“Tell me more”: how semantic technologies can help refining internet image search

Francesco Setti, Daniele Porello and Roberta Ferrario
Istituto di Scienze e Tecnologie Cognitive (ISTC)
Consiglio Nazionale delle Ricerche (CNR)
via alla Cascata 56/C - Trento, Italy
{francesco.setti,daniele.porello,roberta.ferrario}@loa.istc.cnr.it

Sami Abduljalil Abdulhak and Marco Cristani
Department of Computer Science, University of Verona
Cá Vignal 2 - Verona, Italy
{njasbd20,marco.cristani}@univr.it

ABSTRACT

Several branches of computer vision heavily rely (but we could even say depend) on the availability of large datasets of labelled images. While such labeling is usually done by hand, a powerful help can be obtained from Internet and its related tools. In this paper we address the problem of automatically generating a set of images representing an object class, given the name of the class. We exploit semantic technologies, such as lexical resources and ontologies, in order to improve the search performances by using a standard web search engine. We will also discuss an application to the automatic building of a training set for a classification framework. Preliminary experiments are provided for 10 classes from the public CalTech256 dataset and results show an average increment in classification accuracy of about 10%.

Categories and Subject Descriptors

E.1 [**Data**]: Data Structures—*graphs & networks*; H.2.8 [**Information Systems**]: Database Management—*image databases*; H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval—*query formulation*

Keywords

Internet image search, lexical resources, ontologies

1. INTRODUCTION

While a large set of computer vision topics are strongly related to the availability of good quality large image datasets, at the state of the art the task of creating these datasets is still performed by hand. The main reason resides in the fact that building a set of images of a specific object eventually falls in a standard classification problem, which requires, in turn, a proper training set, and this is a *chicken&egg* problem. Thus, the main limitation in automatically building large image datasets of class specific objects is that, while automatic object detection is a very well known and widely studied problem in computer vision and pattern recognition, the recognition of a general object given its name is still an open issue.

Humans are able to perform this task very well, due to the fact that human brain is an amazing device, able to extract a lot of key information from very few data. We can quickly learn the appearance of a new object class from a limited number of images, sometimes even only from a single one, by combining weak sources, focusing on the most interesting features and, particularly, integrating a huge amount of priors coming from our experience. Moreover, our visual and cognitive systems work in tight connection; so we can also discriminate from different objects by associative mechanisms (first you don’t see the chair very well, but you see a familiar shape close to a table and you end up focusing on it and seeing it better) or by inferring information from the context (you have never seen before the object on the table, but you are in the office of an architect, so it should be some kind of tool used for drawing). Implementing these capabilities into a machine is very hard; but, on the other hand, a machine is able to process an incredibly huge amount of data in a short time.

Most of the state of the art methods for object recognition are eventually addressing a classification task, based on learning class descriptors from a specific training set. Although often neglected, the choice of the training set is a crucial point affecting the performances of the detection method. A good training set has to be sufficiently informative to capture the nature of the object under analysis, but at the same time has to be generic enough to avoid overfitting and to cope with new instances of the object of interest.

Few years ago, a competition started posing a new challenge: autonomous mobile robots were located in a room and asked to find a set of objects listed in a text file. The robots had Internet access to download a set of images used to train visual classifiers. It followed an environment exploration phase in which the robots were looking for the objects, running the trained classifiers. This is known as the

VIGTA ’13 San Petersburg, Russia

Semantic Robot Vision Challenge [13]. The proposers of this challenge were evidently aware of the potentialities involved in learning from Internet data. When images are uploaded in the web, they are usually included in websites that also contain portions of written text more or less connected with such images. Moreover, the files containing these images are very often given a name and sometimes users tag the images with more or less relevant information. Webpages are indexed based on all this information, in order to enable search engines to provide the users with information that is relevant relatively to definite queries. The same holds for images, they are also indexed based on textual information and thus, if such textual information is not precise enough, the results of the query do not correspond to what the user was looking for. This is the reason why what usually happens when we perform an image search on the Internet, the top rank results are fairly reliable, while lower ranked results are not, an example of this is provided in Figure 1. How to gather a greater amount of reliable images is the central issue of this work.

We can metaphorically think at the training set as a set of images stored in our brain's memory, so we can think the robot as a human who tries to extract from his/her memory a set of images related to the object, infers the useful information to build a model of it and then performs the detection task. In this case we can easily see how we are missing part of the story: human memories are connected in various ways through associative and inferential mechanisms used to retrieve relevant and useful information. But how can we humans recover information that is stored in our memory, e.g. how can we remember the image of something that we have seen in the past? What most psychological theories tell us is that chunks of memory are connected, and we are able to exploit such connections with associative and inferential cognitive mechanisms. In fact, our brain not only extract a set of images related to the object to search, but it is also able to give us some additional information and relation between different instances of the same object or its properties. Our approach, described in this paper, is an attempt of imitating human cognitive mechanisms by leveraging on information connected to the images on the Internet, in particular to textual information.

The main aim of this paper is to provide a smart way to automatically build a good training set by using images downloaded from the Internet. Our idea is to exploit lexical resources (such as WordNet [9]) and ontologies (such as DOLCE [16] and its application to WordNet OntoClean [12]) in order to automatically associate related concepts and generate a set of words connected to the name of the target class, i.e. the object to look for. These combinations of words are then used as keywords in an image search engine operating on the web. The main purpose connected to the use of semantic technologies is that of refining the results of the image search, by exploiting semantic connections between related words.

In order to evaluate the performances of our algorithm we also implemented an object detection framework. It exploits standard dense SIFT descriptors used in a bags-of-words framework and one-class Support Vector Machine classifier. We tested our proposed algorithm with a set of 10 objects classes, consistently giving improvements in terms of classification accuracy, for an average of about 10%. We will show how the lexical and semantic properties of the object name strongly relates to the performances of the classifier.

The rest of the paper is organized as follows: Sec. 2 presents the state of the art and related works; Sec. 3 presents some reflections about the introduction of lexical resources and ontologies; Sec. 4 illustrates the method presented in this study, while Sec. 5 displays the results of the experiments and, finally, Sec. 6 concludes the paper and draws the lines for future directions of research.

2. RELATED WORK

The first works using Internet for image classification followed the straightforward way to directly use the top ranked images provided by a web search engine for classifier learning [19, 20]. Unfortunately all web search engines are based on *non-visual* features, and in particular they provide images ranked based on the tags assigned by the users and the textual content of the web page in which the image is located. This falls in highly variable images, with a large fraction being unrelated to the query term, posing a challenging learning problem.

Figure 1 reports some example images taken by using Google image search engine with the keyword 'bear'. Over the 100 top ranked images, we got 16 unrelated images representing toys, people or different (often fantasy) animals. In particular, none of the first 10 images and only 6 over the top 50 were bad images.

A series of attempts have been made to overcome the limitations directly related to the web search engine. Berg and Forsyth [2] tackles the problem by using the text on the original web pages to extract contextual cues. They apply this method to gather large sets of animal images from the web, although the system is not completely automatic but requires a human to perform some tasks. Later, Schroff et al. [18] makes the previous method fully automatic with a two steps approach: they first re-rank the images based on text and other metadata, then they learn visual models from the highly ranked images by using a Support Vector Machine (SVM) classifier.

A different approach is used by both [3] and [21]; they use multiple-instance learning techniques to overcome the labeling noise in Internet images. In particular, the first work focuses on handling few positive instances to train the classifier, while the second focuses on handling noisy data.

Fergus et al. [10] uses an approach derived from the probabilistic latent semantic analysis (pLSA) technique for text document analysis, that is able to automatically select a subset of "good" images used to learn object models; while Li & Fei-Fei [14] extends a similar method in an incremental fashion to compile a dataset of a desired class from the Internet.

Collins et al. [6] uses an active learning approach to rapidly build up a large dataset from Internet images, using a human-in-the-loop with the recognition model.

A slightly different kind of works focuses on the joint learning of text and images. Barnard et al. [1] presents a method where models are learned from both visual and textual (labels) features. In this method each image is oversegmented using normalized cuts to give a large number of regions. The regions are then represented by vectors encoding low-level concepts such as color and area. The vectors from each image are modeled jointly with the text labels, establishing a correspondence between the two. Hence, in a recognition scenario, given one the other can be predicted. Carbonetto

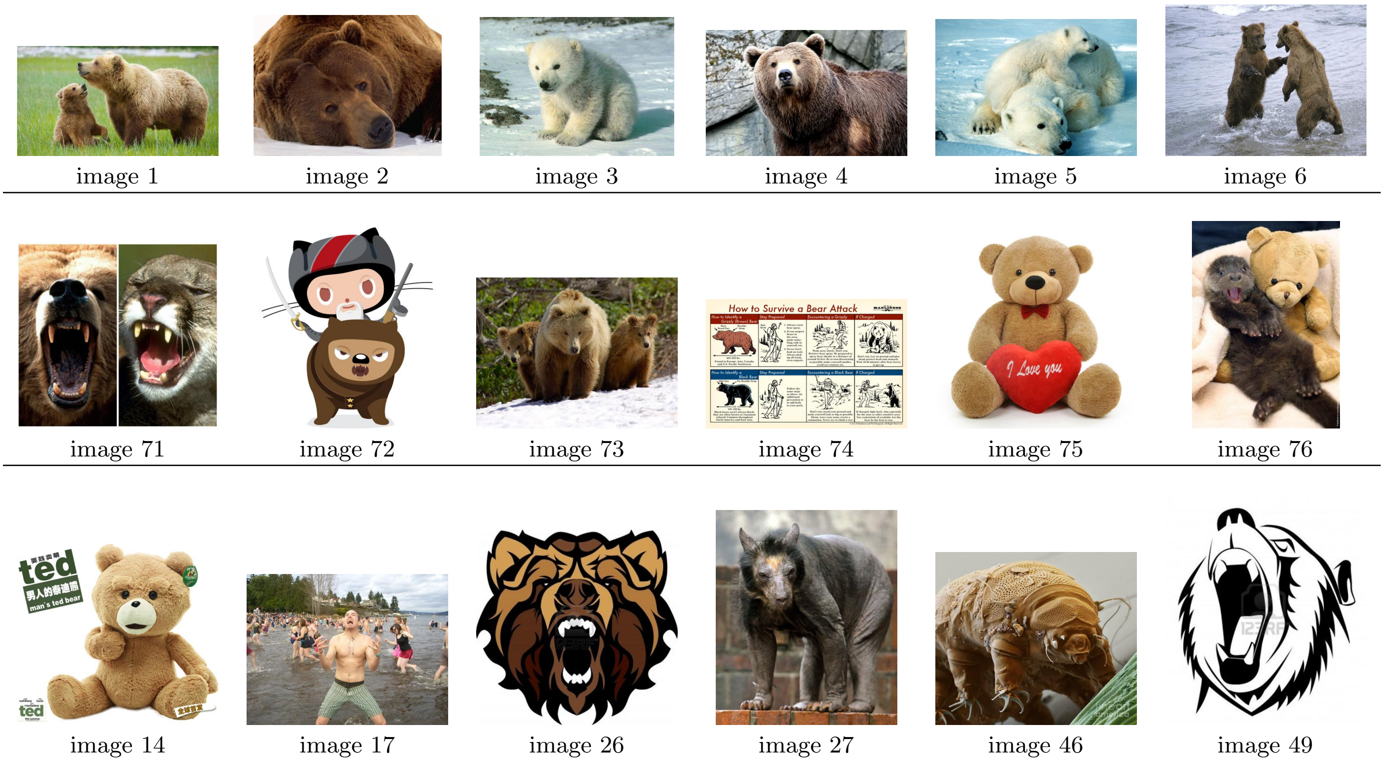

image 1 image 2 image 3 image 4 image 5 image 6

image 71 image 72 image 73 image 74 image 75 image 76

image 14 image 17 image 26 image 27 image 46 image 49

Figure 1: Example of images retrieved by Google Image search engine by using 'bear' as keyword. The first row represents the 6 top ranked images, the second the images ranked 71th to 76th, while the third row shows completely unrelated images ranked within the top 50.

et al. [4] also considers the text and images problem but the authors use sparse kernel methods to determine sets of features related to each object class. Both these last works assume that training images are provided with a reliable and fairly rich text annotation, which is unfortunately not usually the case for images gathered from the Internet.

3. HOW TO REFINE INTERNET SEARCH THROUGH THE USE OF SEMANTIC TECHNOLOGIES

As already mentioned in the Introduction, the issue we are trying to address is that of building a good training set of images for an objects' recognition task by downloading possibly relevant images from the Internet. Image search engines heavily rely on the images' tags (most of the times assigned by simple users) and in general on textual information accompanying the image. These are the reasons why, when we use as keyword for the search the name of the object we are looking for, in general the highest ranked images correctly display such object, while lower ranked images display scenes that are somehow (and often unintelligibly) connected with the object. Finally, they also explain why it is much more difficult to retrieve correct images when the name of the object is heavily polysemous. In order to refine the results of the search and to exclude from the training set undesired images, we propose to leverage on the application of semantic technologies.

When we, human users of the Internet, want to search for an object whose name is polysemous, we usually try to disambiguate by adding to the search a related term, that helps by adding a sort of context and thus restricting its possible interpretations. If we want to perform the same step, but in an automatic manner, a reasonable strategy is, in our opinion, to look at semantic technologies. In particular, since what we are looking for at this stage are terms, lexical resources, like WordNet, appear as good candidates. In WordNet, synonyms are grouped into unordered sets (*synsets*). Each of the $117,000$ synsets is linked to other synsets by means of a small number of *conceptual relations*. In the case of nouns, which is what we are interested in at present, these relations are: *hyperonymy*, which relates a more general class to a specific one; *hyponymy*, a more specific class to a general one; *meronymy*, basically parts. Thus WordNet can be visualized as a network, whose nodes are synsets and whose arcs are semantic relations. This means that, if we are interested in refining the search of an object denoted by a noun, we take the synset that includes that noun and associate to it a noun contained in a node that is reachable by the previous one just traversing one of the arcs that depart from its node.

A further thing to add is that we humans know from experience that, if we want to refine the search when looking for images, the keyword that we add to the one indicating the searched object should preferably be relevant from a visual point of view, as, usually, when people tag images, they do that by adding words describing what is depicted and visible in the image. So, taken the results produced by the search through WordNet, we should find a way to prune away those terms that do not refer to something visible. In the current paper, we have performed this step manually, but a long term goal is certainly that of accomplishing even this filtering activity automatically, and, to do so, it is essential to have a rich enough representation of the object we are looking for, in which we can express statements at a fine-grained level, even statements about properties themselves, like “the property *being furry* is a visible one”, so that we can decide that it is worth to add the term “furry” to “cat” to refine the search, while “being a mammal” is not a visible property, so we don’t expect to get better results by adding such term to the search.

Our proposal is to take a foundational ontology, like DOLCE, and identify which are the properties that are visible and predicate such meta-property on the basic properties. If we then build domain ontologies describing our objects of interest based on this foundational and “visibility-aware” ontology, we can take the terms extracted by WordNet and give them to the ontology reasoner, so that it can select only those terms that are connected to a concept or a property that is visually relevant.

In other words, the selection of the accompanying terms which are visually relevant that we have made manually in the present paper could – and in our opinion should – be performed automatically with the use of a computational ontology. Furthermore, even if in the current work we are only concerned with the retrieval of training images on the Internet, the same ontology could be used, for example by the robot of the SRVC contest, for then searching the real object in the environment.

4. METHOD

In this section we will present our innovative procedure to build a class specific image set automatically from the Internet by exploiting three components: a **lexical resource** (i.e. WordNet [8, 17]), to generate a set of keywords connected to the name of the object class; an **ontology**, to select a subset of “visually meaningful” words within the previously generated set; and an **image search engine**, to find images on the Internet. A schematic representation of the method is shown in Figure 2, while a detailed explanation of each step follows.

4.1 Lexical Search

The lexical search we propose to adopt is based on WordNet [17]. This is a large lexical database of English where terms are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. While in WordNet different kinds of words are used (nouns, verbs, adjectives and adverbs), for the purpose of this paper we will only focus on nouns. WordNet differs from a thesaurus in that whereas in a thesaurus words are grouped only based in their similarity of meaning, words in WordNet can be connected by various semantic relations, as hyperonimy, hyponimy, meronimy etc.

Thus, given a noun in english, WordNet can automatically generate three distinct sets of related words, connected to the original one by means of three basic conceptual relations, the just mentioned hyperonimy, hyponimy and meronimy.

4.2 Ontology Search

Given the original word (the name of the object class) and the set of related words generated in the lexical search step, the ontology is called to filter out all the related words that do not involve any visual meaning. For instance, within meronyms we can have internal or external parts: thinking about a generic species of animals, while legs are external parts, usually visible, that can give a contribution to the characterization of the class, heart or stomach are internal organs that do not provide any information about the shape of the object. In this example, the ontology would tell that the terms “heart” and “stomach” have not to be accepted as keywords for the search, while the term “legs” should be.

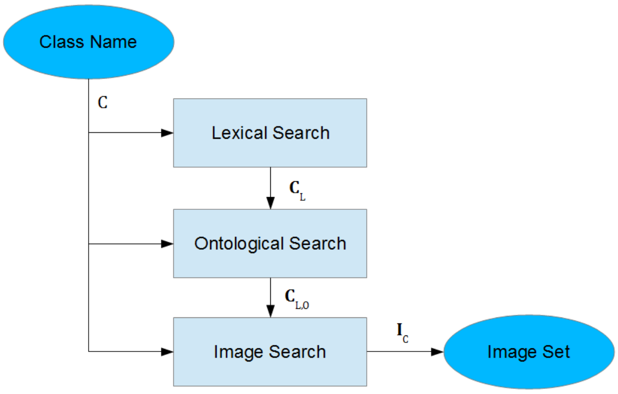

Figure 2: Schematic representation of the proposed algorithm. Given the class name C, the lexical search provides a set of candidate additional words $\mathbf{C}_L$, ontological search selects only the words with a visual meaning $\mathbf{C}_L, O$, while an internet image search engine is used to download images $\mathbf{I}_C$ by using these words as keywords.

4.3 Image Search

Given the name of the object class C and the set of connected words $\mathbf{C}_{L,O}$ we can perform an internet image search with any web search engine which allows image specification. The core of our algorithm is to use as keywords for the search engine the set of i composed keywords $K(i)$ defined as the pairs $K(i) = C + \mathbf{C}_{L,O}(i)$, where the symbol $+$ is the string concatenation operator.

5. EXPERIMENTS

Since a direct evaluation of performances for the dataset generation task is hard to define, we present in this section an indirect evaluation based on a bag-of-words [7] classification framework with SIFT features [15] and one-class Support Vector Machine (SVM) [5].

We performed our experiments on a set of 10 classes taken from the CalTech256 dataset [11]: *airplane*, *bear*, *frog*, *goat*, *goose*, *guitar*, *owl*, *snake*, *socks* and *zebra*. For each class, a list of connected words has been automatically generated with WordNet and a subset of them has been selected to simplify the tests. Although in general all the conceptual relations (i.e. hyperonymy, hyponymy and meronymy) can be useful for our purpose, in our preliminary experiments we used only the hyponymy relation. In future works we are going to extend the method by including also hyperonymy and meronymy relations in the search. The list of all the classes and the additional words used in the experiments

Class name	Additional words
airplane	airliner, fighter, jet
bear	cub, ice, polar
frog	robber, tree, toad
goat	wild, nanny, billy
goose	blue, chinese, gosling
guitar	acoustic, classic, electric
owl	horned, screech
snake	colubrid, elapid
socks	anklet, athletic, tabi
zebra	Equus, mountain

Table 1: List of all the additional words used in our experiments.

are shown in Table 5.

As testing set, we considered all the images provided by CalTech256 dataset for each class taken into account. Only the class *airplane* has been reduced to 100 images randomly selected. Two training sets of 100 images per class have been automatically generated by downloading the top ranked images from the Internet by using Google image search. For the first one, we used the class name as keyword, while for the second, the keywords were generated by concatenating the class name with each additional word (see Sec. 4).

Each image in training and testing sets has been resized to 300 pixels of maximal dimension (proportions have been kept) and processed by extracting dense SIFT with patch dimension of 20×20 pixels and step of 10 pixels, resulting in an average of about 700 SIFT descriptors per image. The images have been then described in terms of bag-of-words by means of a universal codebook provided by the ImageNet Large Scale Visual Recognition Challenge 2011 (ILSVRC2011), generated by processing over millions of images with the same parameters we used and clustering SIFT descriptors into 1000 words. Finally, one-class linear SVM algorithm has been used for the classification of testing images. The classification performances are considered in terms of accuracy, defined as the percentage of correctly detected objects over the total number of testing images.

Table 5 shows how we are better performing for most of the classes, with an average improvement of about 10%. As expected, the method we are proposing is much more effective for classes whos name is general and can be specified by a small number of hyponims: this is the case of the class *guitar*, since guitars can have very different shapes (particularly the electric ones) but they can be split in only three categories: electric, acoustic and classic guitars. When the original class name is already very specific, we can only partially improve results (as for *zebra*). On the other hand, some classes like *socks* and *snake* have so many different hyponims that our experiments can only take into account a small subset of them, generating a training set not general enough to show big improvements. A particular case is the class *goose*, which generates a very good training set even in the original case.

In order to explore the contribution of the additional words in the classification procedure, we also ran experiments with a mixed approach. We tested 11 different compositions of the training set by means of a parameter R, which represents the portion of images in the training set taken by using the class name and the additional words (equally splitted) as keyword with respect to the total number of images in the training set; the rest of the images are taken by only using the class name. This parameter can vary from 0, which means only the class name is used as keyword, to 1, only the composed keywords are used. Results are shown in Figure 3.

Class	# images	simple search	our method
airplane	100	0.72	0.83
bear	102	0.76	0.91
frog	116	0.84	0.91
goat	112	0.83	0.92
goose	110	0.96	0.92
guitar	122	0.52	0.84
owl	120	0.84	0.92
snake	112	0.87	0.88
socks	112	0.72	0.77
zebra	96	0.92	0.94
Average	–	*0.79*	*0.89*

Table 2: Detection accuracy for training set generated by simply searching for the class name (simple search) and by using connected keywords (our method).

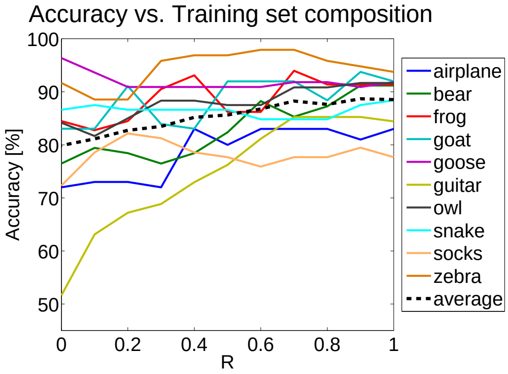

Figure 3: One-class detection accuracy chart when varying the composition of the training set by means of the parameter R. (Best viewed in colors)

Two main considerations can be drawn: first, our method is consistently improving the detection performances, with a stronger increment in accuracy for classes that perform poorly in the standard case. Second, the best performances are usually achieved with $R = 0.9$; which means, considering only composed keywords can reduce the generality of the training set. This is probably due to the fact that in our experiments we used only a small number of additional words, and thus just a subset of all the hyponyms for each class.

6. CONCLUSIONS

The Internet is the hugest repository of images, always freely available. On the other hand, unfortunately, since images are uploaded with a wide variety of purposes, their indexing criteria are not really clear and search engines leverage on the text accompanying the images to try and retrieve results that are as correct as possible. In order to gain independence from available datasets for selecting good training sets of images, we have tried in this paper to figure out a procedure that could allow to build a training set for a task of object recognition, by downloading in a principled way images from the Internet. The procedure we have proposed uses lexical resources for singling out terms that are semantically connected to the term naming the object one is interested in; to this it follows a filtering phase, in which, among these firstly selected terms, only those that refer to objects or properties that can be visualized in an image are kept. This phase has been performed manually in the present paper, but the use of a foundational and "visibility aware" ontology is foreseen, in order to automate the pruning. Finally, a standard search engine is run using as keywords the term denoting the object of interest associated to other terms, that are semantically related and visibility-relevant. The results of the experiments show a considerable improvement in performances, especially for some categories that performed badly in the single term search. We believe that this is a very promising line of research, worth of pursuing, especially if we succeed in automatizing each step of the procedure and this is in fact the direction towards which the present paper represents a first step.

7. ACKNOWLEDGMENTS

F. Setti, D. Porello, and R. Ferrario are supported by the VisCoSo project grant, financed by the Autonomous Province of Trento through the "Team 2011" funding programme.

8. REFERENCES

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

[2] T. L. Berg and D. A. Forsyth. Animals on the web. In *CVPR*, volume 2, pages 1463–1470, Washington, DC, USA, 2006. IEEE Computer Society.

[3] R. C. Bunescu and R. J. Mooney. Multiple instance learning for sparse positive bags. In *ICML*, pages 105–112, New York, NY, USA, 2007. ACM.

[4] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, volume 3021 of *Lecture Notes in Computer Science*, pages 350–362. Springer Berlin Heidelberg, 2004.

[5] Y. Chen, X. S. Zhou, and T. Huang. One-class svm for learning in image retrieval. In *ICIP*, volume 1, pages 34–37, 2001.

[6] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. In *ECCV*, pages 86–98, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. A. Bray. Visual categorization with bags of keypoints. In *ECCV workshop*, pages 1–22, 2004.

[8] C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[9] C. Fellbaum. Wordnet and wordnets. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford, 2005. Elsevier.

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from internet image searches. *Internet Vision*, 98(8):1453–1466, 2010.

[11] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

[12] N. Guarino and C. Welty. An overview of ontoclean. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 201–220. Springer Berlin Heidelberg, 2009.

[13] S. Helmer, D. Meger, P. Viswanathan, S. McCann, M. Dockrey, P. Fazli, T. Southey, M. Muja, M. Joya, L. Jim, D. G. Lowe, and A. K. Mackworth. Semantic robot vision challenge: Current state and future directions. In *IJCAI workshop*, 2009.

[14] L.-J. Li and L. Fei-Fei. Optimol: Automatic online picture collection via incremental model learning. *Int. J. Comput. Vis.*, 88(2):147–168, 2010.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[16] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. Wonderweb deliverable d18. Technical report, CNR, 2003.

[17] Princeton University. Wordnet online. `http://wordnet.princeton.edu/`, May 2010.

[18] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[19] A. T. Setz and C. G. M. Snoek. Can social tagged images aid concept-based video search? In *ICME*, pages 1460–1463, Piscataway, NJ, USA, 2009. IEEE Press.

[20] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel. Learning automatic concept detectors from online video. *Comput. Vis. Image Underst.*, 114(4):429–438, 2010.

[21] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization. In *CVPR*, 2008.