Confidence-Rated Multiple Instance Boosting for Object Detection

Karim Ali[1,2] and Kate Saenko[2]

[1] University of California Berkeley, Berkeley, CA, USA
[2] University of Massachusetts Lowell, Lowell, MA, USA
karim.ali@berkeley.edu, saenko@cs.uml.edu

Abstract

Over the past years, Multiple Instance Learning (MIL) has proven to be an effective framework for learning with weakly labeled data. Applications of MIL to object detection, however, were limited to handling the uncertainties of manual annotations. In this paper, we propose a new MIL method for object detection that is capable of handling the noisier automatically obtained annotations. Our approach consists in first obtaining confidence estimates over the label space and, second, incorporating these estimates within a new Boosting procedure. We demonstrate the efficiency of our procedure on two detection tasks, namely, horse detection and pedestrian detection, where the training data is primarily annotated by a coarse area of interest detector. We show dramatic improvements over existing MIL methods. In both cases, we demonstrate that an efficient appearance model can be learned using our approach.

1. Introduction

Multiple Instance Learning (MIL) has emerged as a powerful paradigm for learning with labeling uncertainties. In contrast with Supervised Learning which requires unambiguously labeled training data, the MIL framework allows for a weaker form of supervision whereby training examples are no longer distinctly labeled singletons. Instead, training examples come in labeled bags. Typically, a negative bag is known to contain exclusively negative instances while a positive bag is known to contain *at least one* positive instance. In MIL, true labels for instances in positive bags are therefore latent and are estimated during model learning.

A number of MIL methods, implementing various forms of weak supervision, have been proposed in literature. However, when applied to object detection, these methods are only effective in the presence of mild labeling ambiguities. This is the case because learning an object detector is a task that is particularly sensitive to label noise and therefore the labeling errors that may result during MIL learning. Applications to object detection have thus far been limited to coping with slight translational and scale uncertainties around a manually annotated bounding box [17, 4, 8]. In these cases, many instances in a positive bag are relatively close to a true positive and MIL has been shown to reduce alignment error and improve detection performance.

We propose a new approach for MIL that is capable of handling strong labeling ambiguities. We rely on MILBoost, a simple and effective method, and propose to reduce the potential for label estimation errors by way of a two step procedure. First, a committee of randomized MILBoost learners is used to obtain confidence estimates on the labels of instances belonging to positive bags. Next, we incorporate the estimates within a new Boosting procedure, built by generalizing the MILBoost loss to incorporate a prior over the latent space and applying Friedman's gradient Boosting [11]. The resulting method is shown to be particularly effective in the case where most positive bags contain *little to no* positive instances, see Figure 7.

Our approach is validated on two tasks, namely horse detection and pedestrian detection. In both cases, we obtain the large majority of our weakly labeled data by performing a category-related text query such as "horses" or "pedestrian" on a commercial web image search engine. We process all returned images with a region of interest detector and form a positive bag per image with the returned bounding boxes. For any given image or bag, the returned set of bounding boxes consists either exclusively of background patches, or a small number of positives. We show improvements of more than 100% in average precision over MILBoost and demonstrate that an efficient appearance model can be learned in this manner.

2. Related Work

A variety of MIL algorithms have been developed since the idea was first introduced by Keeler et al. [14]. Central

This work has been supported in part by the Swiss National Science Foundation, by DARPA Mind's Eye and MSEE programs and by the US National Science Foundation Award 1212928.

to all methods is the formulation of a mechanism through which latent labels are estimated during training. We briefly review some of the key methods.

In the Normalized Set Kernel approach [12], a positive bag is represented as a normalized sum of the representations of all instances within it. The problem is thus reduced to a Supervised Learning problem and a Support Vector Machine (SVM) is learned over the new representation. In a similar, albeit Boosting based approach [18], the probability of a bag being positive is assumed to be the mean of the probabilities that its instances are positive. While effective at some tasks, the approaches in [12, 18] implicitly assume that the majority of instances in the bags are positive.

In [3], Andrews et al. propose two heuristic MIL algorithms. Both approaches alternate the training of an SVM with the relabeling of instances in positive bags. In mi-SVM, the prediction of the learned classifier is directly used to relabel the data, while in MI-SVM, the instance with maximum score in each bag is selected as the sole positive at each iteration. Both methods are initialized assuming that the majority of instances in positive bags are positive and therefore suffer from the same drawbacks as [12, 18].

The assumption that positive bags are mostly composed of positives can be violated in practice. In [5], Bunescu et al. attempted to address this issue by approximating the MIL constraint that a positive bag contain *at least one* positive more strictly. This is done via the introduction of appropriate constraints to the standard soft-margin SVM loss. The approach performs well when compared to [12, 18].

A number of other methods also forgo the assumption that positive bags are mostly composed of positives through an even stricter modeling of the MIL constraint. A relevant example is the Diversity Density approach of Maron et al. [15]. There, the probability of a bag being positive is formulated as a Noisy-OR [13] over the probabilities of its instances. In [17], Viola integrated the Noisy OR Criterion into the AnyBoost [16] framework to derive MILBoost. The proposed method was shown to reduce alignment error while training Viola-Jones people detectors. Following this line of work, Babenko et al. [4] apply the Gradient Boosting framework and proposed two new variants on MILBoost.

The performance of different MIL methods is ultimately tied to the sensitivity of the task to label noise and to the composition of the positive bags. For the noise-sensitive object detection task, even strict approaches such as MILBoost, which exactly assume *at least one* positive per bag, have only been shown to be successful in the case where positive bags are mostly composed of positives [17, 4]. Indeed, successful applications of MIL to object detection have been limited to dealing with slight translational and scale ambiguities around a manually annotated bounding box.

In this paper, we propose a new MIL method that is capable of handling strong labeling ambiguities for noise sensitive tasks. We consider the case where the weak labeling is such that positive bags are not even guaranteed to have a positive instance. We therefore enforce a stricter constraint whereby positive bags contain *either no positive or at least one* positive. Our method builds on MILBoost, chosen for its simplicity, its exact modeling of the standard MIL constraint and because it has been shown to compare favorably to other MIL methods on a variety of tasks [17, 4]. By incorporating confidence estimates over the labels of instances in positive bags, we render possible truly weakly supervised applications of MIL to Object Detection.

3. Methodology

We begin by defining the MIL problem in the context of detection. Next, we briefly review the main elements of gradient Boosting and MILBoost that are relevant to the understanding of our approach.

We focus the discussion on detection by classification. Given a scale-normalized image patch P and its associated feature vector $x_p \in \mathbb{R}^D$, we want to build a classifier

$$\varphi : \mathbb{R}^D \to \mathbb{R}$$

such that $\varphi(x_p) \geq 0$ when a target is present in the image patch. Detection on large scenes, at a particular threshold T, then consists of computing a list of alarms $\{s \in \mathcal{S} : \varphi(x_s) \geq T\}$ where $\mathcal{S}$ is typically the set of patches extracted at every possible position for every possible scale of the scene.

Supervised approaches build a fully annotated training set of N samples

$$(x_i, y_i) \in \mathbb{R}^D \times \{-1, 1\}$$

and train a classifier φ with a low empirical error rate on the data. By contrast, MIL approaches consider a weaker form of supervision with the training set consisting of

$$\left(X_i, y_i\right) \in \left(\mathbb{R}^D\right)^K \times \{-1, 1\}$$

where $X_i = \{x_{i1}, \cdots, x_{iK}\}$ is a set of K feature vectors called a *bag*. Instance labels, y_{ij}, are unavailable and

$$y_i = \max_j(y_{ij}). \tag{1}$$

Observe here that for a negative bag, all instances are necessarily negative. On the other hand, instance labels for positive bags must be estimated. The classifier φ is trained to achieve a low empirical error rate on the *bags* such that

$$\max_j \left(2 \cdot \mathbf{1}_{\{\varphi(x_{ij}) \geq 0\}} - 1\right) = y_i.$$

3.1. Gradient Boosting

We momentarily return to the supervised case to review Friedman's Gradient Boosting [11]. Let $h_m : \mathbb{R}^D \rightarrow \{-1, 1\},\ m = 1, \ldots, M$ be a family of "weak learners" and let $L(\varphi; \mathbf{y})$ define a differentiable loss over the data. We wish to construct a strong classifier as a linear combination of the form

$$\varphi(x) = \sum_{k=1}^{K} \alpha_k h_k(x).$$

Gradient Boosting consists of approximating the mapping

$$\varphi^* = \underset{\varphi \in \mathcal{H}}{\operatorname{argmin}}\, L(\varphi; \mathbf{y})$$

via a gradient descent procedure. At every step, the weak learner h_k which correlates the most with the gradient $\frac{\partial L(\varphi;\mathbf{y})}{\partial \varphi}$ is selected as the direction of descent and the weighing coefficient α_k is set to control the step polarity and size. Specifically, define a *weight* on each sample

$$\omega_i \equiv \frac{\partial L\left(\varphi(x_i); \mathbf{y}\right)}{\partial \varphi(x_i)}$$

then the selected weak learner at the k^{th} iteration is $h_k = \operatorname{argmax}_h \sum_{i=1}^{N} \omega_i h(x_i)$ and α_k is found using a line search to minimize $L(\varphi; \mathbf{y})$. By defining various loss functions, one can define a variety of Boosting procedures. In particular, by setting the loss to the logit,

$$L(\varphi; \mathbf{y}) = \sum_{i=1}^{N} \log(1 + e^{-y_i \varphi(x_i)}) \tag{2}$$

we obtain the LogitBoost algorithm which is minimized at

$$\varphi^*(x) = \log \frac{P\left(y = 1|x\right)}{P\left(y = -1|x\right)}, \tag{3}$$

and can be shown to maximize the likelihood of the data.

3.2. MILBoost

MILBoost can be viewed as a generalization of logistic Boosting where the standard MIL constraint is enforced exactly. Following Equation (3), the probability that an example is positive is given by

$$p_{ij} \equiv P(y_{ij} = 1|x_{ij}) = \frac{1}{1 + e^{-y_{ij}\varphi(x_{ij})}} \tag{4}$$

Next, the probability that a bag is positive is given by the Noisy OR criterion

$$p_i \equiv P(y_i = 1|X_i) = 1 - \prod_j (1 - p_{ij}) \tag{5}$$

enforcing the constraint that there exist at least one positive. By setting the loss to minimize the negative log-likelihood over the bags,

$$L(\varphi; \mathbf{y}) = -\log \prod_{i=1}^{N} p_i^{\bar{y_i}} (1 - p_i)^{1 - \bar{y_i}} \tag{6}$$

where $\bar{y_i} = \frac{y_i + 1}{2}$ and applying gradient Boosting we obtain the MILBoost procedure. Instance weights are given by

$$\omega_{ij} = \frac{\bar{y_i} - p_i}{pi} p_{ij} \tag{7}$$

At each Boosting iteration, high scoring instances within each bag are assigned higher weights and are in essence selected as positives. As seen in §5.7, when positives bags are formed by an area of interest detector and are hence subject to severe noise, performance can be poor: In these cases, the learning procedure readily converges to a multi-modal distribution wherein the selected examples consist of a mixture of true positives and background patches.

4. Confidence-Rated MILBoost

They key idea of our approach is to mitigate label estimation errors by way of a two-step procedure. In the first step, the consistency of instance responses across a committee of randomized MILBoost learners is used to obtain confidence estimates on the latent labels. In the second step, the obtained confidence estimates are incorporated into a generalized MILBoost procedure.

4.1. A Committee of Randomized Learners

In order to obtain confidence estimates over the latent labels, we train a number of predictors and form a committee

$$\mathcal{Q} = \{\varphi_1(x), \cdots, \varphi_Q(x)\}$$

where each $\varphi_q : \mathbb{R}^D \rightarrow \mathbb{R}$ is a randomized MILBoost Learner. Given these predictors, for each training instance, we define a confidence score

$$c(x) \equiv \log \frac{P\left(y = 1|\mathcal{Q}\right)}{P\left(y = -1|\mathcal{Q}\right)} = \sum_{q=1}^{Q} \varphi_q(x)$$

where the equality on the right is obtained from Equation (3) under the assumption that the predictors are statistically independent. For each instance, we can now define an instance confidence estimate as the probability that the instance shares its bag's label

$$\eta_{ij} \equiv P\left(y_{ij} = y_i|\mathcal{Q}\right) = \frac{1}{1 + e^{-y_i c(x)}}.$$

Finally, for each bag, we define a bag confidence estimate as the probability that the bag label is correct

$$\eta_i \equiv P(y_i|\mathcal{Q}) = \max_j \eta_{ij}.$$

where the equality on the right is obtained from Equation (1) assuming the bag instances are statistically independent. Finally, given that negative bags only contain negative instances, we set the confidence estimates of negative instances to $\eta_{ij} = 1$ and negative bags to $\eta_i = 1$.

4.2. Confidence-Rated MILBoost

We now integrate the obtained confidences into a generalized MIL Loss. As in MILBoost, the probability that an example is positive is given by

$$p_{ij} = \frac{1}{1 + e^{-y_{ij}\varphi(x_{ij})}}. \tag{8}$$

The probability that a bag is positive is given by the extended Noisy OR

$$p_i = 1 - \prod_j (1 - p_{ij})^{\frac{\eta_{ij}}{\eta_i^2}} \tag{9}$$

where the exponent is equivalent to repeating the x_{ij} instance η_{ij} times in the training set while ensuring that the repetitions are comparable across the bags. Next, we define an extended negative log-likelihood over the bags as,

$$L(\varphi; \mathbf{y}) = -\log \prod_{i=1}^{N} p_i^{\eta_i \bar{y}_i} (1 - p_i)^{\eta_i (1 - \bar{y}_i)}. \tag{10}$$

Applying gradient Boosting, we obtain the following instance weights

$$\omega_{ij} = \frac{\bar{y}_i - p_i}{\eta_i pi} \eta_{ij} p_{ij}$$

Recall that for negative bags $\eta_i = 1$ and $\eta_{ij} = 1$: the weight on a negative instance is therefore the same as would result from MILBoost, which is also the same as would result from LogitBoost. In particular, a negative example weight follows the logit: it increases linearly for negative margins and decreases exponentially for positive margins. A positive instance's weight can be interpreted as the product of bag weight $\frac{1-p_i}{\eta_i pi}$ and an instance weight $\eta_{ij} p_{ij}$. Within a bag, highly scoring instances with high confidence are assigned higher weight. The bag weight follows the reciprocal: as p_i approaches 1 the bag weight is reduced and all instances within the bag are collectively weighted down. Observe from Equation (9) that for bags with lower confidence η_i, pi will lie closer to 1: a bag with lower confidence will therefore have a lower bag weight. Finally, we note that if $\eta_{ij} = 1 \forall i, j$ implying that $\eta_i = 1 \forall i$, the procedure reduces to standard MILBoost.

4.3. Implementation Details

4.3.1 Numerical Stability

The quantity $1 - pi = \prod_j (1 - p_{ij})^{\frac{\eta_{ij}}{\eta_i^2}}$ in equation (9) approaches 0 as the bag size is increased or as the bag confidence η_i is decreased. Note that this is meaningful: large positive bags and low confidence positive bags are uninformative. In order to avoid underflow, the quantity $1 - pi$ must be stored separately from pi. Note that this is also necessary for the case of MILBoost.

4.3.2 Low-Confidence Positive Bags

Ideally, a positive bag whose confidence η_i is below 0.5 should be considered as a negative bag with increasing confidence as η_i approaches 0. This behavior is desirable from a learning perspective in order to make full use of the available data. We did not, however, experiment with such a setting. Instead we introduced a simple heuristic whereby all bags with confidence $\eta_i < 0.5$ are discarded from training.

4.3.3 Weak Learner Optimization

As mentioned in §3.1, at each iteration, Gradient Boosting selects the weak learner h_k which correlates the most with the gradient. It is easy to show [4] that in the binary $y_i \in \{-1, 1\}$ and discrete $h_k \in \{-1, 1\}$ case, one can equivalently minimize the weighted classification error:

$$\begin{aligned} h_k &= \operatorname*{argmax}_h \sum_{i=1}^{N} \omega_i h(x_i) \\ &= \operatorname*{argmin}_h \sum_{i=1}^{N} |\omega_i| \cdot \mathbf{1}_{\{h(x_i) = -y_i\}} \end{aligned} \tag{11}$$

This allows us to readily use existing learning routines designed to minimize a weighted classification error.

5. Experimental Results

5.1. Image Features and Weak Learners

We use the Gradient Histogram features of [2]. These features are obtained by computing q gradient orientation maps $O_\theta(u, v) = G(u, v) \cdot \mathbf{1}[\Theta(u, v) = \theta]$ where $G(u, v)$ and $\Theta(u, v)$ are the gradient magnitude and orientation at location (u, v). In other words each gradient orientation map $O_\theta(u, v)$ is formed with the magnitudes of gradients whose quantized orientation is θ. Let R denote an arbitrary subwindow in the image. Our features are entirely parameterized by the subwindow R and orientation θ as follows:

$$f_{R,\theta} = \frac{\sum_R O_\theta(u, v)}{\sum_{R,\theta} O_\theta(u, v)} \tag{12}$$

By varying R and θ, we obtain a very large feature representation x where the d^{th} coordinate $x^d = f_{R_d, \theta_d}$. This feature representation essentially computes Histogram of Oriented Gradients [6] in dense and arbitrary subwindows R (by contrast to a regular grid). It can be computed in constant time using q integral images, one for every gradient orientation map. In all our experiments we used $q = 8$.

5.2. Learning

From our feature representation x, we define weak learners as stumps (depth-1 decision trees) of the form:

$$h(x) = 2 \cdot \mathbf{1}_{\{x^d > \rho\}} - 1 \tag{13}$$

In all experiments, a single Boosting stage is trained with the bootstrapping procedure described in [10]. The selection of stumps at every Boosting iteration is done by examining 1000 weak learners whose thresholds ρ are optimized by exhaustive search. Feature parameters R and θ are chosen uniformly at random by enforcing a minimum size of 4 pixels for R. Learning was carried up to 500 stumps.

A committee of $|\mathcal{Q}| = 10$ MILBoost classifiers is learned for all experiments with the same procedure outline above with the exception that learning was only carried up to 100 stumps. Randomization is obtained primarily by sampling different negative samples for each learners. Observe also that restricting the search space for the optimal stump to 1000, using the bootstrapping procedure described in [10], as well as stoping the learning at 100 all contribute additional randomization.

5.3. Testing Data

We validate our approach on two view-based publicly available detection datasets, namely INRIA Horses [9] and INRIA Person [6].

INRIA Horses This dataset consists of 170 positive images containing 184 horses annotated with bounding boxes.. The horses are generally unoccluded, imaged from approximately the side viewpoint and face the same direction. The main challenges are clutter and intra-class variations.

INRIA Person This dataset consists of 288 positive images containing 589 pedestrians annotated with bounding boxes. The people are imaged approximately at eye-level, are upright with no particular bias in terms of frontal or side pose. There is significant clutter and intra-class variations.

5.4. Performance

In all experiments, we evaluate our trained detectors by multi-scale scanning of the full test images and compute error rates using the Pascal VOC [7] bounding box overlap criteria. We report performance with Precision-Recall curves and compute average precision (AP) as per the Pascal VOC criteria [7]. Detectors of size 68×68 and 64×128 are learned for the INRIA Horses and Person test set respectively. Images are scanned using scale strides of size $2^{\frac{1}{10}}$ and space strides of 4 pixels. All results were averaged with 5 independent runs and error bars plotted indicating minimum and maximum performance.

5.5. Training Data

Our weakly supervised training data is obtained by querying a web search engine with the keywords "horses" and "people walking" respectively. In both cases 200 images are collected. Example images sampled uniformly at random are shown in Figures 1 and 2. Note that only a fraction of the images (less than half) contain the target from the viewpoint corresponding to our test data, namely side view for horses and eye-level upright for pedestrians.

5.5.1 Bag Composition

Our overall weakly supervised data consists of 210 bags: 200 bags subject to severe ambiguity and acquired automatically and 10 bags subject to the typical scale and translational ambiguity as described below.

The 200 difficult bags are generated automatically from each internet image by running the interest area detector of [1], retaining the bounding boxes that fall within 25% of the desired detector's aspect ratio, randomly sampling 100 of those windows and finally mirror imaging the subimages. In this manner, we obtain a total of 200 bags, each with 200 samples, which we resize to the detector's aspect, for a total of $40,000$ subimages. Some example images are shown in Figure 7. Note that while some bags as in Figure 7(b) contain a few positives, other bags as in Figure 7(c) contain none. Overall, the level of noise across all $40,000$ patches obtained in this manner is significant. In order to prevent the MIL classifiers from learning repeating background structures, we augment the above 200 bags with 10 bags of the variety used in [17, 8, 4] obtained by perturbing a manually labeled bounding box on 10 of the images.

5.6. Baselines

In all our experiments we compare the performance of our algorithm against four baselines. The first three consist of detectors trained with the MIL methods MILBoost [17], ISRBoost [17] and GMBoost [4][1]. All three baselines are trained using the same data, namely 200 bags subject to severe ambiguity and the 10 approximately initialized bags subject to slight scale and translation ambiguity. The fourth consist of a LogitBoost classifier trained in a fully supervised manner with 110 manually annotated examples. This baseline is an indication of the best possible performance that a MIL method can achieve on the data assuming that a half of the 200 internet images contain a usable example and accounting for the additional 10 annotated samples. It establishes the *quality* of the samples selected by the MIL procedures.

[1] Both ISRBoost and GMBoost define smooth non-probabilistic approximation of the max as a loss and can be derived within a Gradient Boosting framework. We use an order 5 GMBoost loss.

5.7. Results

Qualitative results for the INRIA Person dataset are shown in Figure 3 and Figure 4 while results for the INRIA Horse dataset are shown in Figure 5 and Figure 6. Overall, results on both data sets are good, confirming the soundness of our approach.

In particular, for the INRIA Person dataset, CR-MILBoost achieves a relative gain of 68.2% in AP compared to MILBoost. The comparaison to the other MIL baselines is also substantially favorable with a 33% relative improvement over ISRBoost and a 29.2% improvement over GMBoost. The gains for the INRIA Horse dataset are even more substantial. CR-MILBoost achieves a relative gain over MILBoost of 111.3%, which translates a false alarm reduction rate of factor 10 when looking at the corresponding ROC (not shown here). Relative gains over ISRBoost and GMBoost are of 55% and 100.3% respectively. Note that the gains reported above are averaged over all recall values. Looking at recall 0.6, we note a relative improvement over MILBoost of 147.3% for the Person data and 504% for the Horse data.

Performance compared to the fully supervised LogitBoost baseline is often worse with approximate absolute losses of 10% AP for both datasets. The losses are greatest at high recall indicating that the training samples selected by CR-MILBoost still suffer from too much scale and translational ambiguity when compared with manually annotated ones. This is also evidenced by carefully examining Figure 8 and Figure 9: a careful examination does reveal slight scale and translational ambiguities. Such ambiguities are indeed known to be detrimental at high recall.

Interestingly, as can be seen Figure 4 for the INRIA Person data, CR-MILBoost *matches* the performance of the fully supervised baseline for up to 200 stumps at which point the performances of the two methods begin diverging. This indicates that the selected samples or at the very least the selected weak learners early in the learning process are better aligned with the test data than those selected subsequently.

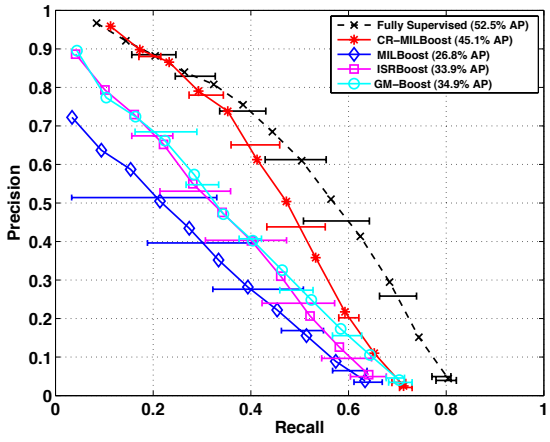

Figure 3: Performance of our method compared with baselines for the INRIA Person data. The figure displays precision as a function of recall.

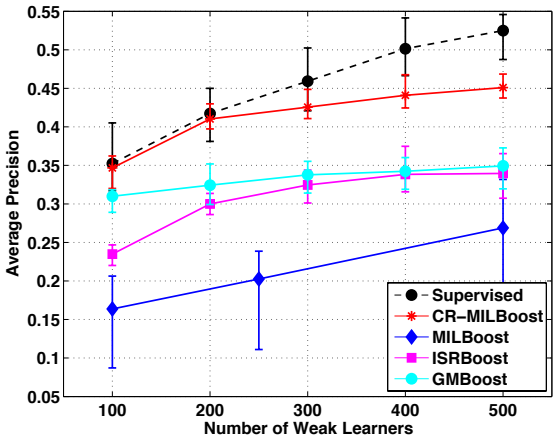

Figure 4: Performance of our method compared with baselines for the INRIA Person data. The figure displays average precision vs number of weak learners.

6. Concluding Remarks

We proposed a new MIL method for object detection that is capable of handling noisy automatically obtained annotations. Experiments demonstrate that our method strongly outperforms existing MIL methods and in some cases, can perform nearly as well as a supervised baseline.

This works admits two natural extensions. The first one is to exploit the known structure of the predictor to avoid the multiple randomized training runs. CR-MILBoost handles the MILBoost classifier as a black box with an unknown structure, and re-interpret the prediction itself, or its distribution across runs, to get an estimate of confidence. However, statistical by-products of the learning may also shed light on the confidence: The ambiguity of the optimization in general, and more precisely the uncertainty on a weak learner behavior, given its score in the Boosting process, may give an estimate of the uncertainty of the learning process, without the need for multiple runs.

The second is to use CR-MILBoost's confidence estimates to re-query the weakly supervised data and hence recompose the positive bags. The current experiments suggest that the selected samples by CR-MILBoost are inferior in quality to manually labeled one. We believe that such a strategy holds the key to improving the quality of the selected samples and thus performance.

Figure 1: Examples images from our horse training dataset selected uniformly at random. Typically, fewer than half of the 200 images contain a horse example from the same (side) view as in the INRIA Horse test data

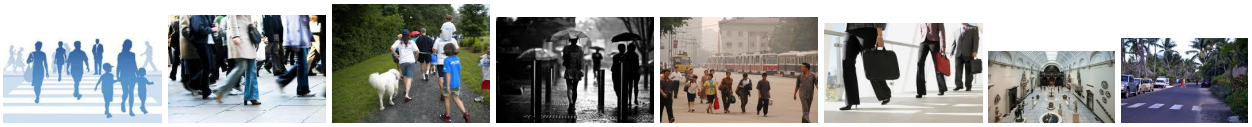

Figure 2: Examples images from our people training dataset selected uniformly at random. Typically, fewer than half of the 200 images contain a person example from the same (upright) view as in the INRIA Person test data

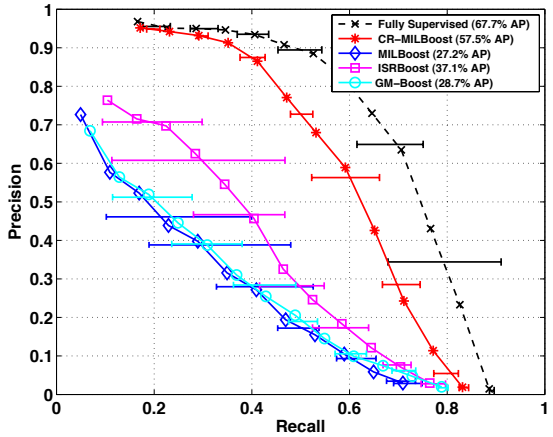

Figure 5: Performance of our method compared with baselines for the INRIA Horse data. The figure displays precision as a function of recall.

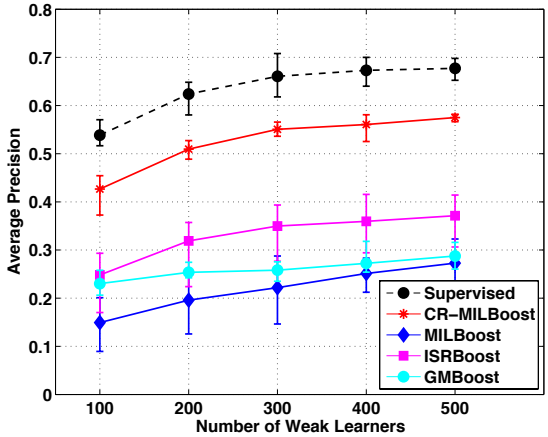

Figure 6: Performance of our method compared with baselines for the INRIA Horse data. The figure displays average precision vs number of weak learners.

Aknowledgment

The authors thank Trevor Darrell for his invaluable part in funding and supporting this work.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2189–2202, 2012.

[2] K. Ali, F. Fleuret, D. Hasler, and P. Fua. A Real-Time Deformable Detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):225–239, February 2012.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 561–568, 2003.

[4] B. Babenko, P. Dollár, Z. Tu, and S. Belongie. Simultaneous learning and alignment: Multi-instance and multi-pose learning. In *ECCV: Faces in Real-Life Images*, October 2008.

[5] R. C. Bunescu and R. J. Mooney. Multiple instance learning for sparse positive bags. In *International Conference on Machine Learning*, 2007.

[6] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

(a) (b) (c)

Figure 7: Two example bags subject to severe noise hand-picked from our 200 pedestrian internet images. (a) Typical examples from INRIA Person test set. (b) A bag containing a small number of positives. Top: original internet image. Bottom: 12 samples picked uniformly at random from the bag. (c) A bag containing no positives. Top: original internet image. Bottom: 12 samples picked uniformly at random from the bag.

Figure 8: Highest margin examples from our weakly-supervised horses training set in a single CR-MILBoost run.

Figure 9: Highest margin examples from our weakly-supervised people training set in a single CR-MILBoost run.

[9] V. Ferrari, F. Jurie, and C. Schmid. From Images to Shape Models for Object Detection. *International Journal of Computer Vision*, 2009.

[10] F. Fleuret and D. Geman. Stationary Features and Cat Detection. *Journal of Machine Learning Research*, 9:2549–2578, 2008.

[11] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[12] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *International Conference on Machine Learning*, pages 179–186, 2002.

[13] D. Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *UAI*, pages 163–172, 1989.

[14] J. D. Keeler, D. E. Rumelhart, and W. K. Leow. Integrated segmentation and recognition of hand-printed numerals. In *NIPS*, volume 3, pages 557–563, 1990.

[15] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 570–576, 1998.

[16] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent in function space, 1999.

[17] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. *Advances in Neural Information Processing Systems*, 18:1417–1424, 2006.

[18] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 272–281. Springer, 2004.