Weakly Supervised Object Detection with Convex Clustering

Hakan Bilen[1,2], Marco Pedersoli[1] Tinne Tuytelaars[1]
[1]ESAT-PSI / iMinds, KU Leuven. [2]Department of Engineering Science, Oxford University.

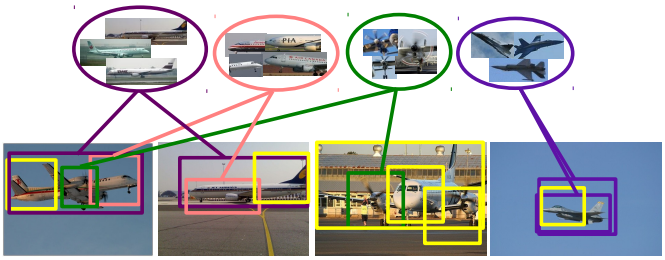

Figure 1: **An illustration of our learning model:** In the top row, we show clusters of objects and object parts that are simultaneously learned with the detectors during training. Our method encourages highly probable windows to be similar among them through the jointly learned clusters during training. The colored lines indicate similarity between windows and clusters. Best viewed in color.

The standard approach for supervised learning of object detection models requires the annotation of each target object instance with a bounding box in the training set. This fully supervised paradigm is tedious and costly for large-scale datasets. In this work, we focus on a more challenging "weakly supervised" case when the annotation at training time is restricted to presence or absence of object instances at image-level.

An ideal weakly supervised learning (WSL) for object detection is expected to guide the missing annotations to a solution that disentangles object instances from noisy and cluttered background. The standard paradigm alternates between labeling the missing annotations and learning a classifier based on these labellings in a spirit similar to Expectation Maximization (EM). Due to the missing annotations, this optimization is non-convex and therefore prone to getting stuck in a local minimum.

In this paper we investigate a possible way to improve the optimization by imposing similarity among objects of the same class. In particular, we propose to couple a smooth discriminative learning procedure as proposed in our earlier work [1] with a convex clustering algorithm [3]. While the discriminative learning estimates a model to best separate positive and negative data, the clustering searches for a small set of exemplars that best describes the discriminative training data. The exemplars are selected based on their similarity to discriminative parts of images and this enforces discriminative parts among training images to be similar. It can be seen as an alternative efficient way to enforce local similarity without the need of the expensive Conditional Random Fields (CRF) as in [2] (see Fig.1).

Our goal is to detect the locations of the objects of a target class (*e.g.* "bicycle", "person"), if there is any, in a previously unseen image. To do so, we learn an object detector for the target class by using a set of positive images (images where at least one object of the target class is present) and negative images (images where there is no object of the target class present). As the locations of the target objects in the positive images are not given, we formulate the task in a latent support vector machine (LSVM) formulation [5] where we aim to find the latent parameter (object window(s)) for each training sample that best discriminates positive images from negative ones. Assuming that there exist some pairs of instances from the target class that are similar to each other, we jointly learn the location of object instances for each positive training image and a detector that is able to localize that object.

Let $x \in \mathcal{X}$, $y \in \{-1,1\}$ and $h \in \mathcal{H}$ denote an image, its binary label and the object location (bounding box) respectively. To learn a detector w, we define our objective objective function $\mathcal{L}$ on a set of training samples

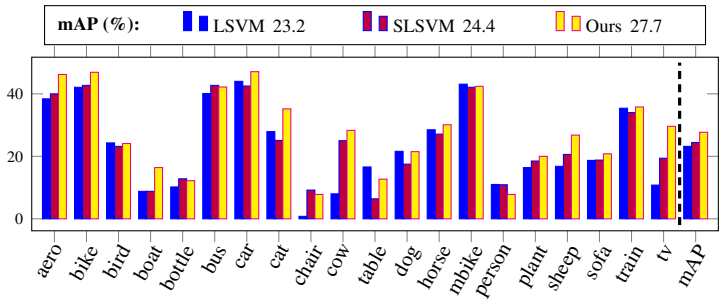

Figure 2: A comparison of our method to the baselines LSVM and SLSVM in terms of average precision on the VOC 2007 dataset. Our method with convex clustering significantly outperforms the baselines for most of the categories. Best viewed in color.

$\mathcal{S} = \{(x^i, y^i), i = 1, \ldots, N\}$ and minimize it with respect to w:

$$\mathcal{L}(w;\mathcal{S}) = \mathcal{L}_R(w) + \lambda \mathcal{L}_{sm}(w,\mathcal{S}) + \gamma \mathcal{L}_c(w,\mathcal{S}^+). \quad (1)$$

The first term is the standard l_2 regularization and the second is the soft-max latent svm loss as defined in [1]. The third one is our main contribution which measures the level of similarity among the positive samples of the same class $\mathcal{S}^+$. Thus the objective $\mathcal{L}$ is a weighted sum of regularization, discriminativity and similarity.

Similarity between pairs of samples in previous work is enforced through a CRF or a global similarity model. Instead we introduce a local similarity through a clustering procedure in an efficient and principled manner. Inspired by [3], we propose a clustering term that enforces such similarity:

$$\mathcal{L}_c = -\sum_{\substack{x\in\mathcal{S}^+ \\ h\in\mathcal{H}}} p(h|x,w) \log\Big(\sum_{\substack{x'\in\mathcal{S}^+ \\ h'\in\mathcal{H}}} q_{x',h'} e^{-\alpha\|\phi(x,h)-\phi(x',h')\|_2}\Big), \quad (2)$$

where $\phi(x,h)$ is a feature vector that represents the window h in image x. α is a positive temperature parameter and controls the sparseness of the scalar weights q. $p(h|x,w)$ estimates the probability of a window h of x to be the target class y based on the learned parameter vector w. $q_{x,h}$ measures how representative a window h of image x. In words, the clustering term penalizes configurations with high probability ($p(h|x,w)$) far from the important clusters (windows h with high $q_{x,h}$). $\mathcal{L}_c$ has two desirable properties: (i) it is convex given w so it is guaranteed to find the optimal solution, and (ii) it results in a sparse selection of clusters (windows h with $q_{x,h}$ greater than zero) and thus it automatically finds the optimal number of clusters for a given α.

We present the performance of our method and several baseline in Figure 2. First we compare the standard LSVM to its soft variation SLSVM and see that smoothing the max formulation leads to an improvement of 1.2 points. The convex clustering formulation (Ours) achieves a significant improvement of 3.3% in mAP (mean average precision) over SLSVM. The results suggest that similarity is a valuable channel of information and helps to better localize objects.

[1] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, 2014.

[2] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012.

[3] D. Lashkari and P. Golland. Convex clustering with exemplar-based models. In *NIPS*, pages 825–832, 2007.

[4] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *CVPR*, volume 1, pages 993–1000. IEEE, 2006.

[5] C.J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176, 2009.

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