree shaped hierarchy [4]. A method to automatically build classification taxonomy in order to increase classification rapidity is proposed in [12]. These hierarchies serve to provide a visual taxonomy, and a major problem with them is how they can be interpreted in higher levels of abstraction. Therefore, building meaningful semantic hierarchies should be done upon both semantic and visual information.

Among approaches for building semantic hierarchies, Li & al. [15] proposed a method based on visual features and tags to automatically build the "semantivisual" image hierarchy. A Semantic hierarchy based on contextual and visual similarity is proposed in [9]. Fan & al. [10] proposed an algorithm to integrate the visual similarity contexts between the images and the semantic similarity contexts between their tags for topic network generation. Flickr distance is proposed in [25], which is a novel measurement of the relationship between semantic concepts in visual domain. A visual concept network (VCNet) based on Flickr distance is also proposed [25]. Semantic hierarchies have great potential to improve image annotation, particularly through their explicit representation of concepts relationships that may help to understand image semantics.

### 2.1 Discussion

Many approaches for hierarchical image annotation use WordNet as a hierarchy of concepts [18, 8]. However, WordNet is not very appropriate to model image semantics. Concepts organization in WordNet follows a psycholinguistic structure, which may be useful for reasoning about concepts and understand their meaning, but is limited and inefficient to reason about image context or its content. Indeed, distances between related concepts in WordNet do not necessarily reflect an appropriate semantic measure for reasoning about images, i.e. distances between concepts is not proportional to their semantic relatedness with respect to image domain. For example, according to the shortest path in WordNet the semantic relatedness of "shark" and "whale" is 11 (nodes), and of "man" and "whale" is 7. This is meant that concept "whale" is closer to "human" than to "shark". This is coherent from a biological point of view because "whale" and "human" are mammal while "shark" is not. However, in image domain it is more accurate to have higher similarity between "shark" and "whale" as they live in the same environment, share many visual features, and it is more common that they co-appear in a photo, unlike with humans. Then, an appropriate semantic hierarchy should represent this information or allow it to be deducted to help understand image semantics.

## 3 Building of the Hierarchy

Based on the previous discussion, we define the following assumptions underlying our approach: *A suitable semantic hierarchy for image annotation should: 1) model images context (as defined in the previous section), 2) allow grouping visually similar concepts in order to obtain better performance of classifiers, 3)*

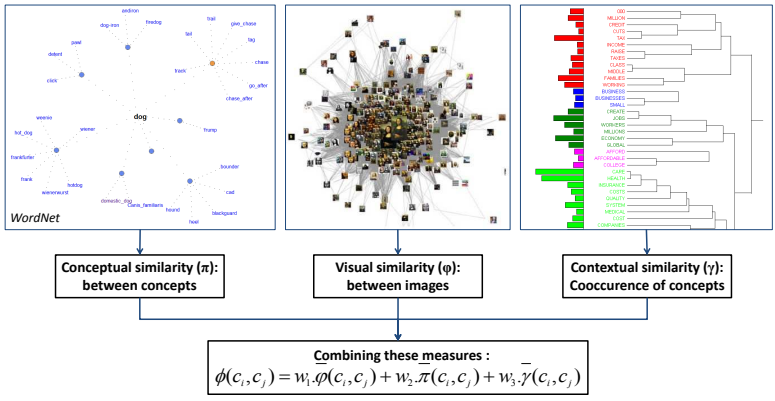



**Fig. 1.** The *SVRC* is based on visual, conceptual and contextual similarities.

*reflect image semantics, i.e. the organization of concepts into the hierarchy and their semantic relatedness reflect image semantics.*

Following the above assumptions, we propose in this paper a new method for building appropriate semantic hierarchies to images 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. 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 concepts definition in WordNet, and 3) a contextual similarity which measures the distributional similarity between each pair of concepts (cf. Fig.1). *SVRC* is then used in *TRUST-ME*, a set of heuristic rules that allow deciding the likelihood of the semantic relatedness between concepts, and help building the hierarchy.

Given a set of pairs image/annotation, where each annotation describes a set of concepts associated with an image, our approach allows to automatically build a semantic hierarchy suitable for image annotation. Formally, we consider $I =< i_1, i_2, \cdots, i_{\mathcal{L}} >$ all images of a considered database, and $C =< c_1, c_2, \cdots, c_{\mathcal{N}} >$ the annotation vocabulary of these images, i.e. the set of concepts associated with these images. The approach we propose consists in identifying $\mathcal{M}$ new concepts that link all the concepts of $C$ in a hierarchical structure that best represents image semantics.

### 3.1 Visual Similarity

Let $x_i^v$ be any visual representation of an image $i$ (a visual features vector), we learn 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) [6] (one-versus-all) with a decision function $\mathcal{G}(x^v)$:

$$\mathcal{G}(x^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x^v) + b \tag{1}$$

where $\mathbf{K}(x_i^v, x^v)$ is the value of a kernel function for the training sample $x_i^v$ and the test sample $x^v$, $y_i \in \{1, -1\}$ the class label of $x_i^v$, $\alpha_i$ the learned weight of the training sample $x_i^v$, and $b$ is a learned threshold parameter. Notice that the training samples $x_i^v$ with weight $\alpha_i > 0$ are the *support vectors*.

After several tests on the training sample, we decided to use a radial basis function kernel:

$$\mathbf{K}(x, y) = exp\Big(\frac{\|x - y\|^2}{\sigma^2}\Big) \tag{2}$$

Now, given these $\mathcal{N}$ trained SVMs where inputs are images visual features and outputs are concepts (image classes), we want to define a centroid $\vartheta(c_i)$ for each concept class $c_i$ that best represent it. These centroids should then minimize the sum of squares within each set $S_i$:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{\mathcal{N}} \sum_{x_j^v \in S_i} \|x_j^v - \mu_i\|^2 \tag{3}$$

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

The objective being to estimate a distance between these classes in order to assess their visual similarities, we compute the centroid $\vartheta(c_i)$ of each visual concept $c_i$ using:

$$\vartheta(c_i) = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j^v \tag{4}$$

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

$$\varphi(c_i, c_j) = \frac{1}{1 + d(\vartheta(c_i), \vartheta(c_j))} \tag{5}$$

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

### 3.2 Conceptual Similarity

Conceptual similarity reflects the semantic relatedness between two concepts from a linguistic and a taxonomic point of view. Several conceptual similarity measures have been proposed [5, 21, 1]. Most of them are based on a lexical resource, such as WordNet [11]. 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 [5]. However, as aforementioned, the structure of these resources does not necessarily reflect image semantics, and therefore such measures does not seem suited to our problem. An alternative approach to measure the semantic relatedness between concepts is to use their provided definition. In the WordNet case, these definitions are known as the glosses and are provided by the synsets associated to each concept. For

example, Banerjee and Pedersen [1] 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 [20], in which they suggest to exploit "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 $\overrightarrow{w}_{c_j}$:

$$\eta(c_i, c_j) = \frac{\overrightarrow{w}_{c_i} \cdot \overrightarrow{w}_{c_j}}{|\overrightarrow{w}_{c_i}||\overrightarrow{w}_{c_j}|} \tag{6}$$

Some concepts definitions in WordNet are very concise and thus make the measure unreliable. Consequently, [20] proposed extending the glosses of concepts with the glosses of adjacent concepts (located in their immediate neighborhood). Hence, for each concept $c_i$ the set $\Psi_{c_i}$ is defined as all the adjacent glosses connected to $c_i$ ($\Psi_{c_i}$=$\{$gloss($c_i$), gloss(hyponyms($c_i$)), gloss(meronyms($c_i$)), etc.$\}$). Then each element $x$ (gloss) of $\Psi_{c_i}$ is represented by $\overrightarrow{w}_x$ as explained above. The similarity measure between two concepts $c_i$ and $c_j$ is then defined as the sum of the individual cosines of the corresponding gloss vectors:

$$\theta(c_i, c_j) = \frac{1}{|\Psi_{c_i}|} \sum_{x \in \Psi_{c_i}, y \in \Psi_{c_j}} \frac{\overrightarrow{w}_x \cdot \overrightarrow{w}_y}{|\overrightarrow{w}_x||\overrightarrow{w}_y|} \ , \quad where |\Psi_{c_i}| = |\Psi_{c_j}|. \tag{7}$$

Finally, each concept in WordNet may match several senses (synsets) that differ from each other in their position in the hierarchy and their definition. A disambiguation step is then necessary to identify the good synset. For example, the similarity between "Mouse" (Animal) and "Keyboard" (device) differs widely from the one of "Mouse" (device) and "Keyboard" (device). Therefore, we first compute the conceptual similarity between the different senses (synset) of $c_i$ and $c_j$. The maximum value of similarity is then used to identify the most likely meaning of these two concepts, i.e. disambiguate $c_i$ and $c_j$. Thus, the conceptual similarity is calculated as following:

$$\pi(c_i, c_j) = \underset{\delta_i \in s(c_i), \delta_j \in s(c_j)}{\operatorname{argmax}} \theta(\delta_i, \delta_j) \tag{8}$$

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

### 3.3 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, 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, this one is a "corpus-dependent" measure, and more precisely depends on the distribution of concepts in the corpus.

In our approach, we define the contextual similarity between two concepts $c_i$ and $c_j$ as the Pointwise Mutual Information (PMI) $\rho(c_i, c_j)$:

$$\rho(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \tag{9}$$

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 cooccurrence 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 $c_i$ et $c_j$, the above probabilities can be estimated by: $\widehat{P(c_i)} = \frac{n_i}{\mathcal{L}}$, $\widehat{P(c_i, c_j)} = \frac{n_{ij}}{\mathcal{L}}$.

$$\Rightarrow \rho(c_i, c_j) = \log \frac{\mathcal{L} * n_{ij}}{n_i * n_j} \tag{10}$$

$\rho(c_i, c_j)$ 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 si $c_i$ et $c_j$ 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 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:

$$\gamma(c_i, c_j) = \frac{\rho(c_i, c_j)}{-\log[\max(P(c_i), P(c_j))]} \tag{11}$$

### 3.4 Semantico-Visual Relatedness of Concepts (*SVRC*)

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. Then, the Semantico-Visual Relatedness $\phi(c_i, c_j)$ of these concepts $c_i$ and $c_j$ is defined as:

$$\phi(c_i, c_j) = \omega_1 \cdot \overline{\varphi}(c_i, c_j) + \omega_2 \cdot \overline{\pi}(c_i, c_j) + \omega_3 \cdot \overline{\gamma}(c_i, c_j)\ ,\ \sum_{i=1}^{3} \omega_i = 1 \tag{12}$$

The choice of weights $\omega_i$ 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 hierarchy is rather to build a hierarchical framework to image classification, it may be advantageous to assign a higher weight to the visual similarity ($\omega_1 \nearrow$).

### 3.5 Heuristic Rules for Hierarchy Building

Once we have estimated the semantic relatedness between each pair of concepts, it is important to regroup 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:

$$Closest(c_i) = \underset{c_k \in \mathcal{C} \setminus \{c_i\}}{\operatorname{argmax}} \phi(c_i, c_k) \tag{13}$$

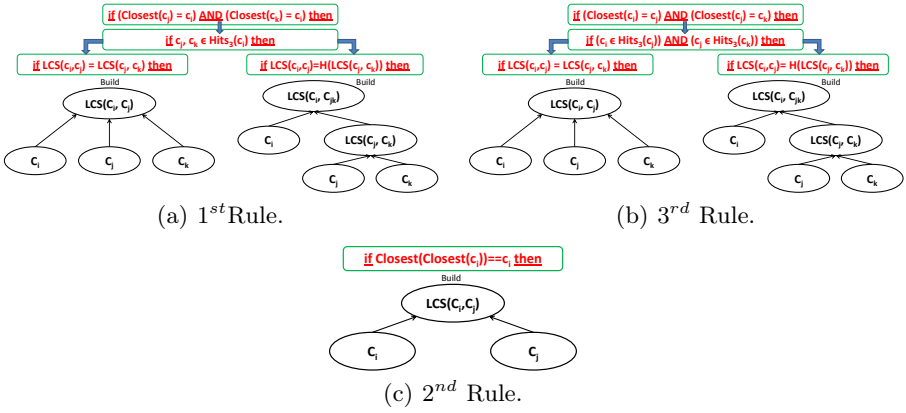
- $LCS(c_i, c_j)$ allows to find the *Least Common Subsumer* of $c_i$ and $c_j$ in WordNet:

$$LCS(c_i, c_j) = \underset{c_l \in \{H(c_i) \cap H(c_j)\}}{\operatorname{argmin}} len(c_l, root) \tag{14}$$

  where $H(c_i)$ allows to find all of 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 $c_x$ and *root*.
- $Hits_3(c_i)$ returns 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 reasoning about the Least Common Subsumer (LCS) to select concepts to be connected to each other. These rules are illustrated and executed in the order described in Fig.2. First rule checks whether a concept $c_i$ is classified as the closest relative to more than one concept ($(Closest(c_j) = c_i), \forall j \in \{1, 2, \cdots\}$). If so and if these concepts $\{c_j\}$ are reciprocal in $Hits_3(c_i)$, then according to their LCS they will be connected either directly to their LCS or in a tow level structure as illustrated in Fig.2(a). In the second, 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. The third rule covers the case when ($Closest(c_i) = c_j$) and ($Closest(c_j) = c_k$) - cf. Fig.2(b).

The building of the hierarchy is bottom-up (starts from leaf concepts) and uses an iterative algorithm until it reaches the root node. Given a set of tags




(a) $1^{st}$Rule. (b) $3^{rd}$ Rule.

(c) $2^{nd}$ Rule.

**Fig. 2.** Rules in *TRUST-Me* allowing to infer the relationship between the different concepts. Preconditions (in red) and actions (in black).

associated with 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*. 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 concepts are linked to a root node. Fig.3 illustrates the built hierarchy on Pascal VOC dataset.

## 4 Experimental Result

As part of this work, we evaluate our semantic hierarchy by comparing the performance of a flat image classification versus a hierarchical based one. Pascal VOC'2010 dataset (11 321 images, 20 concepts) is used for building the hierarchy and evaluating the classification.

### 4.1 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. The used BoF model is built as following: feature detection using Lowe's DoG Detector [17], feature description using SIFT descriptor [17] and codebook generation. The generated codebook is a set of features assumed to be representative of all images features. Given the collection of detected patches from the training images of all categories, we generate a codebook of size $D = 1000$ by performing k-means algorithm. Thus, each patch 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.

### 4.2 Weighting

As this paper aims to build a hierarchy suitable for image classification/annotation, we set the weighting factors in an experimental way as follows: $\omega_1 = 0.4$, $\omega_2 = 0.3$, and $\omega_3 = 0.3$. Our experimentations on the impact of weights ($\omega_i$) showed also that the visual similarity is more representative of concepts similarity, as it will be illustrated with the produced hierarchies in Fig.3.

### 4.3 Evaluation

To evaluate our approach, we used 50% of VOC images for learning concepts and the others for testing. Each image may belong to one or more of the 20 existing classes. For the flat classification we used $\mathcal{N}$ SVM one-against-all, where the inputs are the BoF images representations and outputs are the desired SVM responses for each image (1 or -1) - for details cf. Section 3.1. However Pascal VOC dataset is unbalanced, i.e. many concepts are represented by few hundred of images among the 11321 images in the database (much more negative data than the positive ones for many concepts). To overcome this problem we used cross-validation, taking at each fold as many positive as negative images. Hierarchical classification is made by training a set of ($\mathcal{N}$+$\mathcal{M}$) hierarchical classifiers consistent with the structure of the hierarchy in Fig.3. $\mathcal{M}$ is the number of new concepts created during the building of the hierarchy. For training the classifier of each concept in the hierarchy, we took all images of children nodes (of a given

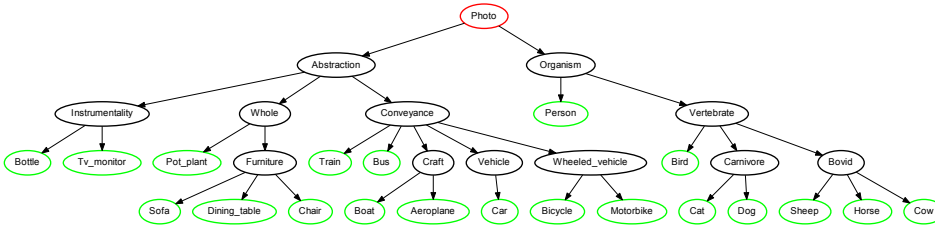



**Fig. 3.** The semantic hierarchy built on Pascal VOC'2010 dataset. Green nodes are original concepts, and the red one is the root of the produced hierarchy.

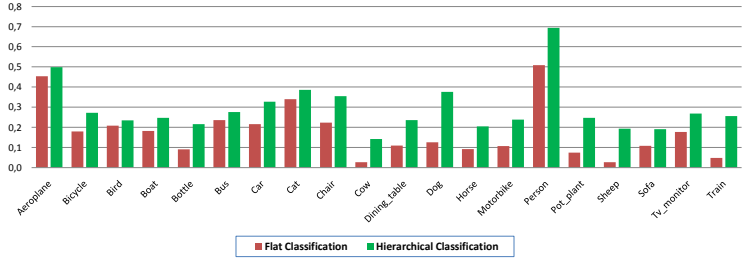
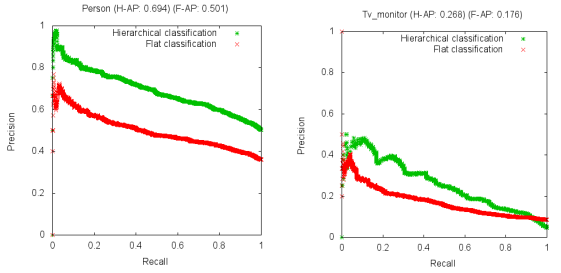



**Fig. 4.** Average precision of flat and hierarchical classification on Pascal VOC concepts.




(a) Concept Person. (b) Concept Tv_monitor.

**Fig. 5.** Precision/recall curves for hierarchical (green) and flat (in red) classification on concepts "Person" and "TV_Monitor".

concept) as positive and all images of children nodes of its immediate ancestor as negative. For example, to train a classifier for "Carnivore" all images of "Dog" and "Cat" are taken as positive while images of "Bird", "Sheep", "Horse" and "Cow" as negative. Thus, each classifier is trained to distinguish one class from others in the same category. For testing the hierarchical classification, a given image can take one (or more) path in the hierarchy based on classifiers responses, and starting from the root node until reaching a leaf node. Results are evaluated with the recall/precision curves and the average precision score.

Fig.4 compares the performance of our semantic hierarchic classifier with the performance of a flat classification. Our approach performs a better classification than the flat one, with a mean improvement of +8.4%. Using half of the training images from the VOC challenge (we have used the validation set for testing) and including the images marked as difficult, hierarchical classification achieves an average precision of 28.2% when the flat one achieves 19.8%. Fig.5 shows the recall/precision curves for concepts "Person" and "Tv_Monitor" using hierarchical and flat classification. This comparison shows that hierarchical classification has the best performance at all levels of recall.

## 5 Conclusion

This paper proposes a new approach to automatically build a suitable semantic hierarchy for image annotation. Our approach is based on a new measure of semantic relatedness, called *SVRC*, that takes into account the visual similarity, the conceptual and the contextual ones. *SVRC* allows estimating the semantico-visual relatedness of concepts. A new heuristic, *TRUST-ME*, is also proposed for reasoning about concepts relatedness, and to link together concepts that are semantically related in a semantic hierarchy. Our experiments showed that the built semantic hierarchy improves significantly the classification performance on Pascal VOC dataset. Our future research will concern the evaluation of our approach on larger datasets (MirFlicker and ImageNet), and the assessment of our hierarchy in terms of structure and contribution of knowledge.

## References


1. S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, 2003.
2. H. Bannour and C. Hudelot. Towards ontologies for image interpretation and annotation. In *CBMI*, 2011.
3. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.
4. E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, 2008.
5. A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32:13–47, March 2006.
6. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.
7. J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.
8. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
9. J. Fan, Y. Gao, and H. Luo. Hierarchical classification for automatic image annotation. In *SIGIR*, 2007.
10. J. Fan, H. Luo, Y. Shen, and C. Yang. Integrating visual and semantic contexts for topic network generation and word sense disambiguation. In *CIVR*, 2009.
11. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
12. G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *CVPR*, 2008.
13. A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *CIVR*, 2007.
14. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*. MIT Press, 2003.
15. L.-J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *CVPR*, 2010.
16. Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.
17. D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.
18. M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007.
19. M. Naphade, J. R. Smith, J. Tesic, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 2006.
20. S. Patwardhan and T. Pedersen. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *EACL*, 2006.
21. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, 1995.
22. J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, 2008.
23. A. W. M. Smeulders, S. Member, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE PAMI*, 22, 2000.
24. X.-Y. Wei and C.-W. Ngo. Ontology-enriched semantic space for video search. In *MULTIMEDIA*, pages 981–990, 2007.
25. L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *MM*, 2008.
26. B. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu. I2t: Image parsing to text description. In *Proceedings of IEEE*, 2009.