Spot On: Action Localization from Pointly-Supervised Proposals

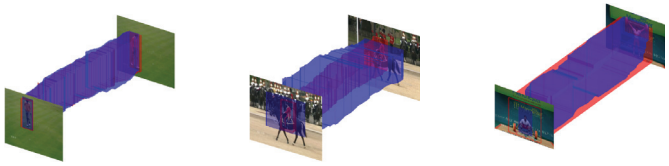
Pascal Mettes
University of Amsterdam

Jan C. van Gemert
Delft University of Technology

Cees G.M. Snoek
University of Amsterdam

Introduction

Discover *what* actions occur *when* and *where* in videos.



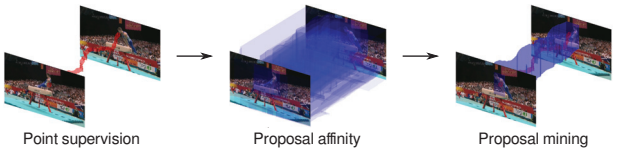
Current bottleneck in action localization:
Annotating bounding boxes for each frame of each action is expensive.
Therefore, current datasets are sparse and with few videos.

Our hypothesis:
Training on bounding box annotations is not required.
Training on unsupervised proposals with fast point annotations is as effective.

Method

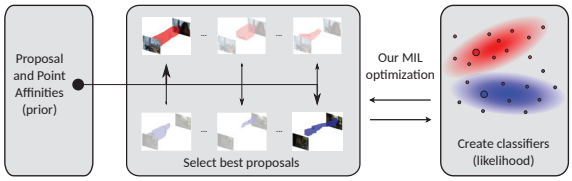
Method overview

Start from points annotated on the frames containing the action.
Train classifier on best action proposals, guided by point annotations.



Mining the best proposals

We start from Multiple Instance Learning and incorporate information from points.
We compute an affinity between each proposal and the point annotations.
We iteratively select best proposals (using affinities as prior) and retrain classifiers.



Formal objective:

$$\min_{\mathbf{w},b,\xi} \tfrac{1}{2}||\mathbf{w}||^2 + \lambda \textstyle\sum_i \xi_i$$

Max-margin objective

$$\text{s.t.} \quad \forall_i : y_i \cdot (\mathbf{w} \cdot \operatorname{argmax}_{\mathbf{z} \in x_i} P(\mathbf{z}|\mathbf{w}, b, A_i, C_i, V_i) + b) \geq 1 - \xi_i,$$
$$\forall_i : \xi_i \geq 0$$

Proposal selection function

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

Select the proposal with the best combined classifer score and affinity score.

Proposal and Point Affinity

Novel overlap measure between a proposal (A) and points (C) in video (V).
Used to guide the selection of best proposals to train on.

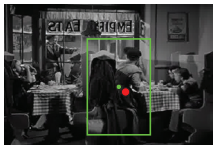
$$O(A, C, V) = M(A, C) - S(A, V)$$

Match score:
Notion: point annotations should be near the center of the proposal boxes.
The match score is the inverted distance between points and proposal centers.
The distance is divided by the distance between the proposal center and edge.

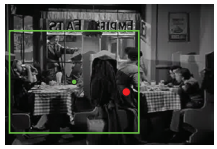
Distance between point and center of proposal box.

$$M(A, C) = \tfrac{1}{|C|} \textstyle\sum_{i=1}^{|C|} \max(0, 1 - \frac{||(x_i, y_i) - \mathrm{c}(BB_{k_i})||}{\max_{(u,v) \in \mathrm{e}(BB_{k_i})} ||(u, v) - \mathrm{c}(BB_{k_i})||_2})$$

Distance between proposal box center and edge.



High match.



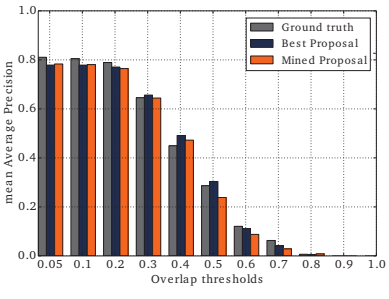
Low match.

Size regularization:
The relative size of the proposal (A), compared to the whole video (V).
This penalty alleviates the bias of large proposals towards the video center.
Actions also tend to occur near the center of the video.

$$S(A, V) = \left(\frac{\sum_{i=m}^{n} |A_i|}{|V|}\right)^2$$

Results (UCF Sports, UCF 101 in paper)

1: Training on proposals vs. ground truth boxes

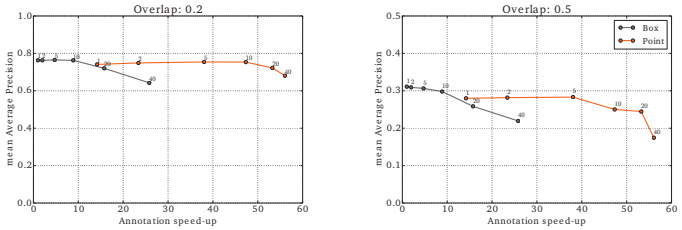


Best possible proposal performs similar to full ground truth boxes.

Scores maintained with pointly-supervised proposals.

Results holds across overlaps and datasets.

2: Lowering the annotation frame rate



Points are 15 to 50 times faster to annotate than bounding boxes.
Scores are maintained when annotating 10% of the frames.

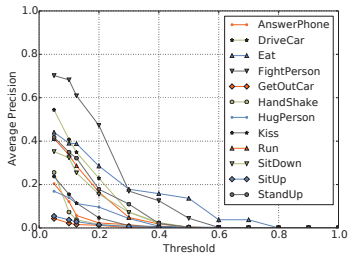
3: Introducing Hollywood2Tubes


Multi-label videos


Group interactions


Contextual actions


Occlusions



New dataset to demonstrate how easy action annotation becomes.
Contains several actions and instances new to action localization.

4: Comparison to state-of-the-art

Method	Supervision	AUC
Lan et al. *ICCV'11*	box	0.380
Tian et al. *CVPR'13*	box	0.420
Jain et al. *CVPR'14*	box	0.489
van Gemert et al. *BMVC'15*	box	0.546
Soomro et al. *ICCV'15*	box	0.550
Gkioxari et al. *CVPR'15*	box	0.559
Weinzaepfel et al. *ICCV'15*	box	0.559
Jain et al. *ICCV'15*	zero-shot	0.232
Cinbis et al. *CVPR'14*	video label	0.278
This work	**point**	**0.545**

Both use same action proposals and features.

Points are competitive to boxes, improve over other weak supervisions.