Object Instance Sharing by Enhanced Bounding Box Correspondence

Santosh K. Divvala
santosh@cs.cmu.edu
Alexei A. Efros
efros@cs.cmu.edu
Martial Hebert
hebert@ri.cmu.edu

The Robotics Institute,
Carnegie Mellon University
Pittsburgh, PA. USA.

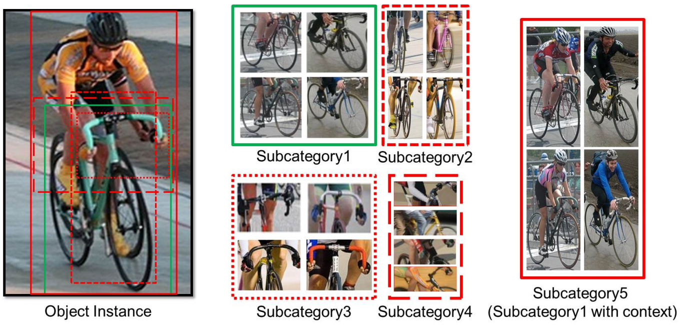

Figure 1: (left) A bicycle instance with its ground-truth bounding box shown in solid green. (center) Four (of the 25) subcategories discovered by our approach (few sample instances within each subcategory are shown). We allow the bicycle instance to be used multiple times with different bounding box representation for training the subcategory models. The different bounding box extents used per subcategory model are color coded accordingly e.g., subcategory3's match is shown using red dotted box, subcategory4's match shown in red dashed box, etc. (right) Subcategory1 shown after adaptively enlarging the bounding box to include local contextual cues around it.

Consider the task of building a sliding-window object detector. The standard learning-based approach is to first turn each human-labeled bounding box into a feature vector using some feature descriptor, e.g. HOG, and then train a classifier, e.g. SVM, on a stack of these feature vectors to discriminate them from the rest of the visual world. This is a reasonable strategy for older datasets, such as "INRIA person", where object instances are largely in correspondence, i.e. aligned such that each feature vector dimension has the same visual meaning for all object instances. However, modern datasets, such as PASCAL VOC, are much less restricted and do not guarantee good correspondence, with often huge variations between annotated bounding box instances.

The way modern approaches usually tackle this problem is by using mixture models [1, 2]. The idea is to somehow segregate instances within a category into disjoint groups (subcategories) and then train separate classifiers for each such subcategory. Each subcategory has reduced appearance diversity (via improved alignment), leading to a simpler learning problem. The recent success of the discriminatively-trained mixture model framework of Felzenszwalb et al., [1] has led to a wide popularity of such models for object detection. While reasonable, this assumes that a lot of training data is available for each subcategory. But this is often not the case, especially for occluded/truncated instances.

Consider the image shown in Figure 1(left). The human-labeled "bicycle" bounding box is indicated by the solid green box. Given this ground-truth framing for the object instance, it is most similar to instances in the "45°-view bicycle" subcategory, so, in a standard mixture-model detector, it would be assigned to subcategory1. However, by relaxing the bounding box framing for this instance, subregions of it can also match to the other subcategory models (subcategory2, subcategory3, subcategory4) as shown using the red bounding boxes. Furthermore, looking *outside* the bounding box might also allow us to capture consistencies in the local context surrounding the object, discovering new subcategories such as 'person riding a bicycle' (subcategory5).

What we propose in this paper is the idea of *training data reuse*. Conceptually, we would like to allow different object subcategories to be able to share (subregions of) each others training instances by providing *extra* correspondences between instances that were not part of the original human-supplied bounding box annotations. We operationalize this by two complementary operations: bounding box shrinking, which aims to find subregions of an instance that could be shared (Figure 2); and bounding box enlarging, which aims to create new subcategories by enlarging instances to include their local context (Figure 3).

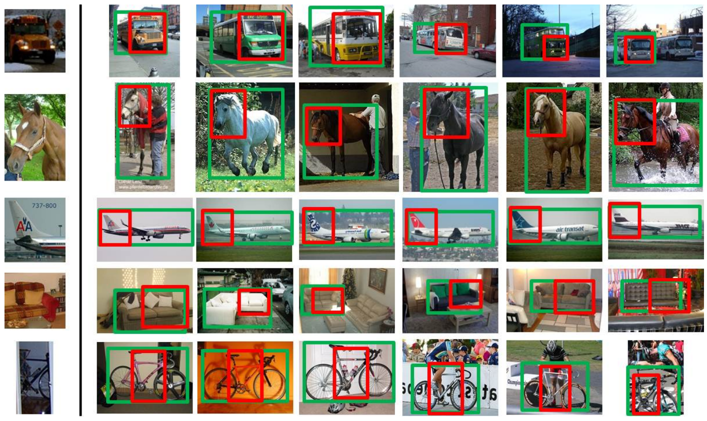

Figure 2: Subcategories composed of only a few instances, specifically in case of truncation, can gather more data from other training examples. Each row displays (left) a sample training instance from a subcategory, (right) new samples generated from existing training instances. Red box is the new sample, green box is the human annotation.

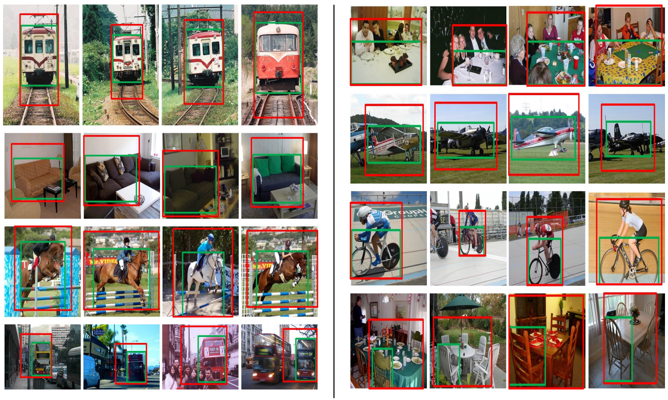

Figure 3: Human-annotated bounding boxes (green box) are automatically enlarged (red box) to leverage local contextual cues (adapted to the subcategory). There is a wide variation in the types of context captured per subcategory e.g., rail tracks under a 'train', people seated at 'dining table', etc.

Approach Overview Our approach builds upon the latent bounding box fitting method introduced in [1, 3], where the human-annotated bounding box is treated as being partially *latent* i.e., the bounding box is allowed to move within a local neighborhood (down to 70% overlap). Intuitively, this can be understood as locally "wiggling" the bounding box representation such that it best *aligns* with the rest of the object instances within a category (or subcategory). In this paper, we apply a very similar mechanism, but rather than just making local adjustments, we use it to *search* for bounding box representations that capture new correspondences between instances in the training data. The main difference is that the latent bounding box fitting assumes that each object instance is represented by a *single* bounding box belonging to a single subcategory, whereas our aim is to find *many different* bounding boxes for the same instance, so that it can be shared across multiple subcategories.

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, September 2010.

[2] Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. Adaptive mixture of local experts. In *Neural Computation*, 1991.

[3] P. A. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2005.