Spot On: Action Localization from Pointly-Supervised Proposals

Pascal Mettes*, Jan C. van Gemert‡, and Cees G. M. Snoek*

*University of Amsterdam
‡Delft University of Technology

Abstract. We strive for spatio-temporal localization of actions in videos. The state-of-the-art relies on action proposals at test time and selects the best one with a classifier trained on carefully annotated box annotations. Annotating action boxes in video is cumbersome, tedious, and error prone. Rather than annotating boxes, we propose to annotate actions in video with points on a sparse subset of frames only. We introduce an overlap measure between action proposals and points and incorporate them all into the objective of a non-convex Multiple Instance Learning optimization. Experimental evaluation on the UCF Sports and UCF 101 datasets shows that (i) spatio-temporal proposals can be used to train classifiers while retaining the localization performance, (ii) point annotations yield results comparable to box annotations while being significantly faster to annotate, (iii) with a minimum amount of supervision our approach is competitive to the state-of-the-art. Finally, we introduce spatio-temporal action annotations on the train and test videos of Hollywood2, resulting in *Hollywood2Tubes*, available at `http://tinyurl.com/hollywood2tubes`.

Keywords: Action localization, action proposals

1 Introduction

This paper is about spatio-temporal localization of actions like *Driving a car*, *Kissing*, and *Hugging* in videos. Starting from a sliding window legacy [1], the common approach these days is to generate tube-like proposals at test time, encode each of them with a feature embedding and select the most relevant one, *e.g.,* [2,3,4,5]. All these works, be it sliding windows or tube proposals, assume that a carefully annotated training set with boxes per frame is available a priori. In this paper, we challenge this assumption. We propose a simple algorithm that leverages proposals at *training* time, with a minimum amount of supervision, to speedup action location annotation.

We draw inspiration from related work on weakly-supervised object detection, *e.g.,* [6,7,8]. The goal is to detect an object and its bounding box at test time given only the object class label at train time and no additional supervision. The common tactic in the literature is to model this as a Multiple Instance Learning (MIL) problem [8,9,10] where positive images contain at least one positive object proposal and negative images contain only negative proposals. During each

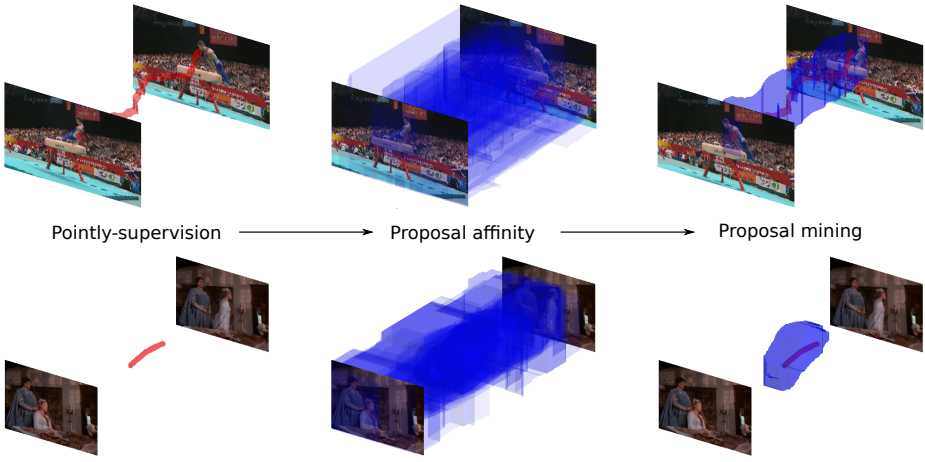


Fig. 1: **Overview of our approach** for a *Swinging* and *Standing up* action. First, the video is annotated cheaply using point-supervision. Then, action proposals are extracted and scored using our overlap measure. Finally, our proposal mining aims to discover the single one proposal that best represents the action, given the provided points.

iteration of MIL, a detector is trained and applied on the train set to re-identify the object proposal most likely to enclose the object of interest. Upon convergence, the final detector is applied on the test set. Methods typically vary in their choice of initial proposals and the multiple instance learning optimization. In the domain of action localization a similar MIL tactic easily extends to action proposals as well but results in poor accuracy as our experiments show. Similar to weakly-supervised object detection, we rely on (action) proposals and MIL, but we include a minimum amount of supervision to retain action localization accuracy competitive with full supervision.

Obvious candidates for the supervision are action class labels and bounding boxes, but other forms of supervision, such as tags and line strokes, are also feasible [11]. In [12], Bearman *et al.* show that human-provided points on the image are valuable annotations for semantic segmentation of objects. By inclusion of an objectness prior in their loss function they report a better efficiency/effectiveness trade off compared to image-level annotations and free-from squiggles. We follow their example in the video domain and leverage point-supervision to aid MIL in finding the best action proposals at training time.

We make three contributions in this work. First, we propose to train action localization classifiers using spatio-temporal proposals as positive examples rather than ground truth tubes. While common in object detection, such an approach is as of yet unconventional in action localization. In fact, we show that using proposals instead of ground truth annotations does not lead to a decrease in action localization accuracy. Second, we introduce an MIL algorithm that is

able to mine proposals with a good spatio-temporal fit to actions of interest by including point supervision. It extends the traditional MIL objective with an overlap measure that takes into account the affinity between proposals and points. Finally, with the aid of our proposal mining algorithm, we are able to supplement the complete Hollywood2 dataset by Marszałek *et al.* [13] with action location annotations, resulting in *Hollywood2Tubes*. We summarize our approach in Figure 1. Experiments on Hollywood2Tubes, as well as the more traditional UCF Sports and UCF 101 collections support our claims. Before detailing our pointly-supervised approach we present related work.

2 Related work

Action localization is a difficult problem and annotations are avidly used. Single image bounding box annotations allow training a part-based detector [1,14] or a per-frame detector where results are aggregated over time [15,16]. However, since such detectors first have to be trained themselves, they cannot be used when no bounding box annotations are available. Independent training data can be brought in to automatically detect individual persons for action localization [3,17,18]. A person detector, however, will fail to localize contextual actions such as *Driving* or interactions such as *Shaking hands* or *Kissing*. Recent work using unsupervised action proposals based on supervoxels [2,5,19] or on trajectory clustering [4,20,21], have shown good results for action localization. In this paper we rely on action proposals to aid annotation. Proposals give excellent recall without supervision and are thus well-suited for an unlabeled train set.

Large annotated datasets are slowly coming available in action localization. Open annotations benefit the community, paving the way for new data-driven action localization methods. UCF-Sports [22], HOHA [23] and MSR-II [24] have up to a few hundred actions, while UCF101 [25], Penn-Action [26], and J-HMBD [27] have 1–3 thousand action clips and 3 to 24 action classes. The problem of scaling up to larger sets is not due to sheer dataset size: there are millions of action videos with hundreds of action classes available [25,28,29,30]. The problem lies with the spatio-temporal annotation effort. In this paper we show how to ease this annotation effort, exemplified by releasing spatio-temporal annotations for all Hollywood2 [13] videos.

Several software tools are developed to lighten the annotation burden. The gain can come from a well-designed user interface to annotate videos with bounding boxes [31,32] or even polygons [33]. We move away from such complex annotations and only require a point. Such point annotations can readily be included in existing annotation tools which would further reduce effort. Other algorithms can reduce annotation effort by intelligently selecting which example to label [34]. Active learning [35] or trained detectors [36] can assist the human annotator. The disadvantage of such methods is the bias towards the used recognition method. We do not bias any algorithm to decide where and what to annotate: by only setting points we can quickly annotate all videos.

Weakly supervised methods predict more information than was annotated. Examples from static images include predicting a bounding box while having only class labels [8,37,38] or even no labels al all [39]. In the video domain, the temporal dimension offers more annotation variation. Semi-supervised learning for video object detection is done with a few bounding boxes [40,41], a few global frame labels [42], only video class labels [43], or no labels at all [44]. For action localization, only the video label is used by [45,46], whereas [47] use no labels. As our experiments show, using no label or just class labels performs well below fully supervised results. Thus, we propose a middle ground: pointing at the action. Compared to annotating full bounding boxes this greatly reduces annotation time while retaining accuracy.

3 Strong action localization using cheap annotations

We start from the hypothesis that an action localization proposal may substitute the ground truth on a training set without a significant loss of classification accuracy. Proposal algorithms yield hundreds to thousands of proposals per video with the hope that at least one proposal matches the action well [2,4,5,19,20,21]. The problem thus becomes how to mine the best proposal out of a large set of candidate proposals with minimal supervision effort.

3.1 Cheap annotations: action class labels and pointly-supervision

A minimum of supervision effort is an action class label for the whole video. For such global video labels, a traditional approach to mining the best proposal is Multiple Instance Learning [10] (MIL). In the context of action localization, each video is interpreted as a bag and the proposals in each video are interpreted as its instances. The goal of MIL is to train a classifier that can be used for proposal mining by using only the global label.

Next to the global action class label we leverage cheap annotations within each video: for a subset of frames we simply point at the action. We refer to such a set of point annotations as *pointly-supervision*. The supervision allows us to easily exclude those proposals that have no overlap with any annotated point. Nevertheless, there are still many proposals that intersect with at least one point. Thus, points do not uniquely identify a single proposal. In the following we will introduce an overlap measure to associate proposals with points. To perform the proposal mining, we will extend MIL's objective to include this measure.

3.2 Measuring overlap between points and proposals

To explain how we obtain our overlap measure, let us first introduce the following notation. For a video V of N frames, an action localization proposal $A = \{\text{BB}_i\}_{\text{i=}f}^{m}$ consists of connected bounding boxes through video frames $(f, ..., m)$ where $1 \leq f \leq m \leq N$. We use $\overline{BB_i}$ to indicate the center of a

bounding box i. The pointly-supervision $C = \{(x_i, y_i)\}^K$ is a set of $K \leq N$ sub-sampled video frames where each frame i has a single annotated point (x_i, y_i). Our overlap measure outputs a score for each proposal depending on how well the proposal matches the points.

Inspired by a mild center-bias in annotators [48], we introduce a term $M(\cdot)$ to represent how close the center of a bounding box proposal is to an annotated point, relative to the bounding box size. Since large proposals have a higher likelihood to contain any annotated point we use a regularization term $S(\cdot)$ on the proposal size. The center-bias term $M(\cdot)$ normalizes the distance to the bounding box center by the distance to the furthest bounding box side. A point $(x_i, y_i) \in C$ outside a bounding box $BB_i \in A$ scores 0 and a point on the bounding box center $\overline{BB_i}$ scores 1. The score decreases linearly with the distance to the center for the point. It is averaged over all annotated points K:

$$M(A, C) = \frac{1}{K} \sum_{i=1}^{K} \max(0, 1 - \frac{||(x_i, y_i) - \overline{BB_{K_i}}||_2}{\max_{(u,v) \in e(BB_{K_i})} ||((u, v) - \overline{BB_{K_i}})||_2}, \quad (1)$$

where $e(BB_{K_i})$ denotes the box edges of box BB_{K_i}.

We furthermore add a regularization on the size of the proposals. The idea behind the regularization is that small spatial proposals can occur anywhere. Large proposals, however, are obstructed by the edges of the video. This biases their middle-point around the center of the video, where the action often happens. The size regularization term $S(\cdot)$ addresses this bias by penalizing proposals with large bounding boxes $|BB_i| \in A$, compared to the size of a video frame $|F_i| \in V$,

$$S(A, V) = \big(\frac{\sum_{i=f}^{m} |BB_i|}{\sum_{j=1}^{N} |F_j|}\big)^2. \quad (2)$$

Using the center-bias term $M(\cdot)$ regularized by $S(\cdot)$, our overlap measure $O(\cdot)$ is defined as

$$O(A, C, V) = M(A, C) - S(A, V). \quad (3)$$

Recall that A are the proposals, C captures the pointly-supervision and V the video. We use $O(\cdot)$ in an iterative proposal mining algorithm over all annotated videos in search for the best proposals.

3.3 Mining proposals overlapping with points

For proposal mining, we start from a set of action videos $\{\mathbf{x}_i, t_i, y_i, C_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^{A_i \times D}$ is the D-dimensional feature representation of the A_i proposals in video i. Variable $t_i = \{\{BB_j\}_{j=f}^m\}^{A_i}$ denotes the collection of tubes for the A_i proposals. Cheap annotations consist of the class label y_i and the points C_i.

For proposal mining we insert our overlap measure $O(\cdot)$ in a Multiple Instance Learning scheme to train a classification model that can learn the difference between good and bad proposals. Guided by $O(\cdot)$, the classifier becomes increasingly more aware about which proposals are a good representative for an

action. We start from a standard MIL-SVM [8,10] and adapt it's objective with the the mining score $P(\cdot)$ of each proposal, which incorporates our function $O(\cdot)$ as:

$$\begin{aligned} & \min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_i \xi_i, \\ \text{s.t.} \quad & \forall_i : y_i \cdot (\mathbf{w} \cdot \arg\max_{\mathbf{z} \in x_i} P(\mathbf{z}|\mathbf{w}, b, t_i, C_i, V_i) + b) \geq 1 - \xi_i, \\ & \forall_i : \xi_i \geq 0, \end{aligned} \tag{4}$$

where $(\mathbf{w}, b)$ denote the classifier parameters, ξ_i denotes the slack variable and λ denotes the regularization parameter. The proposal with the highest mining score per video is used to train the classifier.

The objective of Equation 4 is non-convex due to the joint minimization over the classifier parameters $(\mathbf{w}, b)$ and the maximization over the mined proposals $P(\cdot)$. Therefore, we perform iterative block coordinate descent by alternating between clamping one and optimizing the other. For fixed classifier parameters $(\mathbf{w}, b)$, we mine the proposal with the highest Maximum a Posteriori estimate with the classifier as the likelihood and $O(\cdot)$ as the prior:

$$P(\mathbf{z}|\mathbf{w}, b, t_i, C_i, V_i) \propto (<\mathbf{w}, \mathbf{z}> + b) \cdot O(t_i, C_i, V_i). \tag{5}$$

After a proposal mining step, we fix $P(\cdot)$ and train the classifier parameters $(\mathbf{w}, b)$ with stochastic gradient descent on the mined proposals. We alternate the mining and classifier optimizations for a fixed amount of iterations. After the iterative optimization, we train a final SVM on the best mined proposals and use that classifier for action localization.

4 Experimental setup

4.1 Datasets

We perform our evaluation on two action localization datasets that have bounding box annotations both for training and test videos.

UCF Sports consists of 150 videos covering 10 action categories [49], such as *Diving*, *Kicking*, and *Skateboarding*. The videos are extracted from sport broadcasts and are trimmed to contain a single action. We employ the train and test data split as suggested in [14].

UCF 101 has 101 actions categories [25] where 24 categories have spatio-temporal action localization annotations. This subset has 3,204 videos, where each video contains a single action category, but might contain multiple instances of the same action. We use the first split of the train and test sets as suggested in [25] with 2,290 videos for training and 914 videos for testing.

4.2 Implementation details

Proposals. Our proposal mining is agnostic to the underlying proposal algorithm. We have performed experiments using proposals from both APT [4] and Tubelets [2]. We found APT to perform slightly better and report all results using APT.
Features. For each tube we extract Improved Dense Trajectories and compute HOG, HOF, Traj, MBH features [50]. The combined features are reduced to 128 dimensions through PCA and aggregated into a fixed-size representation using Fisher Vectors [51]. We construct a codebook of 128 clusters, resulting in a 54,656-dimensional representation per proposal.
Training. We train the proposal mining optimization for 10 iterations for all our evaluations, similar to Cinbis *et al.* [8]. Following further suggestions by [8], we randomly split the training videos into multiple (3) splits to train and select the instances. While training a classifier for one action, we randomly sample 100 proposals of each video from the other actions as negatives. We set the SVM regularization λ to 100.
Evaluation. During testing we apply the classifier to all proposals of a test video and maintain the top proposals per video. To evaluate the action localization performance, we compute the Intersection-over-Union (IoU) between proposal p and the box annotations of the corresponding test example b as: $\text{iou}(p,b) = \frac{1}{|\Gamma|}\sum_{f\in\Gamma} IoU_{p,b}(f)$, where Γ is the set of frames where at least one of p, b is present [2]. The function IoU states the box overlap for a specified frame. For IoU threshold t, a top selected proposal is deemed a positive detection if $\text{iou}(p,b) \geq t$.

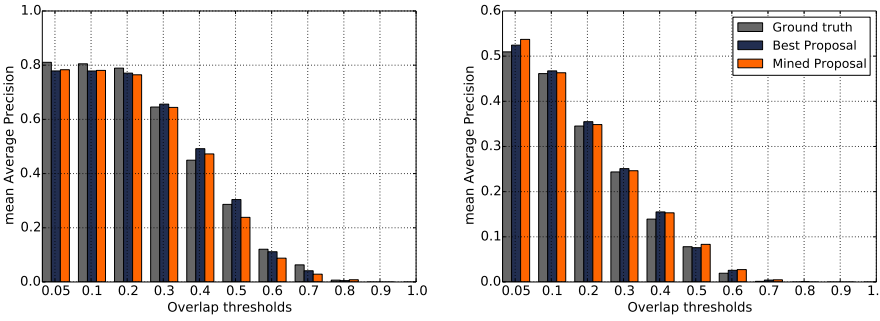
After combining the top proposals from all videos, we compute the Average Precision score using their ranked scores and positive/negative detections. For the comparison to the state-of-the-art on UCF Sports, we additionally report AUC (Area under ROC curve) on the scores and detections.

5 Results

5.1 Training without ground truth tubes

First we evaluate our starting hypothesis of replacing ground truth tubes with proposals for training action localization classifiers. We compare three approaches: 1) train on ground truth annotated bounding boxes; 2) train on the proposal with the highest IoU overlap for each video; 3) train on the proposal mined based on point annotations and our proposal mining. For the points on both datasets, we take the center of each annotated bounding box.

Training with the best proposal. Figure 2 shows that the localization results for the best proposal are similar to the ground truth tube for both datasets and across all IoU overlap thresholds as defined in Section 4.2. This result shows that proposals are sufficient to train classifiers for action localization. The result is somewhat surprising given that the best proposals used to train the classifiers have a less than perfect fit with the ground truth action. We computed the fit



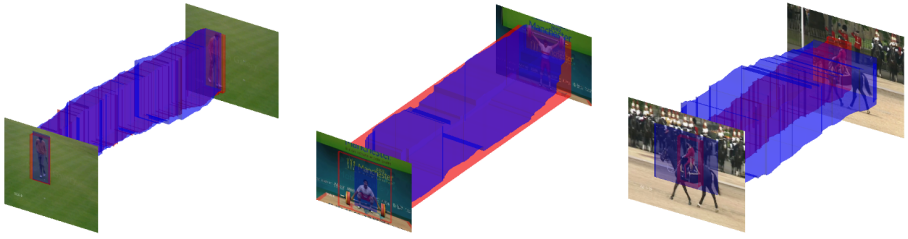
(a) UCF Sports. (b) UCF 101.

Fig. 2: **Training action localization classifiers with proposals** vs ground truth tubes on (a) UCF Sports and (b) UCF 101. Across both datasets and thresholds, the best possible proposal yields similar results to using the ground truth. Also note how well our mined proposal matches the ground truth and best possible proposal we could have selected.

with the ground truth, and on average the IoU score of the best proposals (the ABO score) is 0.642 on UCF Sports and 0.400 on UCF 101. The best proposals are quite loosely aligned with the ground truth. Yet, training on such non-perfect proposals is not detrimental for results. This means that a perfect fit with the action is not a necessity during training. An explanation for this result is that the action classifier is now trained on the same type of noisy samples that it will encounter at test-time. This better aligns the training with the testing, resulting in slightly improved accuracy.

Training with proposal mining from points. Figure 2 furthermore shows the localization results from training without bounding box annotations using only point annotations. On both data sets, results are competitive to the ground truth tubes across all thresholds. This result shows that when training on proposals, carefully annotated box annotations are not required. Our proposal mining is able to discover the best proposals from cheap point annotations. The discrepancy between the ground truth and our mined proposal for training is shown in Figure 3 for thee videos. For some videos, *e.g.,* Figure 3a, the ground truth and the proposal have a high similarity. This does however not hold for all videos, *e.g.,* Figures 3b, where our mined proposal focuses solely on the lifter (*Lifting*), and 3c, where our mined proposal includes the horse (*Horse riding*).

Analysis. On UCF 101, where actions are not temporally trimmed, we observe an average temporal overlap of 0.74. The spatial overlap in frames where proposals and ground truth match is 0.38. This result indicates that we are better capable of detecting actions in the temporal domain than the spatial domain. On average, top ranked proposals during testing are 2.67 times larger than their corresponding ground truth. Despite a preference for larger proposals, our results are comparable to the fully supervised method trained on expensive



(a) *Walking.* (b) *Lifting.* (c) *Riding horse.*

Fig. 3: **Training video showing our mined proposal** (blue) and the ground truth (red). (a) Mined proposals might have a high similarity to the ground truth. In (b) our mining focuses solely on the person lifting, while in (c) our mining has learned to include part of the horse. An imperfect fit with the ground truth does not imply a bad proposal.

ground truth bounding box tubes. Finally, we observe that most false positives are proposals from positive test videos with an overlap score below the specified threshold. On average, 26.7% of the top 10 proposals on UCF 101 are proposals below the overlap threshold of 0.2. Regarding false negatives, on UCF 101 at a 0.2 overlap threshold, 37.2% of the actions are not among the top selected proposals. This is primarily because the proposal algorithm does not provide a single proposal with enough overlap.

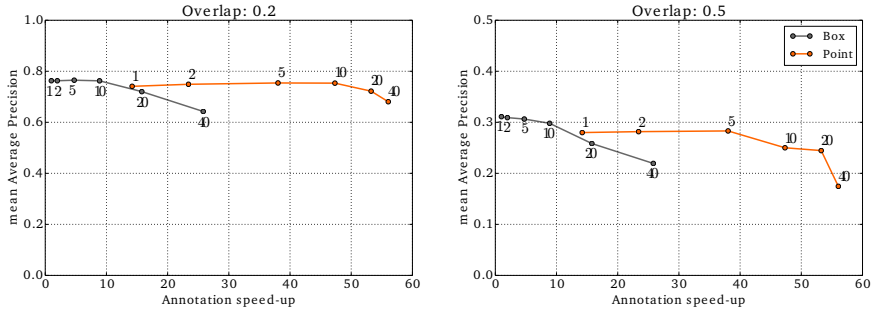
From this experiment we conclude that training directly on proposals does not lead to a reduction in action localization accuracy. Furthermore, using cheap point annotations with our proposal mining yields results competitive to using carefully annotated bounding box annotations.

5.2 Must go faster: lowering the annotation frame-rate

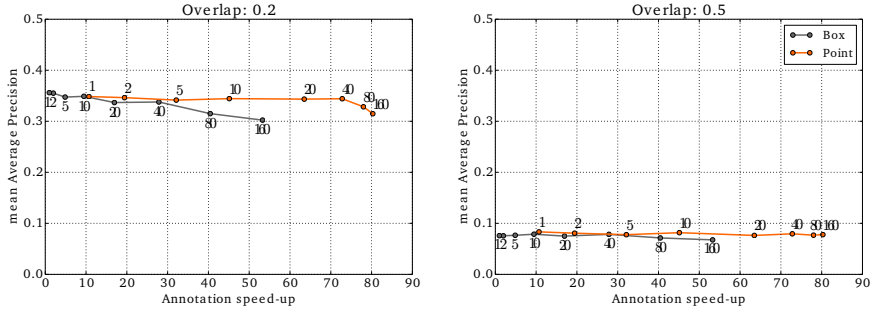
The annotation effort can be significantly reduced by annotating less frames. Here we investigate how a higher annotation frame-rate influences the trade-off between annotation speed-up versus classification performance. We compare higher annotation frame-rates for points and ground-truth bounding boxes.

Setup. For measuring annotation time we randomly selected 100 videos from the UCF Sports and UCF 101 datasets separately and performed the annotations. We manually annotated boxes and points for all evaluated frame-rates $\{1, 2, 5, 10, ...\}$. We obtain the points by simply reducing a bounding box annotation to its center. We report the speed-up in annotation time compared to drawing a bounding box on every frame. Classification results are given for two common IoU overlap thresholds on the test set, namely 0.2 and 0.5.

Results. In Figure 4 we show the localization performance as a function of the annotation speed-up for UCF Sports and UCF 101. Note that when annotating all frames, a point is roughly 10-15 times faster to annotate than a box. The reason for the reduction in relative speed-up between the higher frame-rates



(a) UCF Sports.



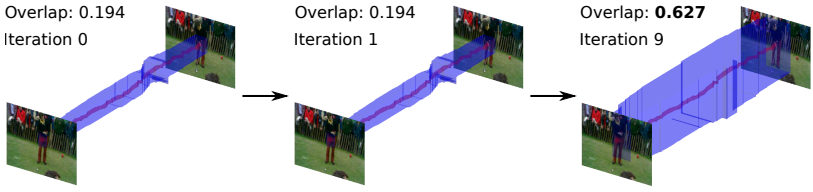
(b) UCF 101.

Fig. 4: **The annotation speedup** versus mean Average Precision scores on (a) UCF Sports and (b) UCF 101 for two overlap thresholds using both box and point annotations. The annotation frame-rates are indicated on the lines. Using points remains competitive to boxes with a 10x to 80x annotation speed-up.

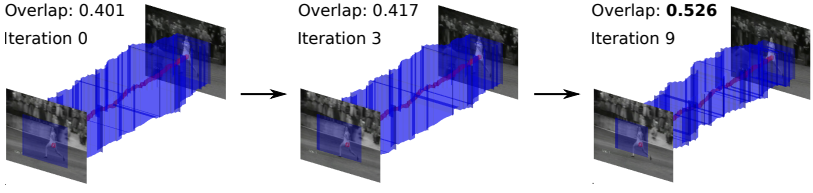
is due to the constant time spent on determining the action label of each video. When analyzing classification performance we note it is not required to annotate all frames. Although the performance generally decreases as less frames are annotated, using a frame rate of 10 (*i.e.,* annotating 10% of the frames) is generally sufficient for retaining localization performance. We can get competitive classification scores with an annotation speedup of 45 times or more.

The results of Figure 4 show the effectiveness of our proposal mining after the iterative optimization. In Figure 5, we provide three qualitative training examples, highlighting the mining during the iterations. We show two successful examples, where mining improves the quality of the top proposal, and a failure case, where the proposal mining reverts back to the initially mined proposal.

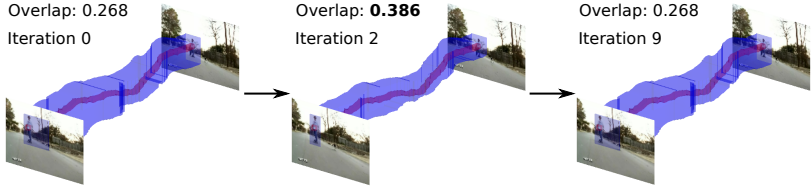
Based on this experiment, we conclude that points are faster to annotate, while they retain localization performance. We recommend that at least 10% of the frames are annotated with a point to mine the best proposals during training. Doing so results in a 45 times or more annotation time speed-up.



(a) *Swinging Golf.*



(b) *Running.*

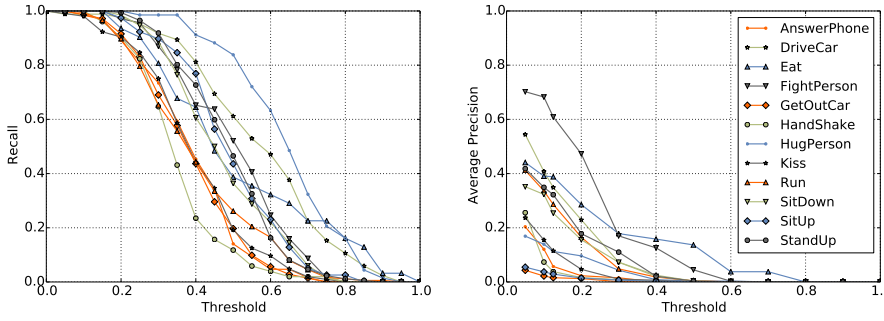


(c) *Skateboarding.*

Fig. 5: **Qualitative examples** of the iterative proposal mining (blue) during training, guided by points (red) on UCF Sports. (a) and (b): the final best proposals have a significantly improved overlap (from 0.194 to 0.627 and from 0.401 to 0.526 IoU). (c): the final best proposal is the same as the initial best proposal, although halfway through the iterations, a better proposal was mined.

5.3 Hollywood2Tubes: Action localization for Hollywood2

Based on the results from the first two experiments, we are able to supplement the complete Hollywood2 dataset by Marszałek *et al.* [13] with action location annotations, resulting in *Hollywood2Tubes*. The dataset consists of 12 actions, such as *Answer a Phone*, *Driving a Car*, and *Sitting up/down*. In total, there are 823 train videos and 884 test videos, where each video contains at least one action. Each video can furthermore have multiple instances of the same action. Following the results of Experiment 2 we have annotated a point on each action instance for every 10 frames per training video. In total, there are 1,026 action instances in the training set; 29,802 frames have been considered and 16,411 points have been annotated. For the test videos, we are still required to annotate bounding boxes to perform the evaluation. We annotate every 10 frames with a bounding box. On both UCF Sports and UCF 101, using 1 in 10 frames yields practically the same IoU score on the proposals. In total, 31,295 frames have been considered, resulting in 15,835 annotated boxes. The annotations, proposals, and localization results are available at `http://tinyurl.com/hollywood2tubes`.



(a) Recalls (MABO: 0.47). (b) Average Precisions.

Fig. 6: **Hollywood2Tubes**: Localization results for Hollywood2 actions across all overlap thresholds. The discrepancy between the recall and Average Precision indicates the complexity of the *Hollywood2Tubes* dataset for action localization.

Results. Following the experiments on UCF Sports and UCF 101, we apply proposals [4] on the videos of the Hollywood2 dataset. In Figure 6a, we report the action localization test recalls based on our annotation efforts. Overall, a MABO of 0.47 is achieved. The recall scores are lowest for actions with a small temporal span, such as *Shaking hands* and *Answer a Phone*. The recall scores are highest for actions such as *Hugging a person* and *Driving a Car*. This is primarily because these actions almost completely fill the frames in the videos and have a long temporal span.

In Figure 6b, we show the Average Precision scores using our proposal mining with point overlap scores. We observe that a high recall for an action does not necessarily yield a high Average Precision score. For example, the action *Sitting up* yields an above average recall curve, but yields the second lowest Average Precision curve. The reverse holds for the action *Fighting a Person*, which is a top performer in Average Precision. These results provide insight into the complexity of jointly recognizing and localizing the individual actions of *Hollywood2Tubes*. The results of Figure 6 shows that there is a lot of room for improvement.

In Figure 7, we highlight a difficult cases for action localization, which are not present in current localization datasets, adding to the complexity of the dataset. In the Supplementary Materials, we outline additional difficult cases, such as cinematographic effects and switching between cameras within the same scene.

5.4 Comparison to the state-of-the-art

In the fourth experiment, we compare our results using the point annotations to the current state-of-the-art on action localization using box annotations on the UCF Sports, UCF 101, and Hollywood2Tubes datasets. In Table 1, we provide a comparison to related work on all datasets. For the UCF 101 and Holly-

(a) Interactions. (b) Context. (c) Co-occurrence.

Fig. 7: **Hard scenarios for action localization** using Hollywood2Tubes, not present in current localization challenges. Highlighted are actions involving two or more people, actions partially defined by context, and co-occurring actions within the same video.

wood2Tubes datasets, we report results with the mean Average Precision. For UCF Sports, we report results with the Area Under the Curve (AUC) score, as the AUC score is the most used evaluation score on the dataset. All reported scores are for an overlap threshold of 0.2.

We furthermore compare our results to two baselines using other forms of cheap annotations. This first baseline is the method of Jain *et al.* [47] which performs zero-shot localization, *i.e.,* no annotation of the action itself is used, only annotations from other actions. The second baseline is the approach of Cinbis *et al.* [8] using global labels, applied to actions.

UCF Sports. For UCF Sports, we observe that our AUC score is competitive to the current state-of-the-art using full box supervision. Our AUC score of 0.545 is, similar to Experiments 1 and 2, nearly identical to the APT score (0.546) [4]. The score is furthermore close to the current state-of-the-art score of 0.559 [15,16]. The AUC scores for the two baselines without box supervision can not compete with our AUC scores. This result shows that points provide a rich enough source of annotations that are exploited by our proposal mining.

UCF 101. For UCF 101, we again observe similar performance to APT [4] and an improvement over the baseline annotation method. The method of Weinzaepfel *et al.* [16] performs better on this dataset. We attribute this to their strong proposals, which are not unsupervised and require additional annotations.

Hollywood2Tubes. For Hollywood2Tubes, we note that approaches using full box supervision can not be applied, due to the lack of box annotations on the training videos. We can still perform our approach and the baseline method of Cinbis *et al.* [8]. First, observe that the mean Average Precision scores on this dataset are lower than on UCF Sports and UCF 101, highlighting the complexity of the dataset. Second, we observe that the baseline approach using global video labels is outperformed by our approach using points, indicating that points provide a richer source of information for proposal mining than the baselines.

From this experiment, we conclude that our proposal mining using point annotations provides a profitable trade-off between annotation effort and performance for action localization.

Method	Supervision	UCF Sports AUC	UCF 101 mAP	Hollywood2Tubes mAP
Lan *et al.* [14]	box	0.380	-	-
Tian *et al.* [1]	box	0.420	-	-
Wang *et al.* [18]	box	0.470	-	-
Jain *et al.* [2]	box	0.489	-	-
Chen *et al.* [20]	box	0.528	-	-
van Gemert *et al.* [4]	box	0.546	0.345	-
Soomro *et al.* [5]	box	0.550	-	-
Gkioxari *et al.* [15]	box	0.559	-	-
Weinzaepfel *et al.* [16]	box	0.559	0.468	-
Jain *et al.* [47]	zero-shot	0.232	-	-
Cinbis *et al.* [8]*	video label	0.278	0.136	0.009
This work	points	0.545	0.348	0.143

Table 1: **State-of-the-art localization results** on the UCF Sports, UCF 101, and Hollywood2Tubes for an overlap threshold of 0.2. Where * indicates we run the approach of Cinbis *et al.* [8] intended for images on videos. Our approach using point annotations provides a profitable trade-off between annotation effort and performance for action localization.

6 Conclusions

We conclude that carefully annotated bounding boxes precisely around an action are not needed for action localization. Instead of training on examples defined by expensive bounding box annotations on every frame, we use proposals for training yielding similar results. To determine which proposals are most suitable for training we only require cheap point annotations on the action for a fraction of the frames. Experimental evaluation on the UCF Sports and UCF 101 datasets shows that: (i) the use of proposals over directly using the ground truth does not lead to a loss in localization performance, (ii) action localization using points is comparable to using full box supervision, while being significantly faster to annotate, (iii) our results are competitive to the current state-of-the-art. Based on our approach and experimental results we furthermore introduce *Hollywood2Tubes*, a new action localization dataset with point annotations for train videos. The point of this paper is that valuable annotation time is better spent on clicking in more videos than on drawing precise bounding boxes.

Acknowledgements

This research is supported by the STW STORY project.

References

1. Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: CVPR. (2013)

2. Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., Snoek, C.G.M.: Action localization with tubelets from motion. In: CVPR. (2014)
3. Yu, G., Yuan, J.: Fast action proposals for human action detection and search. In: CVPR. (2015)
4. van Gemert, J.C., Jain, M., Gati, E., Snoek, C.G.M.: Apt: Action localization proposals from dense trajectories. In: BMVC. (2015)
5. Soomro, K., Idrees, H., Shah, M.: Action localization in videos through context walk. In: ICCV. (2015)
6. Kim, G., Torralba, A.: Unsupervised detection of regions of interest using iterative link analysis. In: NIPS. (2009)
7. Russakovsky, O., Lin, Y., Yu, K., Fei-Fei, L.: Object-centric spatial pooling for image classification. In: ECCV. (2012)
8. Cinbis, R.G., Verbeek, J., Schmid, C.: Multi-fold mil training for weakly supervised object localization. In: CVPR. (2014)
9. Nguyen, M., Torresani, L., de la Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: ICCV. (2009)
10. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS. (2002)
11. Xu, J., Schwing, A.G., Urtasun, R.: Learning to segment under various forms of weak supervision. In: CVPR. (2015)
12. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. ECCV (2016)
13. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
14. Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: ICCV. (2011)
15. Gkioxari, G., Malik, J.: Finding action tubes. In: CVPR. (2015)
16. Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Learning to track for spatio-temporal action localization. In: ICCV. (2015)
17. Lu, J., Xu, R., Corso, J.J.: Human action segmentation with hierarchical supervoxel consistency. In: CVPR. (2015)
18. Wang, L., Qiao, Y., Tang, X.: Video action detection with relational dynamic-poselets. In: ECCV. (2014)
19. Oneata, D., Revaud, J., Verbeek, J., Schmid, C.: Spatio-temporal object detection proposals. In: ECCV. (2014)
20. Chen, W., Corso, J.J.: Action detection by implicit intentional motion clustering. In: ICCV. (2015)
21. Marian Puscas, M., Sangineto, E., Culibrk, D., Sebe, N.: Unsupervised tube extraction using transductive learning and dense trajectories. In: ICCV. (2015)
22. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. In: Computer Vision in Sports. (2014)
23. Raptis, M., Kokkinos, I., Soatto, S.: Discovering discriminative action parts from mid-level video representations. In: CVPR. (2012)
24. Cao, L., Liu, Z., Huang, T.S.: Cross-dataset action detection. In: CVPR. (2010)
25. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012)
26. Zhang, W., Zhu, M., Derpanis, K.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: ICCV. (2013)
27. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.: Towards understanding action recognition. In: ICCV. (2013)

28. Gorban, A., Idrees, H., Jiang, Y., Zamir, A.R., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes. In: CVPR workshop. (2015)
29. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. (2014)
30. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV. (2011)
31. Mihalcik, D., Doermann, D.: The design and implementation of viper. Technical report (2003)
32. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. IJCV **101**(1) (2013) 184–204
33. Yuen, J., Russell, B., Liu, C., Torralba, A.: Labelme video: Building a video database with human annotations. In: ICCV. (2009)
34. Settles, B.: Active learning literature survey. University of Wisconsin, Madison **52**(55-66) (2010)
35. Vondrick, C., Ramanan, D.: Video annotation and tracking with active learning. NIPS (2011)
36. Bianco, S., Ciocca, G., Napoletano, P., Schettini, R.: An interactive tool for manual, semi-automatic and automatic video annotation. CVIU **131** (2015) 88–99
37. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: CVPR. (2015)
38. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? - weakly-supervised learning with convolutional neural networks. In: CVPR. (2015)
39. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: CVPR. (2015)
40. Ali, K., Hasler, D., Fleuret, F.: Flowboost - appearance learning from sparsely annotated video. In: CVPR. (2011)
41. Misra, I., Shrivastava, A., Hebert, M.: Watch and learn: Semi-supervised learning for object detectors from video. In: CVPR. (2015)
42. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. In: ECCV. (2014)
43. Siva, P., Russell, C., Xiang, T.: In defence of negative mining for annotating weakly labelled data. In: ECCV. (2012)
44. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. In: ICCV. (2015)
45. Mosabbeb, E.A., Cabral, R., De la Torre, F., Fathy, M.: Multi-label discriminative weakly-supervised human activity recognition and localization. In: ACCV. (2014)
46. Siva, P., Xiang, T.: Weakly supervised action detection. In: BMVC. (2011)
47. Jain, M., van Gemert, J.C., Mensink, T., Snoek, C.G.M.: Objects2action: Classifying and localizing actions without any video example. In: ICCV. (2015)
48. Tseng, P.H., Carmi, R., Cameron, I.G., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. JoV **9**(7) (2009)
49. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008)
50. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. (2013)
51. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. IJCV **105**(3) (2013) 222–245

Supplementary materials

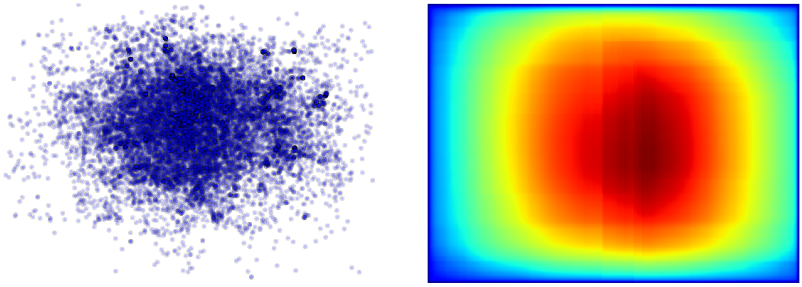
The supplementary materials for the ECCV paper "Spot On: Action Localization from Pointly-Supervised Proposals" contain the following elements regarding *Hollywood2Tubes*:

- The annotation protocol for the dataset.
- Annotation statistics for the train and test sets.
- Visualization of box annotations for each action.

Annotation protocol

Below, we outline how each action is specifically annotated using a bounding box. The protocol is the same for the point annotations, but only the center of the box is annotated, rather than the complete box.

- **AnswerPhone:** A box is drawn around both the head of the person answering the phone and the hand holding the phone (including the phone itself), from the moment the phone is picked up.
- **DriveCar:** A box is drawn around the person in the driver seat, including the upper part of the stear itself. In case of a video clip with of a driving car in the distance, rather than a close-up of the people in the car, the whole car is annotated as the driver can hardly be distinguished.
- **Eat:** A single box is drawn around the union of the people who are joinly eating.
- **FightPerson:** A box is drawn around both people fighting for the duration of the fight. If only a single person is visible, no annotation is made. In case of a chaotic brawl with more than two people, a single box is drawn around the union of the fight.
- **GotOutCar:** A box is drawn around the person starting from the moment that the first body parts exists the car until the person is standing complete outside the car, beyond the car door.
- **HandShake:** A box is drawn around the complete arms (the area between the union of the shoulders, ellbows, and hands) of the people shaking hands.
- **HugPerson:** A box is drawn around the heads and upper torso (until the waist, if visible) of both hugging people.
- **Kiss:** A box is drawn around the heads of both kissing people.
- **Run:** A box is drawn around the running person.
- **SitDown:** A box is drawn around the complete person from the moment the person starts moving down until the person is complete seated at rest.
- **SitUp:** A box is drawn around the complete person from the moment the person starts to move upwards from a laid down position until the person no longer moves upwards..
- **StandUp:** Vice versa to SitDown.



(a) Points (train). (b) Boxes (test).

Fig. 8: Annotation aggregations for the point and box annotations on *Hollywood2Tubes*. The annotations are overall center-oriented, but we do note a bias towards the rule-of-thirds principle, given the higher number of annotations on $\frac{2}{3}$-th the width of the frame.

	Training set	Test set
Number of videos	823	884
Number of action instances	1,026	1,086
Numbers of frames evaluated	29,802	31,295
Number of annotations	16,411	15,835

Table 2: Annotation statistics for *Hollywood2Tubes*. The large difference between the number of frames evaluated and the number of annotations is because the actions in Hollwood2 are not trimmed.

Annotation statistics

In Figure 8, we show the aggregated point annotations (training set) and box annotations (test set). The aggregation shows that the localization is center oriented. The heatmap for the box annotations do show the rule-of-thirds principle, given the the higher number of annotations on $\frac{2}{3}$-th the width of the frame.

In Table 2, we show a number of statistics on the annotations performed on the dataset.

Annotation examples

In Figure 9 we show an example frame of each of the 12 actions, showing the diversity and complexity of the videos for action localization.

(a) Answer Phone. (b) Drive Car. (c) Eat.

(d) Fight Person. (e) Get out of Car. (f) Hand Shake.

(g) Hug. (h) Kiss. (i) Run.

(j) Sit down. (k) Sit up. (l) Stand up.

Fig. 9: Example box annotations of test videos for *Hollywood2Tubes*.