Hierarchical Scene Annotation

Michael Maire[1]
mmaire@caltech.edu
Stella X. Yu[2]
stellayu@icsi.berkeley.edu
Pietro Perona[1]
perona@caltech.edu

[1] California Institute of Technology
1200 East California Blvd
Pasadena, CA 91125

[2] UC Berkeley / ICSI
1947 Center St. Ste. 600
Berkeley, CA 94704

Supervised datasets play a central role as standards against which to benchmark long-term progress in computer vision. Over the past decade, the PASCAL [2] and Berkeley segmentation datasets (BSDS) [3] have filled these roles for the object detection and image segmentation tasks, respectively. The type of annotation available for each dataset determines the particular visual subtasks to which it is applicable. Object bounding boxes can benchmark detection algorithms, but are of limited use for training or evaluating segmentation. Segmented objects are more widely useful, but more time-consuming to annotate. What visual tasks are most important and what level of annotation detail is appropriate?

We present an alternative to thinking about dataset annotation in terms of a restricted set of visual tasks. Our key observation is that a hierarchical groundtruth representation, in the form of a ***doubly ordered region tree***, allows one to subsume disparate aspects of image labeling into a single framework. Specifically, we capture a nearly complete description of any scene in terms of objects, parts, object-part containment, segmentation, and figure-ground or occlusion ordering. Figure 1 illustrates the type of detail our annotation model encompasses for a typical scene.

Our unifying abstraction regards a scene as a set $S = \{R_1, R_2, ..., R_n\}$ where each $R_i \subseteq I$ is a region in the image I. In general, $R_i \cap R_j$ may be nonempty. It then organizes regions $\{R_i\}$ into a tree T. Let $N(R_i)$ denote the node of T corresponding to region R_i. $N(R_j)$ is set to be the parent node of $N(R_i)$ iff: (1) $R_j \supset R_i$, (2) R_j and R_i have an object-part relationship, and (3) $\nexists R_k : R_j \supset R_k \supset R_i$ and R_j, R_k, R_i have an object-part-subpart relationship, respectively. If for R_i, no region R_j satisfies all three conditions, then we set $N(R_i)$ to be a child of the root node. Simply stated, T decomposes the the scene into a multilevel object-part hierarchy.

We exploit one additional degree of freedom within T: the order $O(\cdot)$ in which nodes appear beneath a common parent encodes local occlusion relationships. Given sibling regions R_i and R_j such that $R_i \cap R_j \neq \emptyset$, then R_i occludes R_j if $O(R_i) < O(R_j)$, and R_j occludes R_i if $O(R_i) > O(R_j)$. If $R_i \cap R_j = \emptyset$, then they do not occlude one another and we disregard their relative ordering. Given T and O, preorder tree traversal uses the object-part hierarchy to translate local occlusions into a global figure-ground ordering. It also recovers a groundtruth ultrametric contour map (UCM) [1] weighting visible boundaries by the structural importance (object, part, subpart) of the regions they enclose.

Figure 2 shows a partial scene tree for a self-occluding object. Our full paper describes extensions for representing loops (shirt both behind and in front of arm). To mitigate the cost of creating rich groundtruth, we introduce a ***web-based annotation tool*** with a graphical interface for managing the region hierarchy. Our software eliminates tedious tracing of region boundaries through a dynamic paintbrush that snaps to the shape of underlying superpixels in a precomputed oversegmentation. Combined with a touch-up mode, it guides fast creation of pixel-perfect regions.

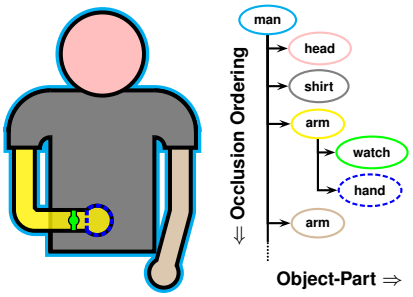

Figure 2: Our model of a scene or object groups pixels into regions and maps regions to nodes in a tree. Parent-child links denote region containment and semantic object-part relationships. Relative ordering of sibling nodes resolves occlusion ambiguities.

We use our annotation tool to groundtruth a dataset of 97 photographs previously used in experiments measuring the importance of objects in complex scenes [4]. Similar in size to the original BSDS test set, but with greater diversity in scene type and wider range of object scale, our labeled image set provides an object-centric context in which to benchmark segmentation algorithms. Instead of binary groundtruth boundary maps, as used in the BSDS, we have ***real-valued hierarchical groundtruth*** in the form of UCMs generated by scene tree traversal.

This permits direct evaluation of the degree to which the hierarchical output of state-of-the-art machine segmentation algorithms, such as gPb-UCM [1], respects the groundtruth object-part hierarchy. Specifically, we examine the order in which boundaries are recalled as one varies an algorithm's boundary detection threshold. The ideal algorithm would recover the occlusion boundaries between the top-level objects in the scene first, followed by large-scale object-part boundaries, followed by finely detailed part-subpart boundaries.

Our ***benchmark*** reveals a mismatch between gPb-UCM and the groundtruth object-part hierarchy. This performance gap serves as a call for further research into object-centric hierarchical scene segmentation and underscores the importance of our annotation tool and dataset.

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[2] Mark Everingham, Luc van Gool, Chris Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 2010.

[3] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2001.

[4] Merrielle Spain and Pietro Perona. Measuring and predicting object importance. *IJCV*, 2010.

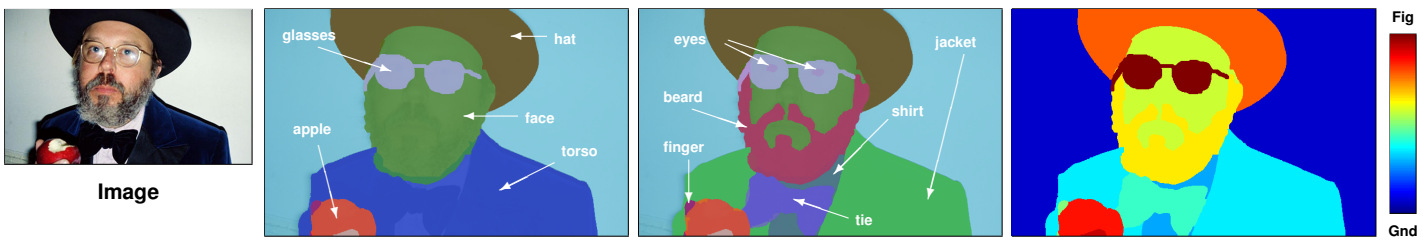

Figure 1: At the coarsest level, this scene contains two objects: a man standing in front of a wall. Looking in detail, we subdivide the man into regions for his hat, glasses, face, torso, and the apple he holds. In even more detail, his face consists of eyes, skin, and beard, and his torso is covered by a shirt, tie, and jacket. Traditional segmentation datasets label a single two-dimensional region partition. Our annotation model captures the object-part-subpart decomposition, as well as the occlusion relationships (*e.g.* apple in front of jacket, glasses in front of face) present in the scene.