Watch and Learn: Semi-Supervised Learning of Object Detectors From Videos

Ishan Misra, Abhinav Shrivastava, Martial Hebert
Robotics Institute, Carnegie Mellon University

The availability of large labeled image datasets [1, 2] has been one of the key factors for advances in recognition. These datasets have not only helped boost performance, but have also fostered the development of new techniques. However, compared to images, videos seem like a more natural source of training data because of the additional temporal continuity they offer for both learning and labeling. The available video datasets lack the richness and variety of annotations offered by benchmark image datasets. It also seems unlikely that human per-image labeling will scale to the web-scale video data without using temporal constraints. In this paper, we show how to exploit the temporal information provided by videos to enable semi-supervised learning.

We present a scalable framework that discovers and localizes multiple objects in video using semi-supervised learning (see Figure 1). It tackles this challenging problem in long video (a million frames in our experiments) starting from only a few labeled examples. In addition, we present our algorithm in a realistic setting of *sparse labels* [3], i.e., in the few initial "labeled" frames, not all objects are annotated. This setting relaxes the assumption that in a given frame all object instances have been exhaustively annotated. It also implies that we do not know if any unannotated region in the frame is an instance of the object category or the background, and thus cannot use any region from our input as negative data. While much of the past work has ignored this type of sparse labeling and *lack of explicit negatives*, we show ways to overcome this handicap.

Contributions: Our semi-supervised learning (SSL) framework *localizes multiple unknown objects* in videos. Starting from *sparsely labeled* objects, it iteratively labels new training examples in the videos. Our key contributions are: 1) We tackle the SSL problem for discovering multiple objects in sparsely labeled videos; 2) We present an approach to constrain SSL [6] by combining multiple weak cues in videos and exploiting decorrelated errors by modeling data in multiple feature spaces. We demonstrate its effectiveness as compared to traditional tracking-by-detection approaches. 3) Given the redundancy in video data, we need a method that can automatically determine the relevance of training examples to the target detection task. We present a way to include *relevance and diversity of the training examples* in each iteration of the SSL, leading to scalable *incremental learning*.

Our algorithm starts with a few sparsely annotated video frames ($\mathcal{L}$) and iteratively discovers new instances in the large unlabeled set of videos ($\mathcal{U}$). Simply put, we first train detectors on annotated objects, followed by detection on input videos. We determine good detections (removing confident false positives) which serve as starting points for short-term tracking. The short-term tracking aims to label unseen examples reliably. Amongst these newly labeled examples, we identify diverse examples which are used to update the detector without re-training from scratch. We iteratively repeat this process to label new examples. We now describe our algorithm.

Sparse Annotations (lack of explicit negatives): We start with a few sparsely annotated frames in a random subset of $\mathcal{U}$. Sparse labeling implies that unlike standard tracking-by-detection approaches, we cannot sample negatives from the vicinity of labeled positives. We use random images from the internet as negative data for training object detectors on these sparse labels. We use these detectors to detect objects on a *subset of the video*, e.g., every 30 frames. Training on a few positives without domain negatives results in high confidence false positives. Removing such false positives is important because if we track them, we will add many more bad training examples, thus degrading the detector's performance over iterations.

Temporally consistent detections: We first remove detections that are temporally inconsistent using a smoothness prior on the motion of detections.

Decorrelated errors: To remove high confidence false positives, we rely on the principle of *decorrelated errors* (similar to *multi-view* SSL [5]). The intuition is that the detector makes mistakes that are related to its feature representation, and a different feature representation would lead to different errors. Thus, if the errors in different feature spaces are decorrelated, one can correct the errors and remove false positives. This gives us a filtered set of detections.

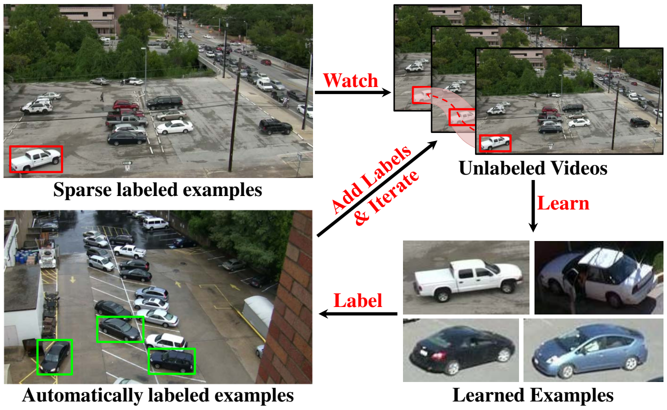

Figure 1: We present a novel formulation of semi-supervised learning for automatically learning object detectors from videos. Our method works with long video to automatically learn bounding box level annotations for multiple object instances. It does not assume exhaustive labeling of every object instance in the input videos, and from a handful of labeled instances can automatically label hundreds of thousands of instances.

Reliable tracking: We track these filtered detections to label new examples. Our final goal is not to track the object over a long period. Instead, our goal is to track reliably and label new examples for the object detector. To get such reliable tracks we design a conservative *short-term tracking* algorithm that identifies *tracking failures*.

Selection of diverse positives for updating the detector: The reliable tracklets give us a large set of automatically labeled boxes which we use to update our detector. Previous work [4] temporally subsamples boxes from videos, treating each box with equal importance. However, since these boxes come from videos, a large number of them are redundant and do not have equal importance for training our detector. Additionally, the relevance of an example added at iteration i depends on whether similar examples were added in the earlier iterations. One would ideally want to train (make an *incremental update*) only on new and diverse examples, rather than re-train from scratch on thousands of largely redundant boxes. We address this issue by selection and show one way of training only on diverse new boxes. After training detectors on diverse examples, we repeat the SSL process to iteratively label more examples.

We demonstrate the effectiveness of all our constraints in preventing semantic drift for SSL in videos. Our experiments also show that such an SSL approach can start with a handful of labeled examples and iteratively label hundreds of thousands of new examples which also improve object detectors.

[1] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007).

[3] S. Oh, A. Hoogs, and A. Perera et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.

[4] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[5] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. In *WACV*, 2005.

[6] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012.

This is an extended abstract. Full paper available at the Computer Vision Foundation webpage.