Semantic Hierarchies for Visual Object Recognition

Marcin Marszałek (marszale@inrialpes.fr) **Cordelia Schmid** (schmid@inrialpes.fr)

INRIA, LEAR / LJK, Grenoble, France

Abstract

We propose to use lexical semantic networks to extend state-of-the-art object recognition techniques. We use the semantics of image labels to integrate prior knowledge about inter-class relationships into the visual appearance learning. We show how to build and train a semantic hierarchy of discriminative classifiers and how to use it to perform object detection. We also demonstrate additional features that become available to object recognition due to the extension with semantic inference tools—we can classify high-level categories, such as animals, and we can train part detectors, for example a window detector, by pure inference in the semantic network.

1. Introduction

Motivation: Most of the existing object recognition systems are trained to recognize only a few categories and can not infer high-level concepts.

Contribution: We address two important limitations for constructing vision systems which deal with a large number of categories:

- inter-class similarities and relationships need to be modeled
- the complexity in the number of object categories has to be reduced

Reasoning example: A *car* and a *motorcycle* are *wheeled vehicles*, therefore both incorporate a *wheel*. Whenever one sees a *car* or a *motorcycle*, one sees a *wheeled vehicle* and should also see a *wheel*.

Method:

1. WordNet is queried with the class labels and knowledge is extracted in form of a semantic hierarchy
2. The hierarchy is used for reasoning and to organize and train the binary SVM detectors
3. The trained hierarchic classifier allows to efficiently recognize a large number of object categories

Advantages:

1. The prior semantic knowledge about object identity is integrated into the visual recognition system, which
   - can help to learn the visual appearance of new object types
   - speeds up the recognition process
2. The reasoning allows to
   - learn new concepts by semantic inference
   - return richer recognition answers

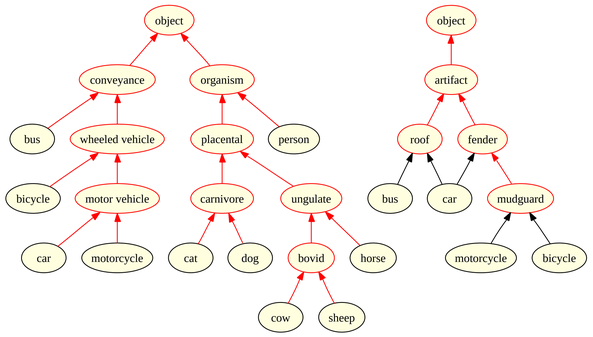
Sample PASCAL VOC'06 images classified with our method.

2. Extracting the semantic hierarchy

WordNet: One of the most popular semantic networks for English language. It models human psycholinguistic knowledge by

- grouping words into sets of synonyms called *synsets*
- recording different semantic relations between synsets

We exploit *hypernymy*/*hyponymy* (*is kind of*) and *meronymy*/*holonymy* (*is part of*) relations.



(a) Full hypernymy subgraph (b) Some shared meronyms

WordNet 2.1 subgraphs for the PASCAL VOC'06 labels. Intermediate nodes were removed for clarity.

Algorithm:

1. Using the WordNet index, the most probable meaning (synset) can be found for each class label; synsets model concepts and are represented with nodes in our semantic graph
2. For each synset a set of semantic links to other synsets can be retrieved; semantic links are represented with directed edges in our semantic graph
3. The hypernymy and meronymy links are followed until a full WordNet subgraph is extracted

3. Constructing the hierarchic classifier

The binary classifier: Image classification approach of Zhang et al. (IJCV'07):

1. Scale-invariant *Harris-Laplace* and *Laplacian* detectors are used to extract salient image regions
2. Appearance-based features are computed for those regions with *SIFT* and *hue* color description
3. Bag-of-features image representation
4. Classification is performed by SVMs with χ^2 kernel

Semantic reasoning: Let us consider images (exemplars) supporting a given concept.

- Trivially, the exemplars that represent the concept will support it
- Due to semantic reasoning, each node of the semantic graph is additionally supported by the union of the exemplars supporting the nodes that point to it through hypernymy or meronymy links

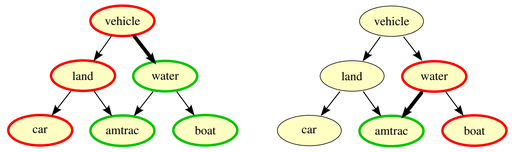
$$\mathrm{supp}(A) = \bigcup_{B_i \to A} \mathrm{supp}(B_i) \cup \mathrm{lbl}(A)$$

where $\mathrm{supp}(A)$ is a set of exemplars supporting the A concept, $B_i \to A$ is true when B_i links to A through hypernymy or meronymy and $\mathrm{lbl}(A)$ is a set of exemplars labeled with the A concept.

Training: $B_i|A$ classifier associated with the $B_i \to A$ edge is trained with

$$P = \mathrm{supp}(B_i) \qquad N = \mathrm{supp}(A) - \mathrm{supp}(B_i)$$

where P is the set of positive exemplars and N is the set of negatives.



Positive (green) and negative (red) support for training of edge classifiers.

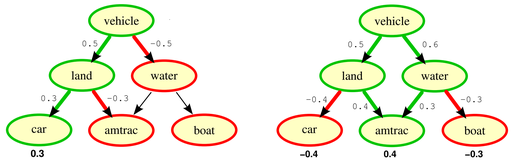
Testing:

- The *base node* is supported by all the exemplars in the dataset
- The hyponymy and holonymy links from A to B_i are descended when the detector associated with the $B_i \to A$ link returns a positive answer

One can combine the decision functions of the underlying binary classifiers. For each concept c a decision function $h_c(x)$ can be defined:

$$h_c(x) = \max_{P \in \mathcal{P}(s,c)} \min_{e \in P} g_e(x)$$

where $\mathcal{P}(s,c)$ is the set of all possible paths from the base synset (starting node) s to synset c, P is an element of this set, e is an edge on the path P and $g_e(x)$ is the decision function of the classifier for the edge e.



Decision functions of edge classifiers may be combined during recognition.

4. Complexity estimate

Estimated number of binary classifiers evaluated for a test sample:

$$T(n) = \frac{c}{a} T\left(\frac{n}{b}\right) + c$$

where n is the number of classes, c is the number of binary classifiers evaluated at a node, a is the subproblem selection factor (c/a defines the number of subproblems that have to be solved) and b is the problem reduction factor (n/b defines the size of the subproblem).

In the case of the PASCAL VOC'06 dataset, on average:

- $c = 2.85$ subproblems examined (binary classifiers run) per node
- one of every $a = 1.94$ subproblems was descended
- for each descent the size of the problem was reduced $b = 1.82$ times

Thus, the estimated complexity of the classifier (for similar datasets):

$$T(n) \in \Theta\left(n^{\log_b(c/a)}\right) \approx \Theta\left(n^{0.64}\right)$$

when $\log_b(c/a) > 0 \Leftrightarrow c > a$ which is true. This is significantly better than $\Theta(n)$ required in a one-against-rest setup with n classifiers.

5. Experimental results

Dataset: PASCAL VOC'06 challenge dataset, classification task

Comparison of the EERs for PASCAL VOC'06.

		baseline		our SH		
		OAR	AVH	SSH	ESH	gain
A	bicycle	79.3%	80.0%	81.4%	**82.8%**	**3.4%**
	bus	90.4%	90.4%	**91.6%**	**91.6%**	**1.2%**
	car	87.9%	87.6%	**88.9%**	88.3%	**0.3%**
	cat	**82.5%**	**82.5%**	80.4%	80.4%	-2.1%
	cow	82.9%	**84.8%**	81.9%	81.9%	-1.0%
	dog	76.4%	**78.7%**	77.0%	77.5%	**1.1%**
	horse	**80.7%**	78.2%	79.8%	79.8%	-0.8%
	motorcycle	84.2%	83.3%	**85.0%**	83.3%	-0.8%
	person	75.1%	73.0%	75.1%	**75.7%**	**0.5%**
	sheep	82.6%	81.8%	**84.1%**	**84.1%**	**1.5%**
	average	82.19%	82.02%	**82.52%**	**82.53%**	**0.34%**
B	conveyance	*89.8%*	*88.4%*	**90.4%**	**90.4%**	**0.6%**
	organism	*76.2%*	*82.1%*	**87.7%**	**87.7%**	**11.5%**
C	window	*62.5%*	*62.5%*	-	**65.8%**	**3.3%**

Baseline:

- OAR - One Against Rest
- AVH - Automatic Visual Hierarchy

Our SH:

- SSH - only hypernymy
- ESH - both relations

Discussion:

A. Results for classifying the ten classes.
   - Our approach leads to slight improvement of performance compared to the methods that do not use the semantics
   - The average EER of the winning challenge method was 86.4%, while our method achieves 82.5% with half of the training images

B. Detecting high-level concepts when training with the original labels.
   - The semantic hierarchic structure of our classifier provides sensible answers in situations of uncertainty, when the precise object identity may be unclear, yet the high-level identity can be determined
   - Comparing ESH to OAR, a gain of 11.5% shows that our classifier goes beyond straightforward reasoning and can successfully detect a living creature concept

C. To further test the generalization ability of our classifier we have collected 120 *vehicle window* images for the positive set and 120 *machine* images for the negative set. Our classifier trained in the original setup:
   - could generalize over the windows of cars and buses
   - was detecting individual windows of different vehicles
   - resulted in some false positives on window-like structures

(a) true positives (b) false positives

Sample images classified by our method as containing a *window*.

6. Summary

- We have proposed a semantic hierarchic classifier that uses the semantics of image labels to extract knowledge about the inter-class relationships and integrates it into the visual appearance learning procedure
- Using semantic hierarchies reduces the classifier complexity in the number of classes and helps to learn the visual similarities
- Our classifier returns valuable information in situation of uncertainty and can learn new classifiers through inference