

Using Semantic Commonsense Resources in Image Retrieval

Adrian Popescu, Gregory Grefenstette, Pierre-Alain Moellic
Commissariat à l'Energie Atomique - LIST
{adrian.popescu, gregory.grefenstette, pa.moellic}@cea.fr

Abstract

Many people use the Internet to find pictures of things. When extraneous images appear in response to simple queries on a search engine, the user has a hard time understanding why his seemingly clear request was not properly satisfied. If the computer could only understand what he wanted better, then maybe the results would be more precise. We believe that the introduction of an ontology, though hidden from the user, into current image retrieval engines would provide more accurate image responses to his query. Coordinating the use of an ontology (OWL representation of WordNet) with image processing techniques, we have developed a system that, given an initial query, automatically returns images associated with the query by specializing the query concept using only its deepest hyponyms from the ontology. We show that picking randomly from this new set of images provides a better representation for the initial, more general query. In addition, we exploit the visual aspects of the images for these deepest hyponyms (the leaves of WordNet) to cluster the images into coherent sets. In this way we can present the results in a structured, and even ontologically labeled, manner to the user.

1. Introduction

Images now represent an important part of the information searched for on the Internet. There are at least two problems with current image search on the internet. First, even when simple and clear queries are formulated, the obtained results are often not representative of the search term. Secondly, there is no semantic structure in the responses offered by popular search engines such as Google, Yahoo, AlltheWeb, Picsearch. Current image retrieval systems are keyword based and make little or no use of image processing techniques. Their image related advanced options are limited to simple image related parameters like: file format, image size or type of picture (color or black and white) and there is no image-related semantic treatment. Even image repositories such as Flickr only use textual user-added tags to structure search results.

Google and Yahoo! propose combinations of keywords that include simple logical operators like AND, OR and NOT. Picsearch proposes a rudimentary ontology, grouping some 120 concepts into 6 higher order categories: animals, classic cars, flowers, landmarks and legend. Yahoo proposes specializations of the initial query, thus introducing some context for the initial concept. For the query *dog*, Yahoo! Image Search also suggests searching for *dog breeds*, *dog names*, *dog the bounty hunter*.

While simple and computationally efficient, the current approach to image search depends on the quality of the text found near images, or in their filename. This text unfortunately can sometimes have no direct relation to the image content.

Image media search has generated research centered on automatically generating keywords for images, such as [1]. In other work [9], [10], [17], predefined ontological knowledge plays an important role in automatic annotation. But, research in image indexing has recognized that the search vocabulary of the user has to correspond to the way that images are indexed [14].

Here we exploit WordNet to improve image retrieval results. WordNet [8] is a lexical resource intended to structure common knowledge into a semantic hierarchy. Initially built for psychological experimentation, and exploiting common dictionaries in its construction, its semantics was hoped to correspond to the way that people see the world. Using of this hierarchy, we hope that a commonly accepted model for organizing natural categories can be reflected in the responses proposed by the system.

In this paper we describe an approach to image-to-language association. We propose a system that uses this simple ontology, WordNet, and image processing techniques to automatically associate picture classes to concepts. In the process we build a large scale image catalogue, considering only picturable entities in our approach. The use of WordNet is two-fold: first, it provides the list of search words that are used to query the Web for images that constitute the raw data in our system, and second, after its transformation into an OWL [16] ontology, it constitutes a taxonomical base for our system. We exploit this taxonomy in the following way. Since specialized

concepts (leaves in the WordNet hierarchy) are usually less ambiguous than higher order concepts in WordNet, given a query term, we use the WordNet hyponyms relation to follow the taxonomy down to the concepts found at the leaf nodes. We then use these concepts to query the web rather than the original query term. For each image query result concerning a leaf node concept, we employ image processing techniques to index and cluster the raw data collected from the Web into visually coherent sets of images.

In Section 2 we describe related work and their relation with our approach. In Section 3 we discuss briefly present some ontological issues related to our work. Section 4 is dedicated to a presentation of the techniques we employed and of some problems encountered. Before concluding, we present and discuss preliminary evaluations of the performances of our system compared to those of a current search engine, AlltheWeb.

2. Related work

As proven by the increasing number of works in the field, exploring the frontier between image and language is an interesting and challenging task. In [1], the authors propose a tree of concepts and probabilistic methods in order to associate words from the tree to images. Their tree of concepts has more general concepts in higher positions and specific words in lower ones but there is no inheritance relation that governs the structure of the tree. From an image point of view, pictures are segmented and characterized in terms of color, texture. Image clustering is compared using three different methods: using only associated words, only visual aspects and combining the two methods and the authors state the last type of clustering gives the best results. The entire work is performed on pictures from the Corel database and we are not aware of the application of a similar method for larger image sets.

In [10], the authors describe M-OntoMat-Annotizer, a complete framework for multimedia documents annotation and an analysis is presented. They structure their system under the form of an ontology that is intended to represent and tie together both low level image descriptors (color, texture, shape), spatial relations in the image and linguistic concepts, concepts extracted from domain ontologies. The system requires the use of a sufficient quantity of images in which the association between a region in the picture and an object name has been manually performed. Given this learning set, the presented results are very good.

In [9], the author describes OntoVis, a domain model that addresses interior crime scenes. The system includes detailed models for a limited number of objects (20 in the current version). There is notably much description of partonomic relations and the 3D modeling of the given objects. Unfortunately, no quantitative evaluations of OntoVis are reported in this paper. Given the level of detail

of the modeling and the associated efforts, it would be hard to extend the approach to much larger domains.

One important difference between our approach and those of [1], [9] and [10] is that we work with images that cover all picturable entities represented in a commonsense knowledge manner while they work on relatively narrow domains. Moreover, what separates the present from those in [1] and [9] is we use raw data from a highly unstructured resource, the Web. We do not aim at finely modeling any particular domain. Consequently, the level of detail of ontological knowledge in our approach is smaller than that in [9] and [10].

In [17], the authors propose the constitution of an image thesaurus using images on the Web. They extract weighted key terms from the text around the image and try to match these keywords to regions in the images. Both low-level descriptors (color and texture information are used) and high level, linguistic concepts are integrated in the system and, consequently, image and keyword queries are supported. Taxonomic relations in WordNet are used to expand queries for given concepts and to filter word senses. Wang's approach is closely related to ours. Both approaches are aimed at constructing image catalogues using raw data collected from the Web. Though we both use different gathering processes, a common point is the use of WordNet. One key difference between Wang's system and ours is our exclusive use of leaves in the hierarchy to collect data from the Web while the authors of [17] use keywords on several levels. Another important difference in Wang's system is that, for polysemic terms (concepts appearing in more than one synset), they retain the first sense only, losing one important advantage offered by the WordNet structure: sense separation. We preserve sense separation in a different way explained below, using query expansion in order to differentiate word senses for ambiguous terms.

An important distinction between our approach and all those described above is that our technique does not imply a learning phase, a time consuming step that becomes critical when working with large data collections.

3. Ontological issues

In this section, we describe some ontology related aspects that are relevant for our work. We justify the choice of WordNet as taxonomical base for the current application and propose a way to separate picturable concepts from the others. We equally discuss current methods in automatic ontology creation.

3.1. Hierarchies

The "IsA" relation is fundamental to the way people organize entities in the world. We currently dispose of a few comprehensive hand built systems based on this relation (i.e.: WordNet in lexicography, Cyc [4] in formal

ontologies). We are aware that it is probably illusory to attempt to construct a hierarchy that performs best in all situations. The choice of one particular hierarchy is directed by the envisioned application. Since we wish to respect commonsense knowledge in our system, it is desirable to use a resource that accounts for the way people organize entities in the world. WordNet2.1 seems a good choice. This is organized as a tangled hierarchy and covers most common English terms. The root concept is *entity*. There are two relations that are fundamental in WordNet, “IsA” and synonymy. The first structures the hierarchy in depth while the second gives rules for constituting its basic units, the synsets. A synset includes one or more terms that describe the same entity. Ambiguity can be resolved by attaching a sense number to all defined terms. Thus, each modeled entity is uniquely described by a WordNet synset.

3.2. Picturable objects

One way to separate high order categories is to distinguish between *nominals*, *biologic concepts* and *artifacts* [6]. The first do not correspond to physical entities in the world and there are no coherent pictural representations of such categories. It would be very hard to imagine a set of pertinent pictures for *truth*, *reason*, or even *association*. Given this situation, we do not associate images to nominal concepts in WordNet. *Biologic concepts* and *artifacts* however are picturable categories and we can think about constructing image classes that properly represent the associated linguistic concepts.

There are differences between the categorical distinctions in [6] and those in WordNet but these differences do not affect the distinction of entities in picturable or not. We decided to associate picture clusters only to concepts ranged under *physical entity* in the WordNet hierarchy. This is an initial choice and it is an open discussion if we should further restrain the categories to which we associate image classes. We think that, for example, subconcepts of *process*, *physical process* like *iteration*, *looping* or *irreversible process* probably do not have a coherent visual representation though they are found as hyponyms of the concept *physical entity*.

3.3. Automatic ontology building

Manual construction of taxonomies is a very time consuming process, especially when we deal with large quantities of data. There exists an important current in ontology engineering that addresses problems related to the automatic construction of ontologies [2]. The standard procedure is to process specific domain textual documents and develop hierarchies for the relevant domain concepts. When one wants to cover broad domains, this approach is, for the moment, impractical. One other possible solution is to reuse existing resources. We adopted this last idea and automatically transformed the WordNet nouns hierarchy

into an Ontology Web Language (OWL) ontology [16]. Our version is similar to that of [14] but differences arise given that we translate for an identified goal: the use of WordNet in image retrieval tasks. The two versions are not contradictory and, given that the translation described in [14] emerged from an official Semantic Web task force, further work will include alignment of our translation to the official one.

4. Visual ontology construction

After describing some aspects related to translation of the WordNet hierarchy, we give here a brief description of the image clustering module in Subsection 4.2. Section 4.3 discusses some problems raised by our approach and by proposing possible solutions.

4.1. WordNet nouns hierarchy

Our translation of the WordNet nouns hierarchy into an OWL format did not include *instance-of* synsets, resulting in 73733 OWL classes, rather than 81246 total synsets. There are about 60000 leaves in the entire hierarchy. We decided that leaf concepts under *physical entity* (the great majority of leaves) would be used. Studies on concept representation [12] show that specialized concepts offer a good visual coherence, which led us to consider the most specialized nodes, the leaf nodes, as possibly providing the most coherent images in visual search, and to take their union for non-leaf nodes..

Since WordNet is a representation of commonsense knowledge, from a domain specialist’s point of view, the ontology is far from complete. For example, the *placental* hierarchy in WordNet contains only 1112 synsets whereas other knowledge sources, vi, include nearly 600 entries for *dog* breeds alone. Nevertheless, WordNet gives us a good coverage of common words for our image search application. To illustrate the level of detail of the knowledge contained in WordNet, figure 1 shows an excerpt from the noun hierarchy for *bear* and its hyponyms. Observe that *bear* includes immediate subtypes such as *ice bear* or *brown bear*. *Brown bear* in turn is subdivided in *Syrian bear*, *grizzly* or *Alaskan brown bear*. We use the most specialized concepts (leaf nodes like *grizzly* or *ice bear* and their equivalents in the respective synsets) to obtain corresponding pictures from a popular picture search engine. The image results for *bear*, an internal node, are obtained by merging the results for all its leaf subtypes.

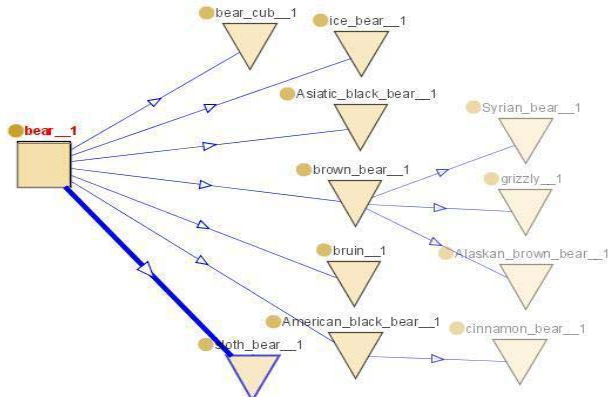


Fig. 1 Subtypes of *bear* in the ontology.

4.2. Image clustering

Described elsewhere [18], we have developed a clustering tool that takes a textual query, fetches Internet images and furnishes clustered pictures as results. We exploited this clustering program to create our large pictorial image dictionary in the following manner. First, we created a dictionary entry for each leaf synset under *physical entity* in WordNet. The dictionary entry is indexed by the same name as in the OWL hierarchy we created. For example, the entry *grizzly_1*, corresponds to the synset *grizzly*, *grizzly bear*, *silvertip*, *silver-tip*, *Ursus horribilis*, *Ursus arctos horribilis*. Secondly, to collect images for this synset, search engine queries for each term in the synset were generated and launched. The union of all gathered images will be connected to the ontology class *grizzly_1*. But before attaching the images, we apply our clustering tool using a border/interior pixel classification algorithm [3] designed to index images from broad domains as we consider here. This tool clusters the indexed images using a *k* – SNN (*k* - Shared Nearest Neighbors) algorithm [13] in which the similarity between two images is related to the number of neighbors they share. The clusters are formed around a group of core images that possess the best connectivity with respect to other images, called aggregated images. This algorithm fits to our purposes as it is flexible: it does not impose a fixed number of picture clusters, or a fixed number of images in each class and, equally important, not all elements in the raw data set have to belong to a class.

The main role of the image clustering algorithm is to group images into visually coherent sets. In figure 2, some raw image data for *grizzly*, obtained by querying the Alltheweb search engine is shown. In figures 3 and 4, we present two image clusters obtained for the same concept after clustering the first 1000 images returned by AllTheWeb using visual characteristics.



Fig. 2 Raw AllTheWeb image data for the query *Grizzly*.



Fig. 3 Image cluster 1 for *Grizzly* using visual characteristics of retrieved images.

Comparing images from figure 2 and figure 3, we observe that the images in figure 3 are thematically structured while the raw images of figure 2 do not present the same kind of coherence. Generally, the clustering step performs well and groups together visually similar images. The clustered images are finally attached to the dictionary entry, and then attached to the OWL hierarchy.

Polysemy in the hierarchy is dealt using query expansion [5]. For ambiguous terms, the associated entry in the leaf dictionary is formed of the term and its immediate hypernym. *Angora* is a polysemic term in WordNet and the associated image queries will be *Angora domestic cat*, *Angora goat* and *Angora rabbit*, each one situated in its original place in the hierarchy.

4.3. Problems encountered

There are two types of problems we encountered: image processing related ones and keyword related. One of its important actual limitations comes of our system from the fact that current image indexers are not capable to perform reasonable outside very narrow domains. As a

consequence, we are not capable to improve precision in the image clusters when passing from the raw data to clustered data. This is to say that, currently, there is no improvement of precision for leaf concepts in the hierarchy.

Another problem is related to the fact that not all leaves are equally represented on the Web. Generally, we obtain hundreds of images for each leaf class in the ontology, but the variation between concepts is large. There are concepts that are represented by very few images or, at worst, not at all represented on the Web, but these are rare cases.

Lexically, the main problems encountered are related to meanings of words that are not included in WordNet and to annotations that are not related to image content. As it uses only specialized terms, usually less ambiguous than higher order ones, our approach is less sensitive to noise than current search engines which return images with terms at all levels in the ontology. But we are not able to treat cases where a meaning of a term is not included in the hierarchy. For example, there are a lot of dolls that represent a *cinnamon bear* and their pictures on the web are annotated with the name of the animal. Similar problems appear for concepts with high symbolic value, like pets.

5. Evaluation

Our main purpose in introducing an ontology into the image search problem and using the leaf nodes as surrogate terms for general queries is to improve the precision in the image sets associated to non-leaf concepts in the WordNet hierarchy. We describe here a limited evaluation of the results obtained with our approach against the results obtained using a commercial search engine, AllTheWeb.

5.1. Experimental settings

As shown in [5], specialized terms do not appear frequently when people are asked to describe image content. They prefer using higher level concepts. For this reason, we decided to test our system using four general languages searches: *bear*, *cat*, *dog*, and *dolphin*.

We decided that a relevant image must include a representation of the category in such a manner that a human should be able to immediately associate it with the assessed concept. We used an independent human evaluator to judge our results, a person who was not previously informed about the methods we used to construct the collections of pictures corresponding to each concept. The images presented to the evaluator during the test were randomly selected from the two picture sets, one derived from the results of using the original query term on AllTheWeb, and the other set derived by our method of using only the hyponyms of these terms to query the web.

5.2. Results

In table 4, we present the results of the evaluation. The values in the table correspond to the percentage of positive assessments in each of the random 200 image sets. The last line contains a mean of the precision for the two approaches.

	Precision[%]	
	Concept merging	Alltheweb query
Bear	44,5	31
Cat	55	46
Dog	77	62,5
Dolphin	52	38,5
Mean	57,1	44,5

Table 4. Evaluation results comparing image results using a general term and using its hyponyms instead.

In all cases, using the hyponyms of general concepts to find images rather than the general class words improve precision. We obtain a mean improvement of the precision of 12.6% with a minimum of 9% for “cat” and a maximum of 14.5% for *dog*. Significant improvement is obtained for all four evaluated categories. It is interesting to note that we obtain the ordering of precision for concepts for the two evaluated methods.

With the use of subtypes to represent higher order concepts, we obtain better representative images of these concepts. Moreover, using the ontology and the clusters associated to leaf concepts, we have the additional advantage of being able to present the results in a structured and visually coherent manner, using the ontology labels as well as the clusters.

Any improvements in semantic image filtering before clustering would, of course, improve current results for both sets. Linguistic filtering might consist in the creation of adaptive dictionaries of undesirable terms that could be excluded during the web search. Co-occurrence tests between these concepts and the queried WordNet leaf could be performed for the text surrounding the image. If both terms would appear, the image should be eliminated. Image filtering might also consist of elimination of pictures containing human faces and text, when searching for non human objects. Dealing with finding response on the web, we are mainly concerned with precision, so the elimination of pictures prior to clustering is harmful only for the rare cases where we get a very small number of answers from search engines. These situations are easy to detect and if they occur, the filters can be switched off.

6. Conclusions

We have presented techniques for automatically associating images to terminal and non terminal nodes in a large scale ontology. Hypothesizing that a joint use of this semantic resource and of image processing techniques can improve image retrieval, and we have shown in a small test that we can improve precision in the image sets associated to general concepts in the ontology. Transforming WordNet nouns hierarchy into an OWL ontology, we used the leaves under picturable concepts to gather raw data from a classical image search engine and indexed the raw data. The indexed images were clustered to provide visually coherent image classes associated to leaf concepts in hierarchy. As to the pertinence of the using leaf node labels compared to more general terms, we compared our results to those obtained by querying image search engines for four familiar concepts and showed that the use of our technique produces improves results for these general terms.

Future work shall primarily concentrate on the implementation of the filtering techniques mentioned in Subsection 5.3. We are also interested in the use of other semantic resources in image-language association efforts. Namely, we want to expand current queries with knowledge from a large scale semantic network, ConceptNet [7] in order to add context (e.g.: localization, related categories) to concepts in WordNet. The system would then be able to respond to more complex queries.

Another future direction is the constitution of a multilingual ontology using WordNets in other languages [11]. The inclusion of other WordNets will allow a multilingual annotation of the images in the catalogue, allowing a user to formulate queries in all the languages in the ontology.

7. References

[1] K. Barnard, and D. Forsyth, "Learning the Semantic of Words and Pictures", In *Proc. of ICCV 2001*, Vancouver, Canada, 2001, pp. 408-415.

[2] P. Cimiano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web3, In *Proc. of WWW 2004*, Manhattan, NY, USA, 2004, pp. 462-471.

[3] L. Ertöz, M. Steibach, and V. Kumar, "Finding Topics in Collections of Documents. A Shared Nearest Neighbor Approach" In *Clustering and Information Retrieval*, Kluwer, 2003.

[4] R. V. Guha, and D. B. Lenat, "Cyc: A Midterm Report", *AI Magazine*, 11, 3, 1990, pp. 32-59.

[5] L.Hollink, G.P.Nguyen, D.Koelma, A.Th.Schreiber, M.Worrying. "Assessing User Behaviour in News Video Retrieval", *IEE proceedings on Vision, Image and Signal*

Processing, 152/6, December 2005, pp. 911-918.

[6] F. C. Keil, *Concepts, Kinds, and Conceptual Development*, Bradford Books, 1992.

[7] H. Liu and Singh, P., "ConceptNet: A Practical Commonsense Reasoning Toolkit", *BT Technology Journal*, 22, 4, Kluwer Academic, 2004, pp. 211-226.

[8] G. A. Miller, "Nouns in WordNet: a Lexical Inheritance System", *Intl Journal of Lexicography*, 3, 4, 1990, pp. 245-264.

[9] K. Pastra, "Image – Language Association: are we looking at the right features?", In *Proc. of Workshop on Language Resources for Content-based Image Retrieval, LREC 2006*, Genoa, Italy 2006, pp. 40-44.

[10] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, Y. Kompatsiaris, and S. Staab, "Knowledge Representation and Semantic Annotation of Multimedia Content", *IEEE Proceedings on Vision, Image and Signal Processing*, 153, 32, June 2006, pp. 55-262.

[11] E. Pianta, L. Bentivogli, C. Girardi, "MultiWordNet: developing an aligned multilingual database", In *Proc. of the 1st Intl Conference on Global WordNet*, Mysore, India, 2002.

[12] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories", *Cognitive Psychology*, 8, 1976, pp. 382-439.

[13] R. O. Stehling, M. A. Nascimento, and A. X. Falcao, "Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification", In : *Proc. of CKIM 2002*, Mc Lean, USA, 2002, pp. 102-109.

[14] T. Thlivitits and I. Kanellos, "SEMINDEX: A Human-Directed Textual Indexing Of Image Content" HCP'99, Brest, Sep 1999, pp.257-263

[15] M. van Assem, A. Gangemi, and G. Schreiber, "RDF/OWL Representation of WordNet", <http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619>, June 2006.

[16] W3C, "OWL Web Ontology Language Overview", www.w3.org/TR/owl-features/, 2004.

[17] X. J., Wang, W. Y. Ma, and X. Li, "Data-driven Approach for Bridging the Cognitive Gap in Image Retrieval", In *Proc. of ICME 2004*, Taipei, Taiwan, 2004, pp. 2231-2234.

[18] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. Hède, and P.-A. Moellic, "Clustering and semantically filtering web images to create a large scale image ontology", In *Proc. of IS&T/SPIE 18th Symposium Electronic Imaging*, San Jose, California, USA, 2006