# Learning Hierarchical, Semantic Priors for Image Boundary Detection

Anthony Hoogs and Roderic Collins
GE Global Research
One Research Circle
Niskayuna, NY 12309
{hoogs,collins}@research.ge.com

Image segmentation is typically performed without knowledge of visual categories, higher-level context or semantics. Recently, learning-based methods have shown that segmentation can be improved by training on human-generated segmentations [5, 4]. However, incorporating the high-level knowledge that humans appear to use in segmentation is not usually considered.

In our previous work [3], we investigated how a large semantic ontology, WordNet [2], could be used for image segmentation. The underlying premise is that a true boundary in an image corresponds to a semantic difference (or "distance") between the objects on either side of the boundary. Hence image edges with a large visual difference, such as high contrast, can be disregarded when the semantic distance is small, and edges with a small visual difference but large semantic difference can be emphasized. The semantic distance across an edge is computed using WordNet by first training WordNet with visual exemplars for many concepts. The semantic distance between two concepts is computed as the weighted graphical distance, where the weights are established using prior probabilities of concept frequency in training.

The image is segmented into regions using a standard visual algorithm such as mean-shift [1], and each region is visually compared to all concepts in WordNet using a feature vector of color and texture. The top $K$ matches are kept for each region $R_i$, forming set $S_i$ of concepts. Since the region segmentation partitions the image, each edge is bounded by exactly two regions $R_1$ and $R_2$. The average semantic distance is computed between all concept pairs $S_1 \times S_2$, and the edge is weighted by this value (after scaling) to produce a continuously-valued boundary map.

Here, we explore the relationship between visual learning and structured, ontological knowledge in the domain of image understanding and segmentation. The framework of WordNet with visual attribution on each concept is constructed using 920 training images from the Berkeley Segmentation Database (BSD) [6], which provides manual segmentations. 100 additional images are held out for testing. We have extended the database with WordNet concept labels for each image segment (region). The enhanced database provides information about the co-occurrence of semantic concepts in images, as well as spatial relationships such as which concepts tend to be adjacent. This database is unique as far as we know, as other databases of labeled segmentations do not label every pixel in every image [7].

The 920 images contain 645 unique labels, distributed among 18 regions per image on average. WordNet includes a taxonomy of noun senses, so that a child concept is a sub-type or *hyponym* of its parent. All noun senses in common English are included - more than 80,000. We reduce the graph to those labels observed in the 920 training images, and their ancestors. The hierarchical structure of WordNet nouns enables semantic generalization in visual learning, which is not possible from visual information alone or from a flat list of concepts. In particular, we learn the following priors:

- The probability $P(B|C_i, C_j)$ that two concepts, or any of their child concepts, share a boundary;
- The probability $P(G|C_i, C_j)$ that two concepts have a spatial relation $G$, such as above/below, left/right, containment, etc.;
- The probability $P(C_i, C_j|I)$ that two concepts will appear in the same image $I$;
- The (marginal) probability $P(C_i|I)$ that a concept will appear in an image $I$;
- The probability $P(D|C_i, C_j)$ that the boundary between two concepts, or any of their child concepts, can be detected automatically;
- The conditional probability $P(E|C_i)$ that a detected edge is a true boundary, given a concept.

The first four priors are based only on the ground truth segmentations and labels, while the last two also depend on computed segmentations. $P(E|C_i)$ is the proportion of detected edges on the boundary of a concept in the training images, and is useful for identifying concepts with lots of clutter edges. The hyponym hierarchy induces constraints on the probabilities, so that, for all of these priors, the probability of a concept is the sum of its children's probabilities plus its own.

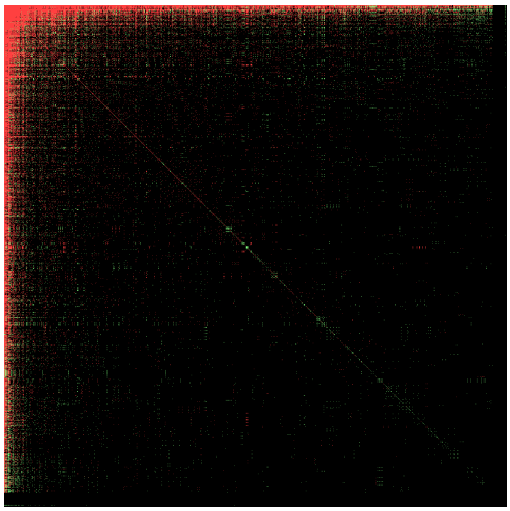

**Figure 1. Co-occurrence matrix for the 1238 WordNet nodes in the tree induced by the 645 training labels. Each row and column is a label in the tree; a red dot indicates an edge shared by the two labels in a training image, while a green dot indicates the nodes appear (but do not share an edge) in a training image. The axes are ordered by number of shared edges; the upper left corner is the root of the tree. Note how co-occurrence falls off as one moves away from the root, indicating the high branching factor in the tree.**

The first and third priors are illustrated in Figure 1, in which each concept pair is a point colored red if they share an edge, and green if they occur in the same image. Sorting the axes by total shared edge count exposes the hierarchy by placing the root in the upper left corner while leaf nodes spread out to the right and bottom.

Without priors, the edge weight used in [3] is:

$$P(E) = \frac{1}{2} - \frac{2}{\pi K(K-1)} \sum_{i \in S_1, j \in S_2} \tan^{-1}(\frac{D_{i,j} - \mu}{\sigma}) \tag{1}$$

where $D_{i,j}$ is the semantic distance between concepts $C_i$ and $C_j$, $\mu, \sigma$ are normalization mapping parameters, and the sum excludes duplicate pairings (hence the normalization is not $K^2$).

The priors are used to condition P(E). Including the priors for two concepts sharing a boundary, the boundary being detectable, and the detected edge being a boundary for each class, the boundary weight becomes:

$$P(E) = \frac{1}{2} - \frac{1}{\pi} \sum_{C_i \in S_1, C_j \in S_2} P(B|C_i, C_j) P(D|C_i, C_j) P(E|C_i) P(E|C_j) \tan^{-1}(\frac{D_{i,j} - \mu}{\sigma}) \tag{2}$$

Although the range of concepts is very large compared to the training set, the learned priors are still useful for improving image segmentation. Results are computed on the test images in the Berkeley Segmentation Database.

## References


[1] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[2] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[3] A. Hoogs and R. Collins. Object boundary detection in images using a semantic ontology. In *Proceedings of the National Conference on Artificial Intelligence*, 2006.

[4] J. Kaufhold and A. Hoogs. Learning to segment images using region-based perceptual features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2004.

[5] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using brightness and texture. In *Neural Information Processing Systems Conference*. MIT Press, 2002.

[6] D. Martin and J. Malik. The ucb segmentation benchmark database. URL: www.cs.berkeley.edu/projects/vision/grouping/segbench/.

[7] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. Technical report, MIT, 2005.