# Weakly Supervised Object Detection with Convex Clustering

Zhang, Liqiang

June 8, 2018

## Abstract

*Weakly supervised object detection, is a challenging task, where the training procedure involves learning at the same time both, the model appearance and the object location in each image. The classical approach to solve this problem is to consider the location of the object of interest in each image as a latent variable and minimize the loss generated by such latent variable during learning. However, as learning appearance and localization are two interconnected tasks, the optimization is not convex and the procedure can easily get stuck in a poor local minimum, i.e. the algorithm "misses" the object in some images. In this paper, the author help the optimization to get close to the global minimum by enforcing a "soft" similarity between each possible location in the image and a reduced set of "exemplars", or clusters, learned with a convex formulation in the training images. The help is effective because it comes from a different and smooth source of information that is not directly connected with the main task.*

## 1. Introduction

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. The alternative but more challenging paradigm is to learn from the growing amount of noisily and sparsely annotated visual data available [2]. In this work, we focus on the specific "weakly supervised" case when the annotation at training time is restricted to presence or absence of object instances at image-level.

In this paper the author propose to couple a smooth discriminative learning procedure as proposed in their earlier work with a convex clustering algorithm. While the discriminative learning estimates a model to best separate positive and negative data, the clustering searches for a small set of exemplars. These exemplars that best describe our training data are not directly forced to be the localization hypotheses but they are selected based on the probability of being part of the object. This indirectly enforces the localized hypotheses to be similar to one another (similar to the cluster centers) and therefore it is a way to enforce local similarity without the need of the expensive pairwise CRF. Furthermore, the optimal number of clusters is automatically selected by the algorithm. This also allows the clustering procedure to optimally adapt to the new localization of object instances at any point of the learning and, due to the convexity of the optimization, it does not depend on the initialization. This idea is illustrated in Fig. 1.

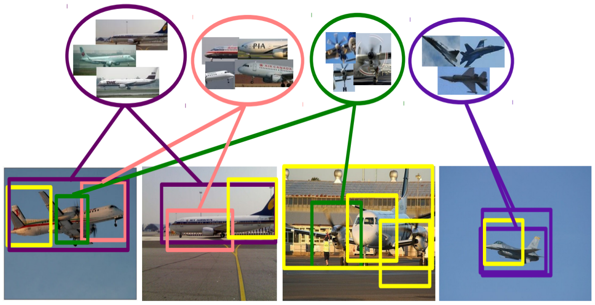

Figure 1. An illustration of our learning model: In the top row, the author show clusters of objects and object parts that are simultaneously learned with the detectors during training. Our method encourages highly probable windows to be similar among them through the jointly learned clusters during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

## 2. Inference and Learning

The author want their object models to score high for positive images (*i.e.*$y = 1$) and low for negative images (*i.e.*$y$ = -1). To train such object models that can separate between positive and negative samples, a common formulation to measure the mismatch between the image, label and window is the max-margin latent SVM (LSVM):

$$l_{mm}(w, x^i, y^i) = \max(w \cdot \Phi(x^i, y, h) + \Delta(y^i, y)) - \max w \cdot \Phi(x^i, y, h). \quad (1)$$

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Our method | 42.9 | 56.0 | 32.0 | 17.6 | 10.2 | 61.8 | 50.2 | 29.0 | 3.8 | 36.2 | 18.5 | 31.1 | 45.8 | 54.5 | 10.2 | 15.4 | 13.2 |
| Shi *et al.* [3] | 43.6 | 50.4 | 32.2 | 26.0 | 9.8 | 58.5 | 50.4 | 30.9 | 7.9 | 36.1 | 18.2 | 31.7 | 41.4 | 52.6 | 8.8 | 14.0 | 26.3 |
| Shi *et al.* [4] | 39.4 | 50.1 | 31.5 | 16.3 | 12.6 | 64.5 | 42.8 | 42.6 | 10.1 | 35.7 | 24.9 | 38.2 | 34.4 | 55.6 | 9.4 | 14.7 | 35.3 |
| Cinbis *et al.* [1] | 57.1 | 52.0 | 31.5 | 7.6 | 11.5 | 55.0 | 53.1 | 34.1 | 1.7 | 33.1 | 49.2 | 42.0 | 47.3 | 56.6 | 15.3 | 12.8 | 14.6 |
| Wang *et al.* [5] | 53.1 | 57.1 | 32.4 | 12.3 | 15.8 | 58.2 | 56.7 | 39.6 | 0.9 | 44.8 | 39.9 | 31.0 | 54.0 | 62.4 | 4.5 | 20.6 | 34.2 |

Table 1. Comparison of WSL object detectors on PASCAL VOC 2007 in terms of correct localization (CorLoc) on positive training images.

where $\Delta(y^i, y)$ is zero-one error, i.e. $\Delta(y^i, y) = 0$ if $y = y^i$, 1 else. This formulation aims to separate the highest scoring window h from the other configurations. The softmax term $l_{sm}$ is given as:

$$\begin{aligned} l_{mm}(w, x^i, y^i) = &\frac{1}{\beta} log \sum_{y,h} exp(\beta w \cdot \Phi(x^i, y, h) + \beta \Delta(y^i, y)) \\ &- \frac{1}{\beta} log \sum_{h} exp(\beta w \cdot \Phi(x^i, y, h)). \end{aligned} \tag{2}$$

) where $\beta$ is a tunable temperature parameter. It can be shown that Eq. 2 reduces to the max-margin formulation of [31], as $\beta \rightarrow \infty$. The author set this parameter to 1 in all their experiments. The margin loss for the training set is then $L_m(w, S) = \sum_{i=1}^{N} l_{sm}(w, x^i, y^i)$.

## 3. Experiments

The results in Table 1 show that the method is comparable to the state-ofthe-art in CorLoc.

## References

[1] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, pages 2409–2416, 2014. 2

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 1

[3] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, pages 2984–2991, 2014. 2

[4] Z. Shi, P. Siva, and T. Xiang. Transfer learning by ranking for weakly supervised object annotation. In *BMVC*, pages 1997–2005, 2012. 2

[5] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, pages 431–445, 2014. 2