Localizing Actions from Video Labels and Pseudo-Annotations

Mettes, P.; Snoek, C.G.M.; Chang, S.-F.

DOI
10.5244/C.31.22

Publication date
2017

Document Version
Final published version

Published in
Proceedings of the British Machine Vision Conference 2017

Localizing Actions from Video Labels and Pseudo-Annotations

Pascal Mettes[1]

Cees G.M. Snoek[1]

Shih-Fu Chang[2]

[1] University of Amsterdam
Amsterdam, NL

[2] Columbia University,
New York, USA

Abstract

The goal of this paper is to determine the spatio-temporal location of actions in video. Where training from hard to obtain box annotations is the norm, we propose an intuitive and effective algorithm that localizes actions from their class label only. We are inspired by recent work showing that unsupervised action proposals selected with human point-supervision perform as well as using expensive box annotations. Rather than asking users to provide point supervision, we propose fully automatic visual cues that replace manual point annotations. We call the cues pseudo-annotations, introduce five of them, and propose a correlation metric for automatically selecting and combining them. Thorough evaluation on challenging action localization datasets shows that we reach results comparable to results with full box supervision. We also show that pseudo-annotations can be leveraged during testing to improve weakly- and strongly-supervised localizers.

1 Introduction

The goal of this paper is to determine the spatio-temporal location of actions such as *Skateboarding* and *Shaking hands* in video content. This challenging problem is typically solved by classifying sliding cuboids [12, 13, 25] action proposals [3, 7, 17, 23, 24], or by linking detectors over time [10, 20, 30, 34]. In all cases, precise box annotations for actions on training video are a prerequisite for localizing actions in test videos. We challenge the need for spatio-temporal box annotations and propose an intuitive and effective algorithm that localizes actions in video from a video label only.

We are inspired by the recent work of Mettes *et al*. [16]. For their training they start from unsupervised action proposals [27], typically about 1,000 sequences of bounding boxes that are generated automatically for a video. Mettes *et al*. [16] show that using the best possible action proposal during training, rather than ground truth annotations, does not lead to a decrease in action localization accuracy. Encouraged by this observation, they introduce a variant of the Multiple Instance Learning algorithm [1] able to mine proposals with a good spatio-temporal fit to actions of interest by letting humans annotate a limited amount of points on the action in relevant training frames. While surprisingly effective, their approach still demands human supervision beyond the action class label. In this paper, we also rely

Figure 1: **We introduce pseudo-annotations from visual cues,** indicated by different colored dots, that simulate supervision in videos. From the pseudo-annotations and action class labels, we automatically select action proposals (blue tube) for training action localizers.

on unsupervised action proposals during training, but rather than selecting the best proposal using manual human point-supervision we prefer a completely automatic alternative.

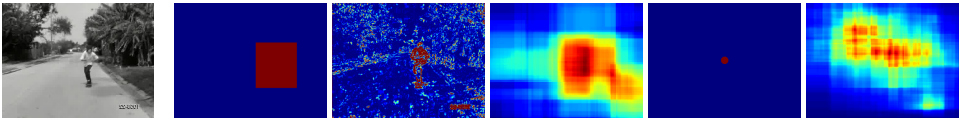
We introduce the notion of pseudo-annotations, see Figure 1, which we define as visual cues that replace point-supervision in video. We investigate five of such pseudo-annotations by exploiting sources such as action proposals [27], object proposals [35], person detections [34], motion [7], and center biases [26] to discover which cues are most informative to point on the action locations. The pseudo-annotations specify the likely location of an action in a video, resulting in the automatic selection of a desirable action proposal during Multiple Instance Learning optimization, where the information from pseudo-annotations is combined with action-specific video labels. To automatically select and combine pseudo-annotations from different cues, we introduce a metric based on correlations between the pseudo-annotations.

Thorough evaluation on multiple action localization datasets shows that individually, each visual cue is informative for localizing actions. Using our correlation metric for selecting and combining annotations, we reach results comparable to action localization from full box supervision with the same proposal and classification settings, while outperforming other weakly-supervised alternatives. Furthermore, we demonstrate how pseudo-annotations can be leveraged during testing, to further improve any localization result, be it trained on pseudo-annotations or manually annotated boxes.

2 Related work

Yu and Yuan [34] introduce supervised actor proposals for action localization. They rely on a person detector on successive frames and generate spatio-temporal proposals by assuring sufficient overlap and appearance consistency. Gkioxari and Malik [8] replace the person detector by an action-specific detector using appearance and motion. They link regions with strong overlap over time. Weinzaepfel *et al*. [30] follow the same scheme, but rather than linking detections they prefer tracking by detection (using boxes and class label) for further fine-tuning over time. It is obvious that by adding more supervision to the action proposal generation, better localization can be achieved, especially with deep learning, see Saha *et al*. [20]. Rather than using class-specific action detectors and box supervision, we prefer to localize an action in video from its class label only.

Jain *et al*. [7] introduce unsupervised action proposals that are likely to include the action, ideally achieving high recall with few proposals. They start from super-voxels and group them based on color, texture, motion, size, fill cues, and independent motion. Van Gemert *et al*. [27] bypass the computationally expensive segmentation step of [7] by creating unsupervised proposals directly from dense trajectories [28] used to represent videos



(a) (b) (c) (d) (e) (f)

Figure 2: **Heatmaps of each of the five pseudo-annotations** for (a) an example frame from *Skateboarding*. From left to right: (b) person detection, (c) independent motion, (d) action proposals, (e) frame center, and (f) object proposals.

during classification. Chen and Corso [3] also advocate clusters of dense trajectories for unsupervised action proposals. We also rely on unsupervised action proposals, but rather than selecting the best proposals using a classifier that learns from box annotations, we learn from a class label only.

Mettes *et al*. [16] propose to train action localization classifiers using unsupervised proposals as positive examples rather than ground truth boxes. They introduce a Multiple Instance Learning (MIL) algorithm that mines proposals with a good spatio-temporal fit to actions by including point supervision. It extends the traditional MIL objective with a measure that takes into account the overlap between proposals and points. Their approach allows to localize actions in video from class labels and point annotations. We also exploit a MIL optimization, but rather than relying on point-supervision, we prefer automated cues that do not require any action localization supervision.

3 Action localization with Pseudo-annotations

For training an action localizer, we are given a set of N training videos $\{X_i, Y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^{|A_i| \times D}$ states the $|A_i|$ action proposals, each of feature dimension D, and $Y_i \in \{-1, +1\}$ indicates the video label, which is $+1$ if the action occurs anywhere in the video and -1 otherwise. Each action proposal $A_i = \{A_i(t)\}_{t=1}^T$ is a tube consisting of T bounding boxes.

Our goal is to train a classifier using a proposal with high spatio-temporal overlap with the action of interest for each video. We employ a Multiple Instance Learning perspective [1, 4, 16]. Each video is a bag and the proposals in each video are the instances. Using a max-margin objective, the Multiple Instance Learning optimization is given as:

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

where $S(\mathbf{z}|\mathbf{w}, b, P)$ specifies a selection function for proposal $\mathbf{z} \in X_i$, conditioned on both the classifier score $(\mathbf{w}, b)$ and (pseudo-)annotations P.

In this work, we only require video labels. Therefore, we are tasked with automatically discovering annotations P, dubbed pseudo-annotations. They exploit sources such as action and object proposals, motion, humans, and center biases, see Figure 2. First, we outline how to obtain each individual pseudo-annotation, after which we show how to compute overlap scores from pseudo-annotations and how to combine the pseudo-annotations.

3.1 Pseudo-annotations

Pseudo-annotations from person detection. Actions are typically human-centered, so a robust detection of people in video frames provides information about the spatio-temporal location of actions. Here, we employ the Faster R-CNN network [18], using the person class after pre-training on MS-COCO [14]. After non-maximum suppression, the network yields roughly 50 box detections per frame, each with a confidence score. We select the bounding box in each frame with maximum confidence score as our pseudo-annotation.

Pseudo-annotations from independent motion. The independent motion at each pixel of a frame F provides information as to where the foreground action occurs in the frame. Here, we employ the interpretation of independent motion from Jain *et al.* [7]. Independent motion states the deviation from the global motion present in a frame. Let $\xi_{(x,y,F)} \in [0,1]$ denote the inverse of the residual in the global motion estimation at pixel (x,y) in frame F. The higher the value of $\xi_{(x,y,F)}$, the less likely that the pixel contributes to the global motion. Then we compute a point-wise pseudo-annotation for frame F as the center of mass over all the pixels in the frame, where the mass is given by their independent motion estimation:

$$p_{im}(F) = \frac{1}{\xi_{(F)}} \sum_{(x,y)\in F} \xi_{(x,y,F)} \cdot (x,y), \tag{2}$$

where $\xi_{(F)}$ denotes the total independent motion in frame F.

Pseudo-annotations from action proposals. We furthermore examine the action proposals themselves as a source of information for pseudo-annotations, using the unsupervised spatio-temporal proposals of [27]. For a frame F and action proposals A^*, we examine the spatial distribution of the proposal boxes of A^* in the frame. We make the following assumption about the spatial distribution of the proposals: the more the action proposals are on the same spatial location, the higher the likelihood that the action occurs in that location. The use of action proposals for pseudo-annotations can be interpreted in two ways. First, it is a form of self-supervision [5], as we employ the action proposals to specify which action proposals to train on. Second, it is a form of outlier detection. If many proposals agree on the same location, we give a penalty to the proposals that are outside that location.

For each pixel $(x,y) \in F$, we denote the number of proposals from A^* that contain (x,y) as $C_{A^*}(x,y,F)$. We compute the pseudo-annotation as the center of mass over these counts:

$$p_{pa}(F) = \frac{1}{C_{A^*}(F)} \sum_{(x,y)\in F} C_{A^*}(x,y,F) \cdot (x,y), \tag{3}$$

where $C_{A^*}(F)$ denotes the sum of the proposal counts over all pixels in F.

Pseudo-annotations from frame centers. In [16, 26], it is noted that both actions and annotators have a bias towards the center of the video. We exploit this bias directly by adding a point-wise pseudo-annotation on the center of each frame of each video:

$$p_{fc}(F) = (F_W/2, F_H/2), \tag{4}$$

where F_W and F_H denote the width and height of frame F.

Pseudo-annotations from object proposals. The presence of objects is also correlated with the presence of actions, as observed in [8, 9]. Object proposals are computed here from EdgeBoxes [35], using the top 1,000 object proposals per frame. Similar to the action proposal pseudo-annotation, we compute the number of proposals containing the pixel for

each pixel in frame F. Let $C_O(x,y,F)$ denote the number of proposals containing pixel (x,y), then the pseudo-annotation is given as:

$$p_{oa}(F) = \frac{1}{C_O(F)} \sum_{(x,y)\in F} C_O(x,y,F) \cdot (x,y). \tag{5}$$

where $C_O(F)$ denotes the sum of the proposal count over all pixels in F. The difference with Equation 3 is in that we assume here that the foreground action is the most dominant object in the scene, as defined by the number of object proposals focusing on the action.

3.2 Computing pseudo-annotation overlaps

Each visual cue outputs an automatic box (person detection) or point (all others) annotation. Given an action proposal A, we compute the overlap with the box annotations using the spatial-temporal intersection-over-union score [7]. The intersection-over-union with a set of box annotations B is computed as: $\frac{1}{|\Gamma|}\sum_{f\in\Gamma}\text{iou}(f_B,f_A)$, where Γ denotes the set of frames with at least one of A and B present. For point-wise pseudo-annotation P, we compute the overlap using the function defined in [16]: $O(A,P,V) = M(A,P) - S(A,V)$. Here $M(A,P)$ states the overlap between action proposal A and pseudo-annotations P and is defined as: $M(A,P) = \frac{1}{|P|}\sum_{i=1}^{|P|}\max(0, 1 - \frac{||(P_{x_i},P_{y_i}) - \overline{A_{P_i}}||_2}{\max_{(u,v)\in e(A_{P_i})}||((u,v)-\overline{A_{P_i}})||_2})$, where $\overline{A_{P_i}}$ denotes the center of the box of proposal A in frame P_i. In turn, $S(A,V)$ is a size regularization on the action proposal itself: $S(A,V) = \left(\frac{\sum_{i=f}^{m}|A_i|}{\sum_{j=1}^{N}|V_j|}\right)^2$, where proposal A runs from frame f to frame m, $|\cdot|$ denotes the area of a box, and V denotes the whole video. Intuitively, the overlap measure for point annotations aims to promote proposals with box centers close to the points while penalizing proposals of large size compared to the whole video.

3.3 Correlation metric for pseudo-annotations

For a video v, let $\{\mathbf{S}_v^{(i)}\}_{i=1}^{|P|}$ denote the overlap scores over all pseudo-annotations P and let $\mathbf{S}_v^{(i)} \in \mathbb{R}^{|A_v|}$ denote the overlap scores for the $|A_v|$ action proposals of the i^{th} pseudo-annotation in the video. Since no supervision within the videos is provided, it is *a priori* unknown how pseudo-annotations from different cues should be used and to what extent. Given the integral importance of people in detecting and localizing actions [34], we propose a correlation metric for pseudo-annotations using the person detection as an anchor for the correlation.

Let $\mathbf{H}_v$ denote the overlap scores of the action proposals in video v given by the person detection. Then we compute the statistical correlation between the pseudo-annotation of the i^{th} cue and the pseudo-annotations from the person detection over all N_t training videos:

$$\eta(P^{(i)}) = \frac{1}{N_t}\sum_{v=1}^{N_t}\frac{\text{cov}(\mathbf{S}_v^{(i)},\mathbf{H}_v)}{\sigma(\mathbf{S}_v^{(i)})\cdot\sigma(\mathbf{H}_v)}. \tag{6}$$

The covariance and standard deviations in Eq. 6 are computed over the pseudo-annotation overlap scores of all action proposals in video v. Intuitively, Eq. 6 assigns a high score to pseudo-annotations that assign similar overlaps scores to the person detection; the more a

pseudo-annotation agrees with the ranking of action proposals, the higher the correlation score. In turn, we can fuse the overlap scores of the pseudo-annotations as:

$$\mathbf{S}_v^{\text{fused}} = \sum_{P^{(i)} \in P} \eta(P^{(i)}) \cdot [\![\eta(P^{(i)}) \geq t]\!] \cdot \mathbf{S}_v^{(i)}, \tag{7}$$

where t is a threshold to remove pseudo-annotations with overlap scores too dissimilar to the person detection. Note that the person detection itself is also in the set P. In accordance with Eq. 6, the person detection yields a correlation score of 1.

The correlation metric for pseudo-annotations provides a way of measuring the quality of pseudo-annotations without the need for manual box or point annotations, nor the need for examining test performance to combine and select pseudo-annotations. By using a single pseudo-annotation per frame for each cue, we assume a single dominant action in each video. This assumption holds throughout our experiments. To handle videos with multiple actions and objects we can extend our approach with a density estimation over the pixel-wise weight of each cue to estimate multiple pseudo-annotations in frames.

4 Experimental setup

4.1 Datasets

UCF Sports. The UCF Sports dataset consists of 150 videos from sport broadcasts covering 10 action categories [19], including *Diving*, *Riding a Horse*, and *Skateboarding*. We employ the train and test data split as suggested in [13].

UCF 101. The UCF 101 dataset has 101 actions categories [22] 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 [22].

Hollywood2Tubes. The Hollywood2Tubes dataset consists of 1,707 videos with ground truth point (training videos) and box (test videos) annotations [16]. The dataset contains the actions from the Hollywood2 dataset [15], including *Getting out of a car*, *Hugging*, and *Fighting*. We use the the train and test data split as suggested in [15].

A2D. The A2D dataset contains 3,782 videos of actions performed both by human actors and other actors, such as dogs, cars, and babies [32]. For a limited number of video frames, box annotations are provided. We use the train and test split as suggested in [32].

We stress that throughout our experiments, we do not employ any manual point or box annotations for our approach.

4.2 Implementation details

Proposals. Following [16], we employ the unsupervised action proposals from [27]. We note that [27] only rely on dense trajectories for creating the proposals and do not use the cues that we employ for the pseudo-annotations.

Proposal representations. On all datasets, we represent each action proposal with a Fisher Vector [21] with 128 clusters over the improved dense trajectories [28] within the proposal. This results in a 54,656-dimensional representation per proposal.

Training. We train the Multiple Instance Learning algorithm for 5 iterations for all evaluations. Following [4], we split the training videos into multiple folds during training for the

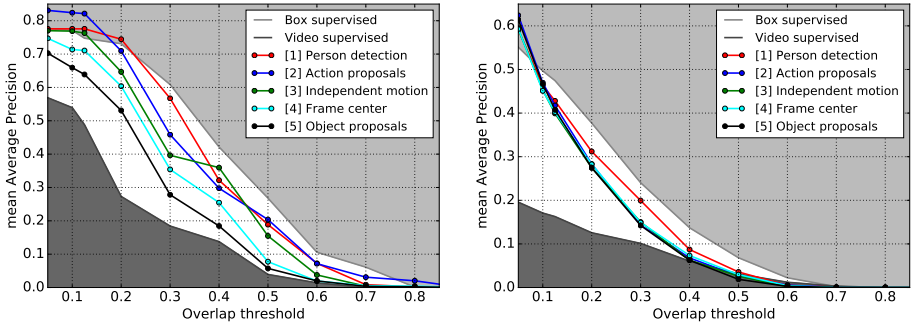


Figure 3: **Pseudo-annotation action localization performance** on UCF Sports (left) and UCF-101 (right), compared to the supervision upper and lower bounds.

classifier and proposal selection steps. We set the regularization parameter λ in the max-margin optimization to 10 in all experiments.

Evaluation. During testing we apply the classifier of an action to all proposals of a test video and keep the proposal with the highest classifier score [7, 27]. To evaluate the action localization performance, we compute the intersection-over-union in space and time between the top proposal and a ground truth tube as defined in [7]. Only proposals whose overlaps with ground truths exceed the threshold are considered correct.

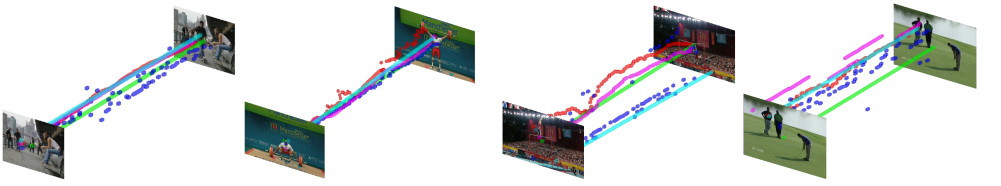
5 Experimental results

5.1 Evaluating the pseudo-annotations

In the first experiment, we evaluate each pseudo-annotation individually for action localization on UCF Sports and UCF-101 with the mean Average Precision score. We compare the pseudo-annotations to two baselines. The first baseline uses full box supervision during training (light gray area). This baseline serves as a supervision upper bound. The second baseline uses the video labels with standard Multiple Instance Learning (dark gray area). This baseline serves as the supervision lower bound. Note that all approaches use the same features and classifier settings.

The scores across all overlap thresholds are shown in Figure 3. On UCF Sports, we observe that each pseudo-annotation performs better than only using the video label, which means that pseudo-annotations provide meaningful information about the location of actions in videos. Furthermore, the person detection performs best, followed by action proposals and independent motion. At low overlap thresholds, these approaches even outperform full supervision. This is surprising, since no human intervention is provided. At higher thresholds, full supervision is still better, while all approaches break down at the highest thresholds.

On UCF-101, the pseudo-annotations similarly all outperform the approach using the video label only. The difference between the approaches is smaller since the dataset is larger, making it more robust against accidental hits and misses of the pseudo-annotations. The person detection pseudo-annotation performs best, followed by using frame centers and action proposals. To highlight the effect and limitations of the pseudo-annotations, we show qualitative results in Figure 4.



(a) Correct pseudo-annotations. (b) Incorrect pseudo-annotations.

Figure 4: **Qualitative examples of pseudo-annotations** in non-trivial action videos. In (a), the pseudo-annotations successfully follow the centrally oriented main action. In (b), some pseudo-annotations are distracted either by complex background (*Swinging on a bar*, left), or due to lack of primary action and the presence of other people (*Golf swinging*, right).

5.2 Combining pseudo-annotations

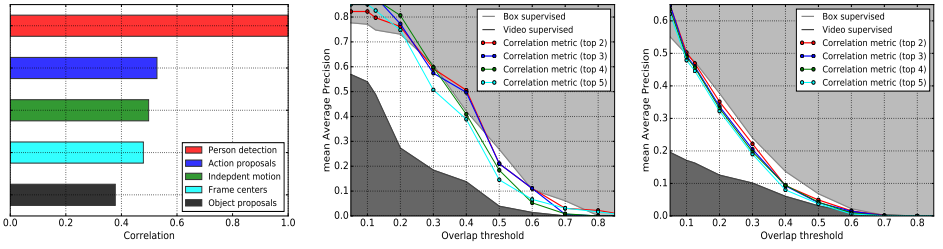
In the second experiment, we evaluate the correlation metric for pseudo-annotations. In Figure 5(a), we show the correlation scores on UCF Sports. The scores show that person detection, action proposals, and independent motion pseudo-annotations are most relevant, while frame centers and object proposals are less relevant. The discovered order is in line with the order of performance from the first experiment. This means that the correlation metric provides a reliable way of measuring the quality of pseudo-annotations, while automatic selection is possible by using the ones with highest average correlation.

We provide the localization performance in Figures 5(b) and 5(c). For UCF Sports, when using the correlation metric with both the top-two and top-three pseudo-annotations yields results comparable but not identical to full box supervision. When using more pseudo-annotations, the performance at higher overlap thresholds degrades, indicating that not all pseudo-annotations should be used in the combination. On UCF-101 (Figure 5(c)), using correlation metric with the top pseudo-annotations also yields results comparable or close to full box supervision. Here, the combination using person detection and frame centers (the second highest correlation pseudo-annotation) performs best. Incorporating more pseudo-annotations slightly degrades the performance.

We conclude from this experiment that a correlation metric from the top pseudo-annotations provides a reliable way to merge different visual cues for action localization. On both datasets, the metric with the top 2/3 pseudo-annotations outperform person detections only, while performing comparable to full box supervision.

Non-human action localization. The datasets typically used in action localization are human-centric [16, 19, 22]. Here, we evaluate how well our pseudo-annotations generalize to actions performed by non-human actors using the A2D dataset [31, 32]. These actors include babies, balls, birds, cars, cats, and dogs. Since this dataset does not have action tube annotations, our approach can not be directly evaluated. Individual box annotations for a set of frames per video are provided instead. Therefore, we investigate whether pseudo-annotations are capable of "pointing at" actions performed both by human and non-human actors. We evaluate how the overlap between proposals and pseudo-annotations relates to the overlap between proposals and ground truth boxes.

Over all actors, we find that the Pearson correlation score is 0.29; a high score for the pseudo-annotations correlates with a high score in action overlap, which strengthens the notion of pseudo-annotations for action localization. We also find that the Pearson correlation score is positive for all actor types individually. We do note that the score is higher for the



(a) Correlations (UCF Sports). (b) UCF Sports. (c) UCF-101.

Figure 5: **Correlation-based combination of pseudo-annotations.** On UCF Sports and UCF-101, automatically combining the top two or three correlated pseudo-annotations yields the best results, even comparable to full box supervision.

	UCF Sports						UCF 101					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
Full box annotations	77.1	73.1	60.9	42.2	26.8	10.6	46.1	34.5	24.4	13.9	7.8	1.9
Pseudo-annotations	88.0	77.2	57.4	49.7	21.0	11.1	50.3	35.1	22.2	9.3	4.9	1.6
Pseudo-annotations ++	87.7	81.7	64.4	54.5	37.8	17.5	49.8	37.4	25.8	13.7	6.2	1.3
Full box annotations ++	86.7	86.7	74.0	61.2	42.3	23.1	50.6	40.8	28.8	17.5	8.3	2.4

Table 1: **Localization performance (%) with pseudo-annotations during testing** (++). Using pseudo-annotations during testing increases the performance across all overlap thresholds and datasets, even outperforming full box supervision. Pseudo-annotations can also be employed to improve models trained with full box supervision.

person actor than the other types, indicating a bias towards persons as actors in our pseudo-annotations. Interestingly, when excluding the person detection as pseudo-annotation, person remains the most postively correlated actor type. We conclude that our pseudo-annotations are not restricted to the person as actor type and handle other actor types as well. The person as actor type does have closest relations to the pseudo-annotations, although this is not solely due to the use of the person detection itself.

5.3 Pseudo-annotations at test time

Since pseudo-annotations are automatically generated for videos, their use is not restricted to training videos only. In the third experiment, we employ pseudo-annotations during testing to help select the best proposal per video. We do this by combining the classifier score with the overlap scores from the pseudo-annotations. We employ the correlation metric with the top pseudo-annotations for this experiment.

Results on UCF Sports and UCF-101 are shown in Table 1. On both datasets, we observe a jump in performance when adding pseudo-annotations during testing, even outperforming the full box supervision results. This performance shows the effectiveness of the pseudo-annotations for action localization. We also evaluate the effect of using pseudo-annotations during testing with a model trained on full box supervision, which yields a similar increase in performance. We conclude from this experiment that pseudo-annotations during testing improves any model trained on unsupervised proposals. With only the video labels as manual annotations, we even outperform the standard full box supervision setup.

Method	Proposal annotations	Classifier annotations	UCF Sports AUC	UCF 101 mAP	H2T mAP
Wang *et al.* [29]	n.a.	video-label + joints	47.0	-	n.a.
Saha *et al.* [20]	video-label + boxes	video-label + boxes	-	66.8	n.a.
Weinzaepfel *et al.* [30]	video-label + boxes	video-label + boxes	55.9	46.8	n.a.
Jain *et al.* [9]	none	video-label + boxes	52.0	-	n.a.
van Gemert *et al.* [27]	none	video-label + boxes	54.6	34.5	n.a.
Mettes *et al.* [16]	none	video-label + points	54.5	34.8	14.3
Cinbis *et al.* [4],[16]	none	video-label	27.8	13.6	0.9
Jain *et al.* [8]	none	zero-shot	23.2	-	-
This paper	none	video-label	53.3	35.1	13.6
This paper ++	none	video-label	55.6	37.4	17.2

Table 2: **Localization results (%)** at an overlap of 0.2. A dash (-) states that results are not provided, while n.a. states that the approach can not be applied due to the dataset's lack of required annotations. The sign (++) denotes the use of pseudo-annotations during testing. We achieve results comparable to approaches that train on unsupervised proposals and box annotations, while outperforming approaches using video labels or zero-shot information considerably.

5.4 Comparison to state-of-the-art

We compare our results on three action localization datasets to the current state-of-the-art. In Table 2, we show the performance of the methods, ordered by their level of supervision. To maximize the number of comparisons, we show the results at a threshold of 0.2.

On all datasets, we perform comparable to approaches that rely on expensive manual box or point annotation during training and unsupervised proposals during testing. This result is encouraging, as it states that video labels and automatic pseudo-annotations can provide enough information for localization. We improve over approaches using only video labels [4] or zero-shot information [8], resulting in state-of-the-art performance on the Hollywood2Tubes dataset. We also compare against the approaches of Weinzaepfel *et al.* [30] and Saha *et al.* [20]. On UCF Sports, we achieve comparable AUC scores. On UCF-101, these approaches report higher scores. While effective, these approaches require full box supervision both for making proposals and training action classifiers. They can therefore not generalize to weaker forms of supervision.

6 Conclusions

In this work, we introduce pseudo-annotations for localizing actions in videos. We investigate pseudo-annotations from person detection, independent motions, action proposals, center biases, and object proposals. Using a correlation metric for pseudo-annotations, we reach results comparable or better to using full box supervision with the same settings, while outperforming other weakly-supervised approaches. As our approach relies on action class labels as the only manual annotations, it enables action localization on any action classification dataset, such as Sports 1M [11], ActivityNet [2], and EventNet [33].

Acknowledgements

This research is supported by the STW STORY project.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[3] Wei Chen and Jason J Corso. Action detection by implicit intentional motion clustering. In *ICCV*, 2015.

[4] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

[5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[6] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *CVPR*, 2015.

[7] Mihir Jain, Jan C van Gemert, Hervé Jégou, Patrick Bouthemy, and Cees G M Snoek. Action localization with tubelets from motion. In *CVPR*, 2014.

[8] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees G M Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, 2015.

[9] Mihir Jain, Jan C van Gemert, and Cees G M Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, 2015.

[10] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016.

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[12] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[13] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[15] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *CVPR*, 2009.

[16] Pascal Mettes, Jan C van Gemert, and Cees G M Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016.

[17] Dan Oneata, Jérôme Revaud, Jakob Verbeek, and Cordelia Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[19] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[20] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip H S Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *BMVC*, 2016.

[21] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012.

[23] Khurram Soomro, Haroon Idrees, and Mubarak Shah. Action localization in videos through context walk. In *ICCV*, 2015.

[24] Waqas Sultani and Mubarak Shah. What if we do not have multiple videos of the same action?–video action localization using web images. In *CVPR*, 2016.

[25] Yicong Tian, Rahul Sukthankar, and Mubarak Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013.

[26] Po-He Tseng, Ran Carmi, Ian GM Cameron, Douglas P Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *JoV*, 9(7), 2009.

[27] Jan C van Gemert, Mihir Jain, Ella Gati, and Cees G M Snoek. Apt: Action localization proposals from dense trajectories. In *BMVC*, 2015.

[28] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[29] Limin Wang, Yu Qiao, and Xiaoou Tang. Video action detection with relational dynamic-poselets. In *ECCV*, 2014.

[30] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, 2015.

[31] C. Xu and J. J. Corso. Actor-action semantic segmentation with grouping-process models. In *CVPR*, 2016.

[32] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, 2015.

[33] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *MM*, 2015.

[34] Gang Yu and Junsong Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015.

[35] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.