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

Goal

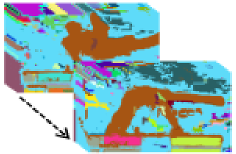
Kissing

Shaking hands

Related work: Action proposals at *test* time

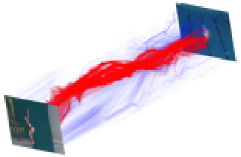
Supervoxels

Jain et al. *CVPR’14*
Oneata et al. *ECCV'14*



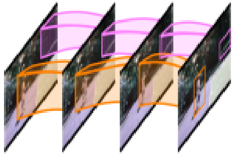
Trajectories

van Gemert et al. *BMVC’15*
Puskas et al. *ICCV'15*

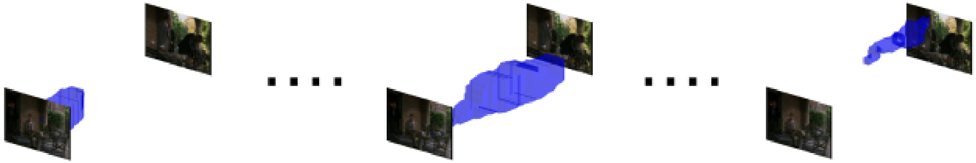


Tracking/detection

Yu et al. *CVPR'15*
Weinzaepfel et al. *ICCV'15*



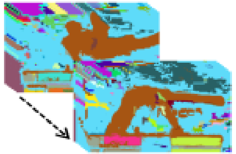
Action proposals



Related work: Action proposals at *test* time

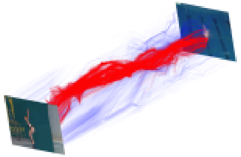
Supervoxels

Jain et al. *CVPR’14*
Oneata et al. *ECCV'14*



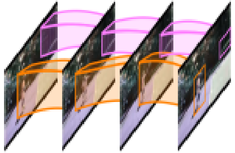
Trajectories

van Gemert et al. *BMVC’15*
Puskas et al. *ICCV'15*

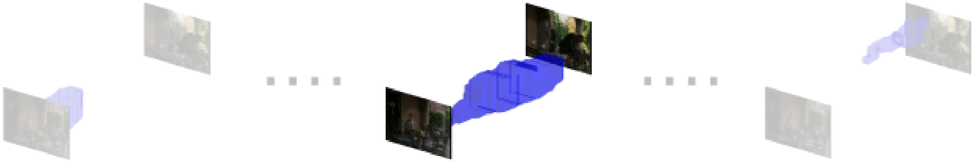


Tracking/detection

Yu et al. *CVPR'15*
Weinzaepfel et al. *ICCV'15*

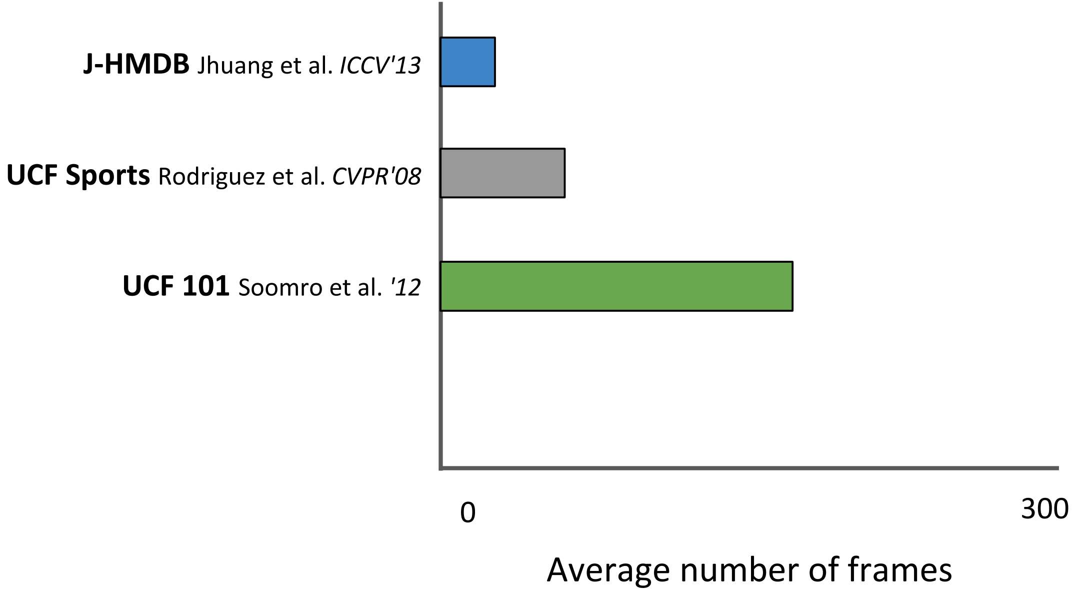


Action proposals



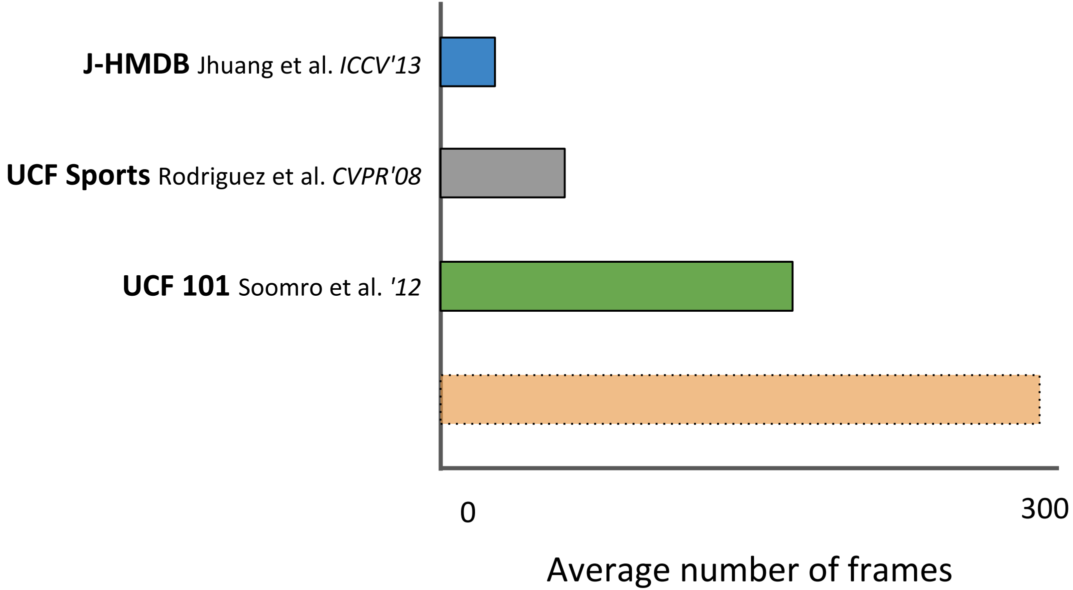
Related work: Training from box supervision

Annotate boxes for *each* frame of *each* train video.



Related work: Training from box supervision

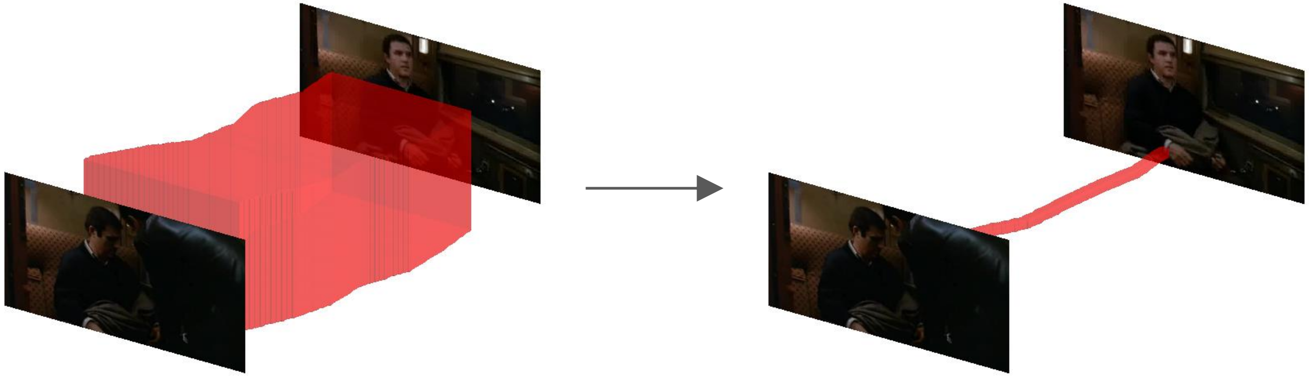
Annotate boxes for *each* frame of *each* train video.



Our hypothesis

Training on bounding boxes not required.

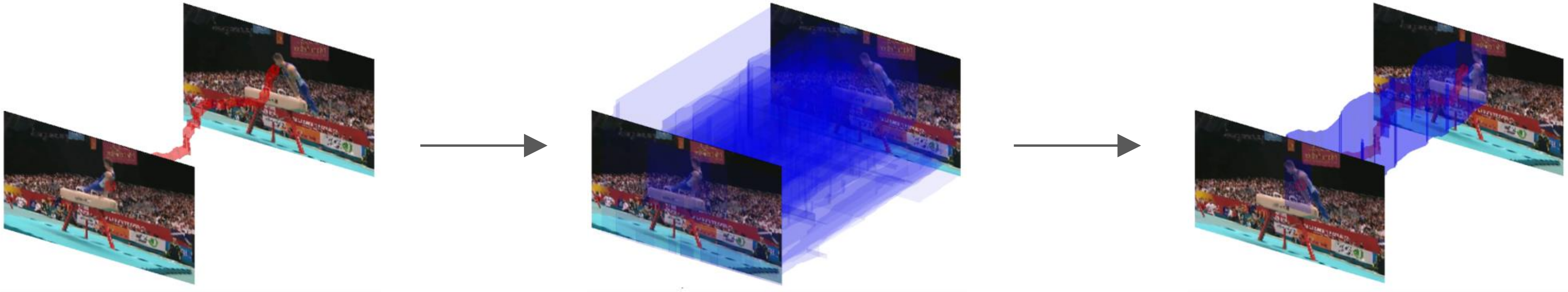
Training on proposals with fast point annotations is as effective.



Annotation time for video: 5 min. 11 sec.

Annotation time for video: 25 sec.

Our contribution

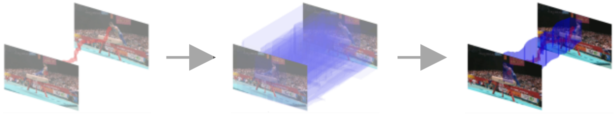


Human point supervision

Compute proposal affinity

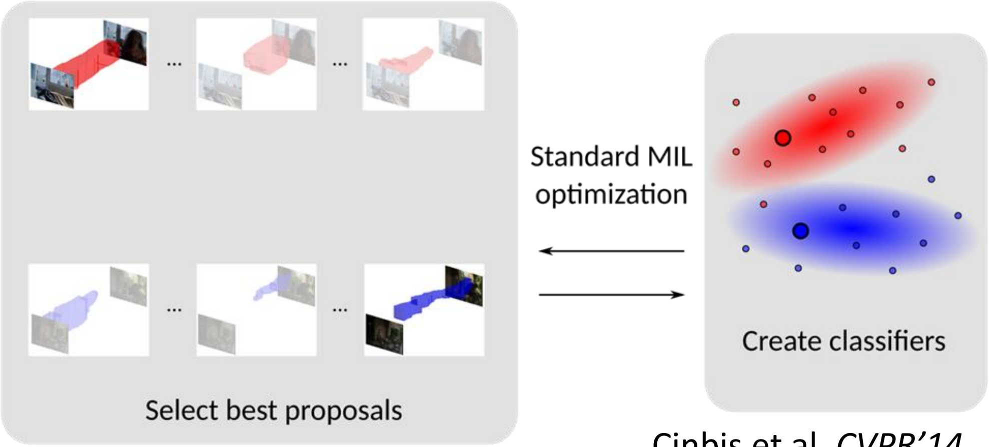
Mine best proposal

Mining the best proposals



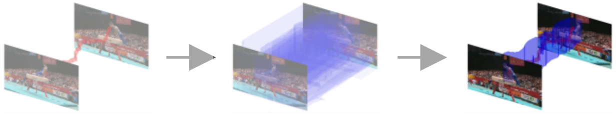
Train action classifiers using only best proposals.

Casted as a Multiple Instance Learning problem.



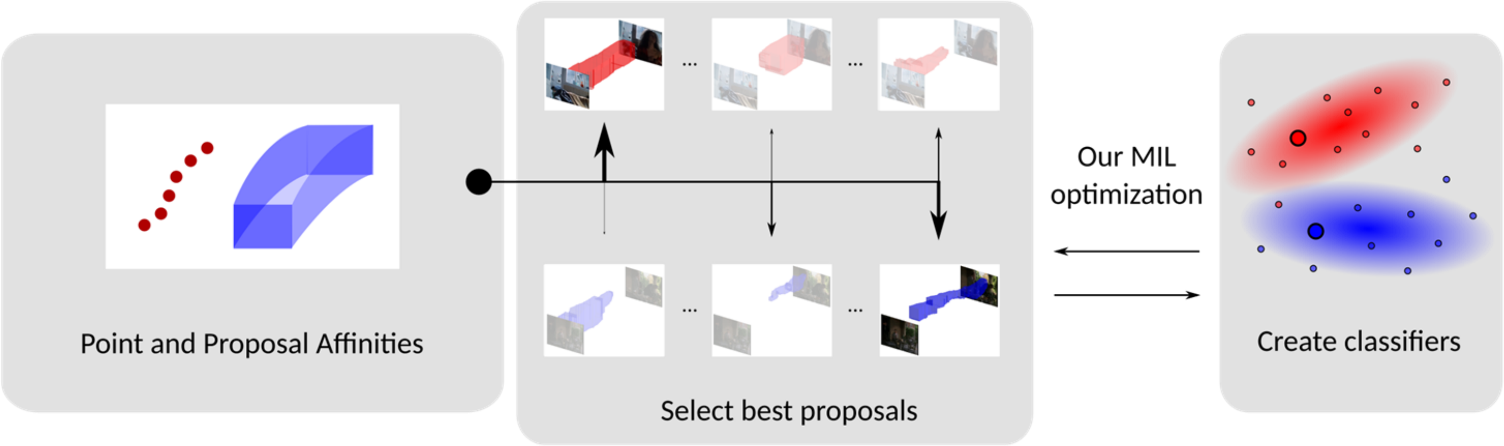
Cinbis et al. *CVPR'14*

Mining the best proposals



Train action classifiers using only best proposals.

Casted as a Multiple Instance Learning problem.



Point and Proposal Affinities

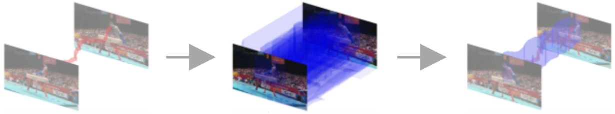
Select best proposals

Our MIL optimization

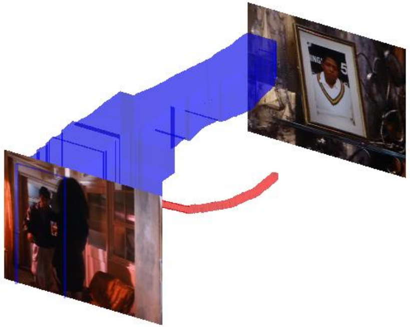
Create classifiers

Use affinity with point annotations to guide the mining.

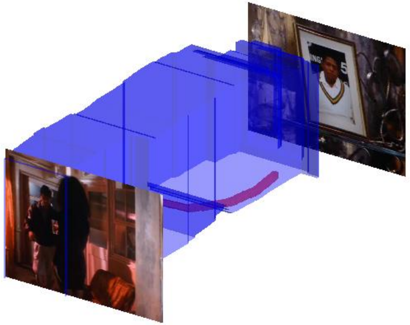
Proposal affinity



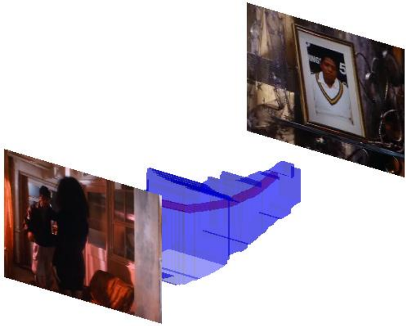
Novel overlap measure between point annotations and proposals.



No overlap

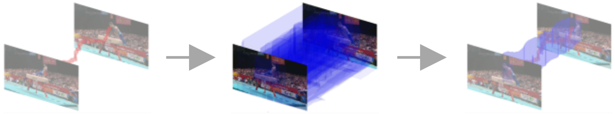


Small overlap



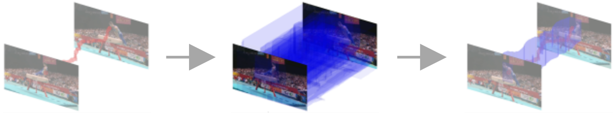
High overlap

Proposal affinity



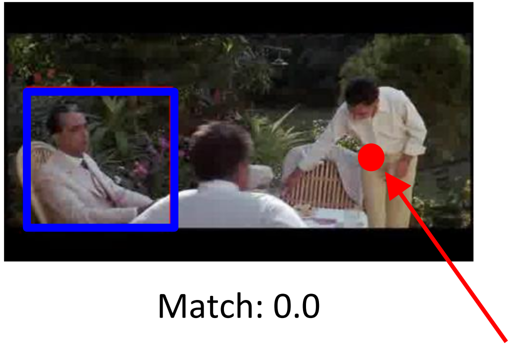
Affinity = Proposal Match - Size Regularization

Proposal match

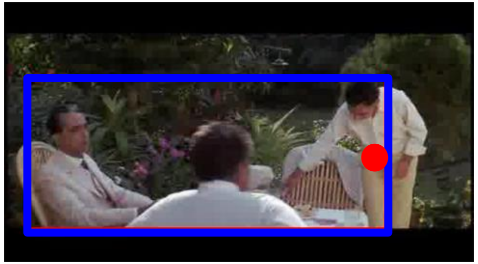


Affinity = Proposal Match - Size Regularization

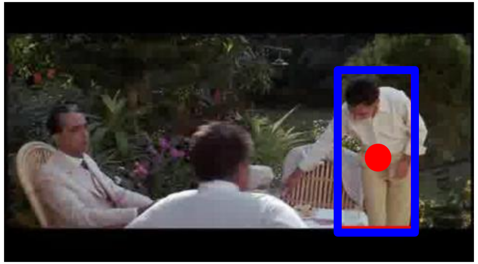
Each point should match with the center of the proposal.
Average the matches over all the points.



Match: 0.0

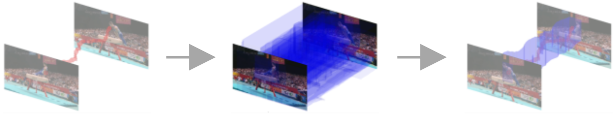


Match: 0.1



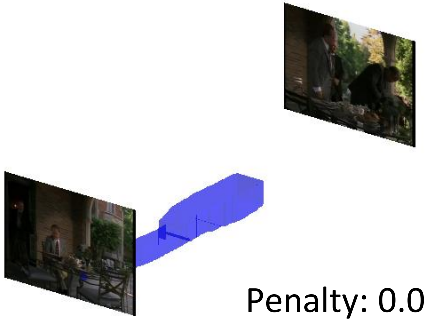
Match: 1.0

Size regularization

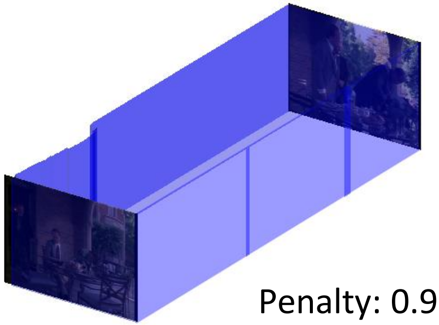


Affinity = Proposal Match - Size Regularization

Subtract the size of the proposal from the match.
To alleviate center bias of large proposals.

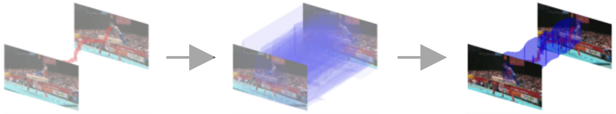


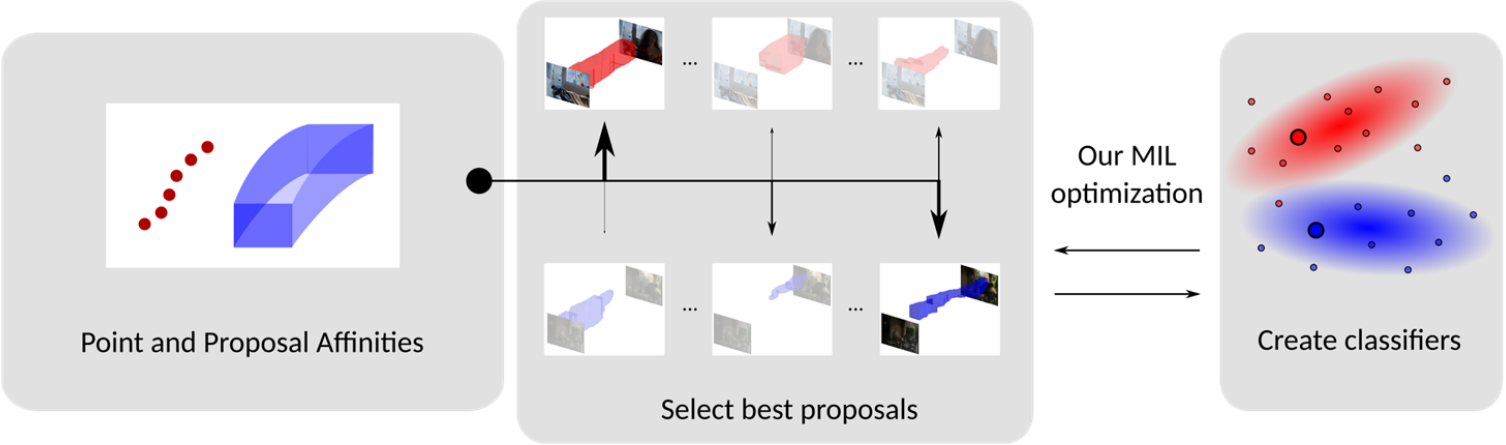
Penalty: 0.05



Penalty: 0.90

Mining recap





Point and Proposal Affinities

Select best proposals

Our MIL optimization

Create classifiers

Experiments

UCF Sports

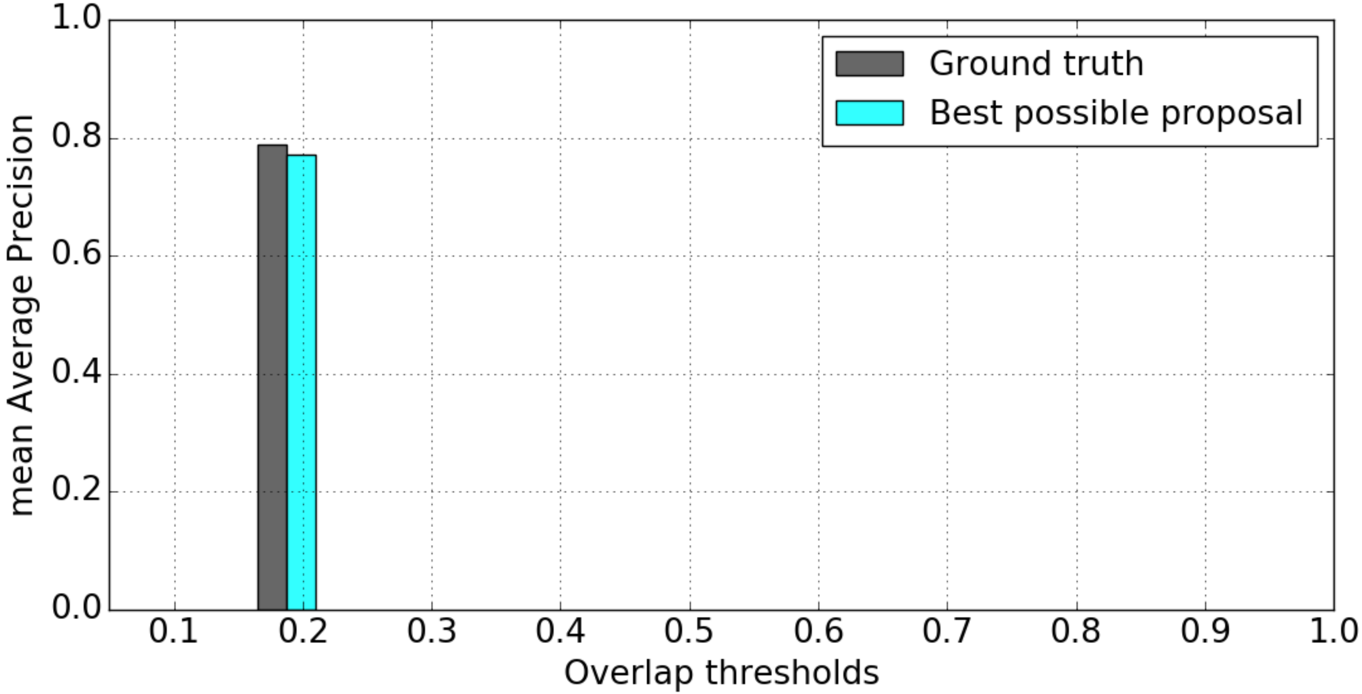
UCF 101

Unsupervised proposals from clustered trajectory features.

Evaluated with Fisher Vectors and SVMs.

van Gemert et al. *BMVC'15*

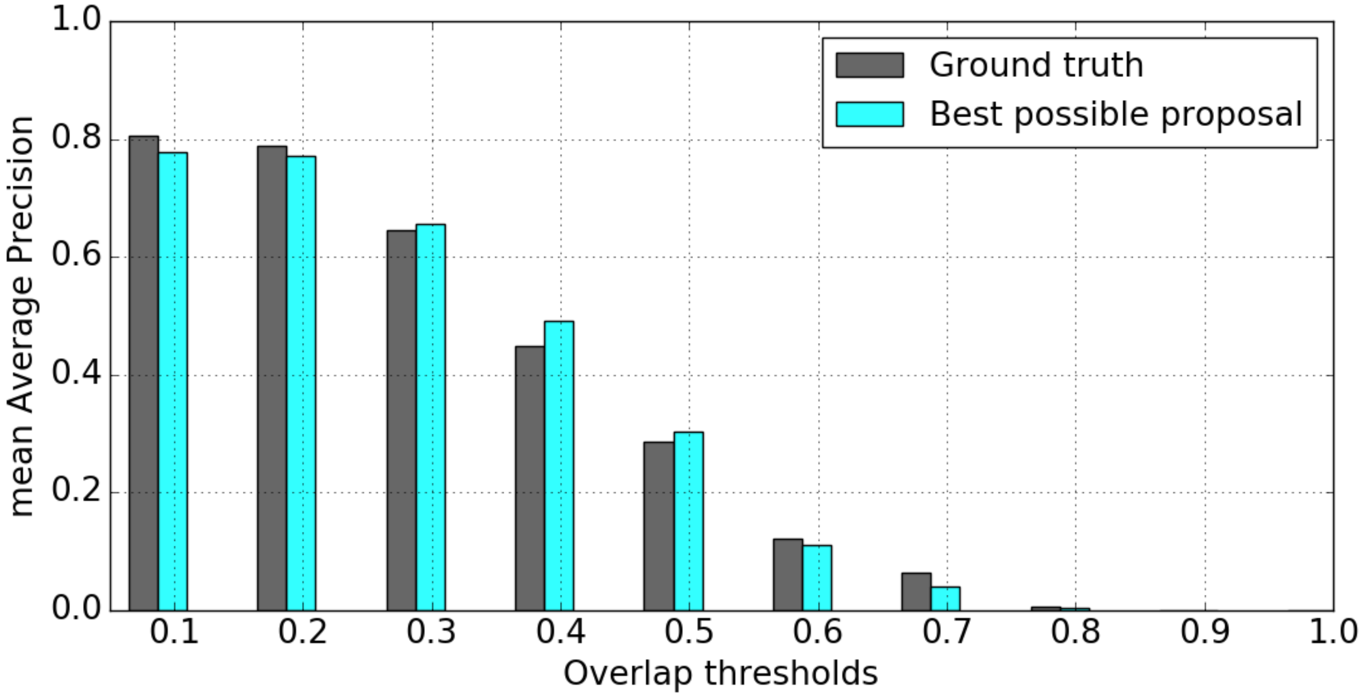
Training without ground truth tubes



UCF Sports

Best possible proposal performs as well as ground truth tubes.

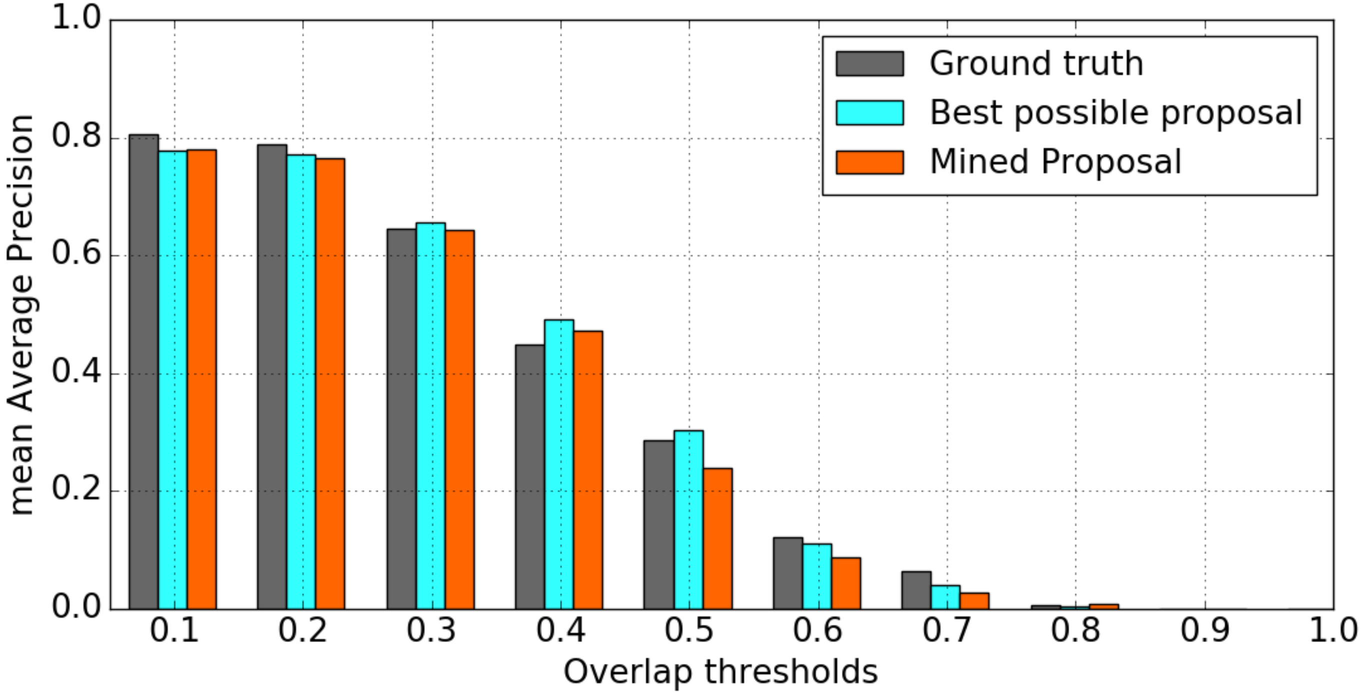
Training without ground truth tubes



UCF Sports

Best possible proposal performs as well as ground truth tubes.

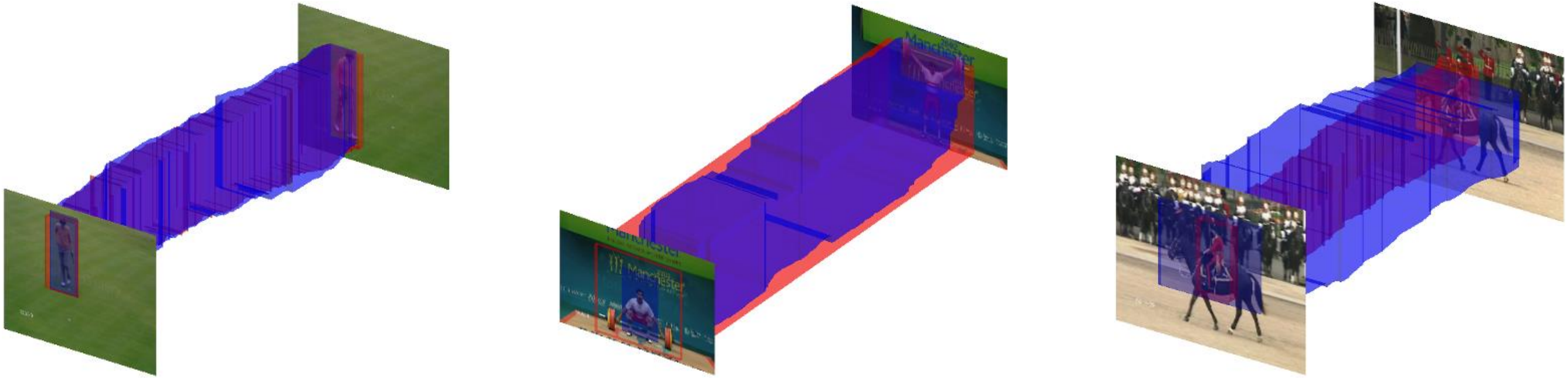
Training without ground truth tubes



UCF Sports

Mean AP maintained using our mined proposals.

Training without ground truth tubes

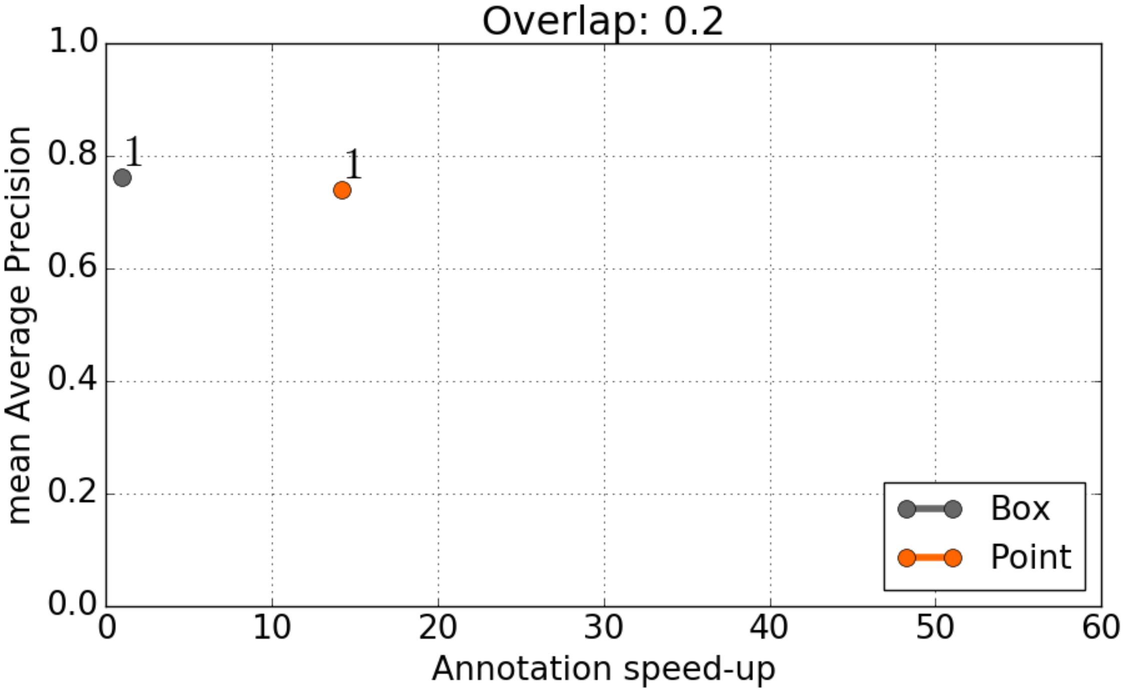


🟥 Ground truth box annotations

🟦 Our mined proposal

Similar performance from different tubes.

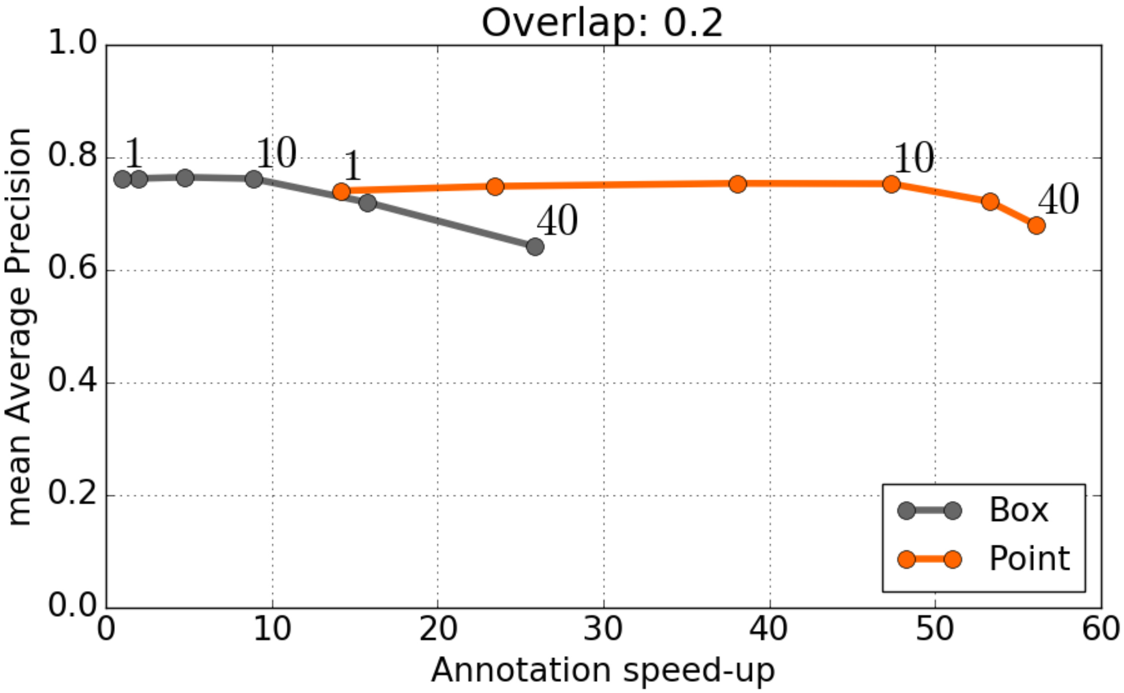
Lowering the annotation frame-rate



UCF Sports

Points as effective as boxes, while faster to annotate.

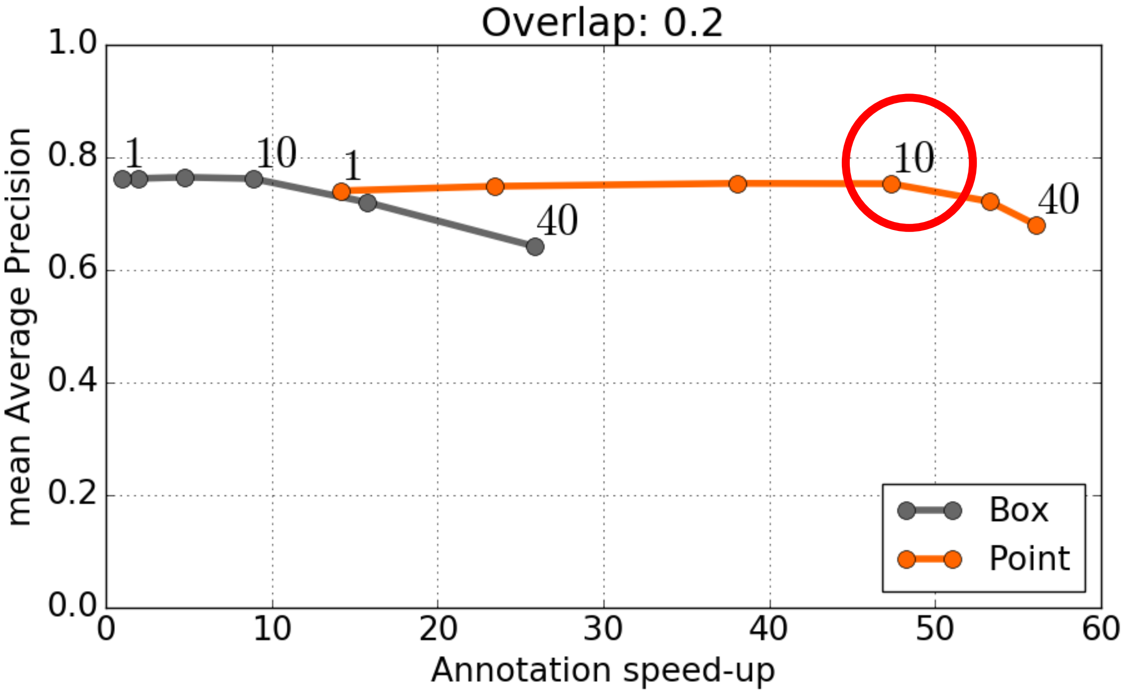
Lowering the annotation frame-rate



UCF Sports

Up to 50 times speed-up at similar performance.

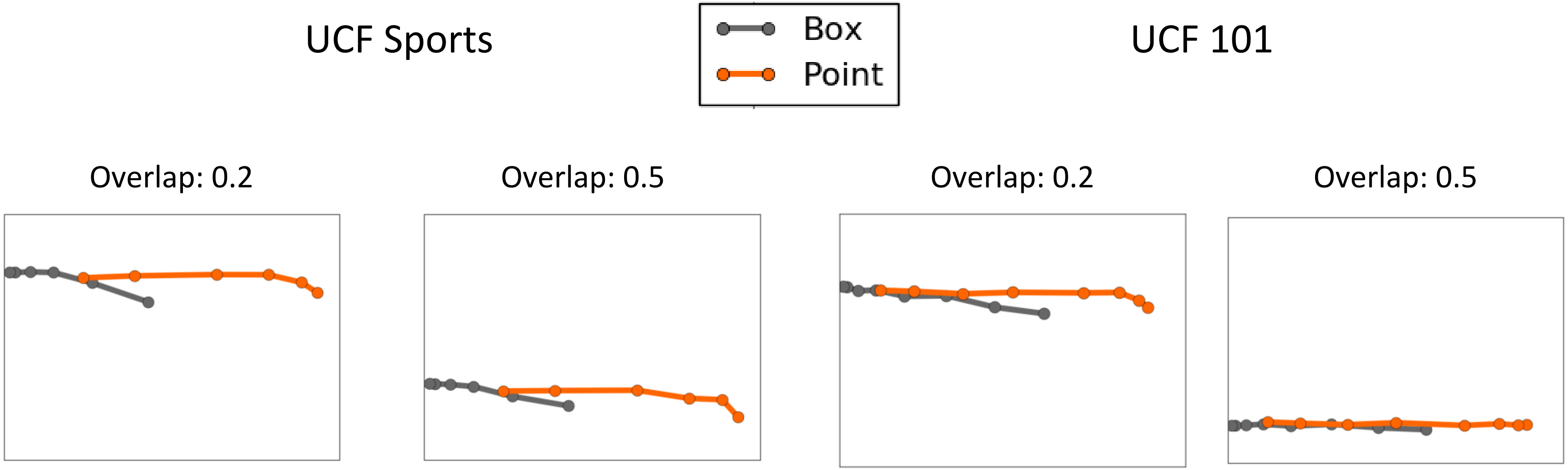
Lowering the annotation frame-rate



UCF Sports

Up to 50 times speed-up at similar performance.

Lowering the annotation frame-rate



Points are fast. Competitive even at 10% annotation effort.

Hollywood2Tubes

Download: tinyurl.com/hollywood2tubes

Dataset to demonstrate how easy action annotation becomes.

Contains actions and instances new to action localization.

a.

Multi-label videos.

b.

Contextual actions.

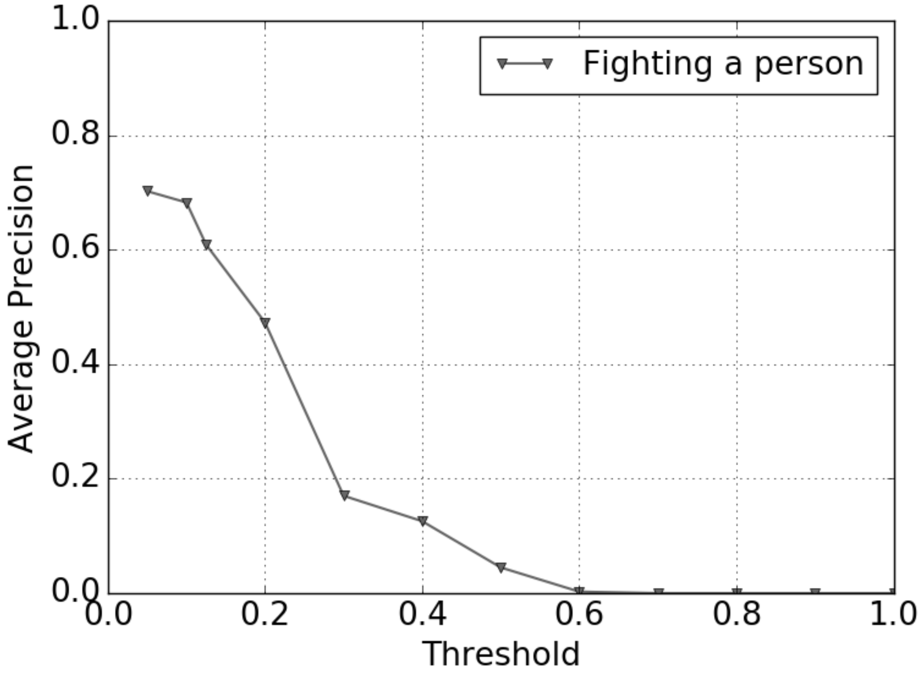
c.

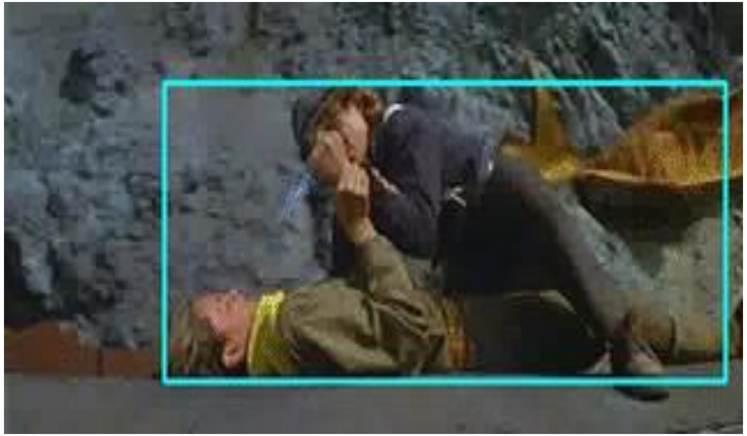
Group interactions.

Videos from Hollywood2 by Marszalek et al. *CVPR'09*

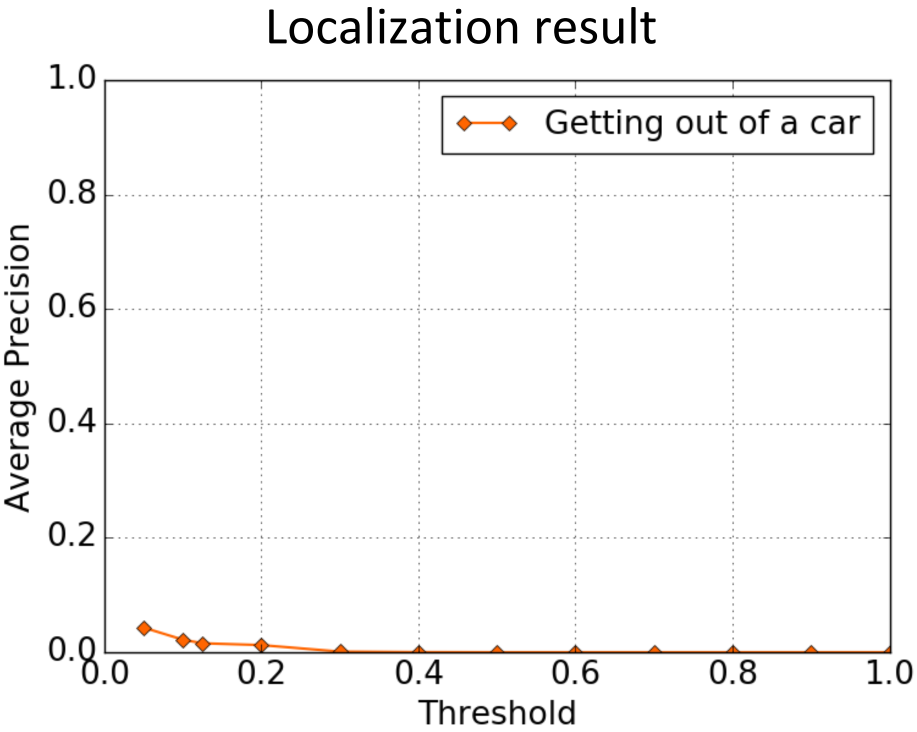
Hollywood2Tubes - good example

Localization result



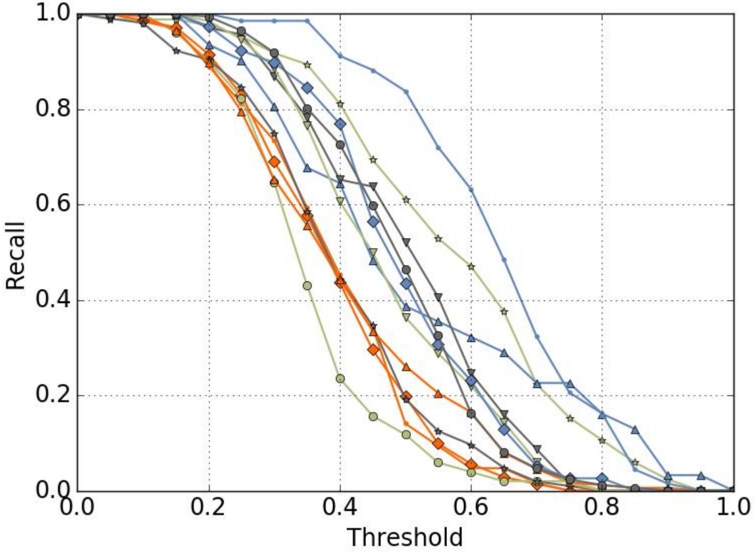


Hollywood2Tubes - bad example

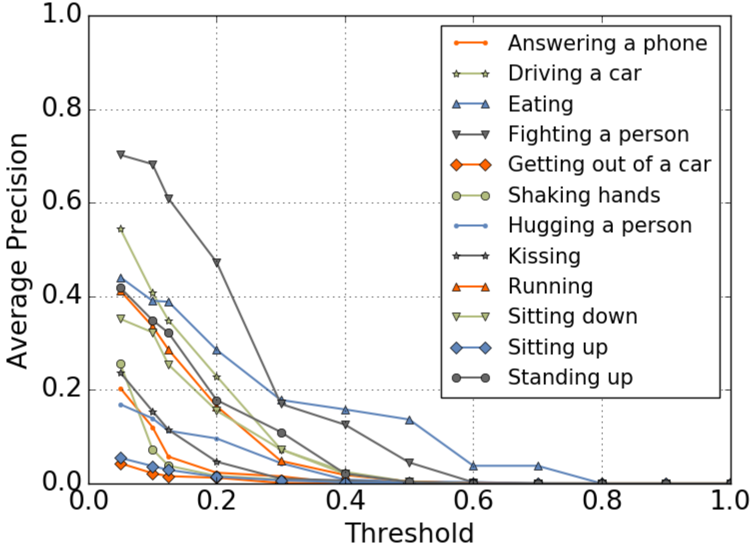


Hollywood2Tubes - Overall performance

Proposal quality



Localization result



Action recall comparable to current datasets.

Action localization leaves much room for improvement.

Conclusions

Carefully annotated boxes not required for action localization.

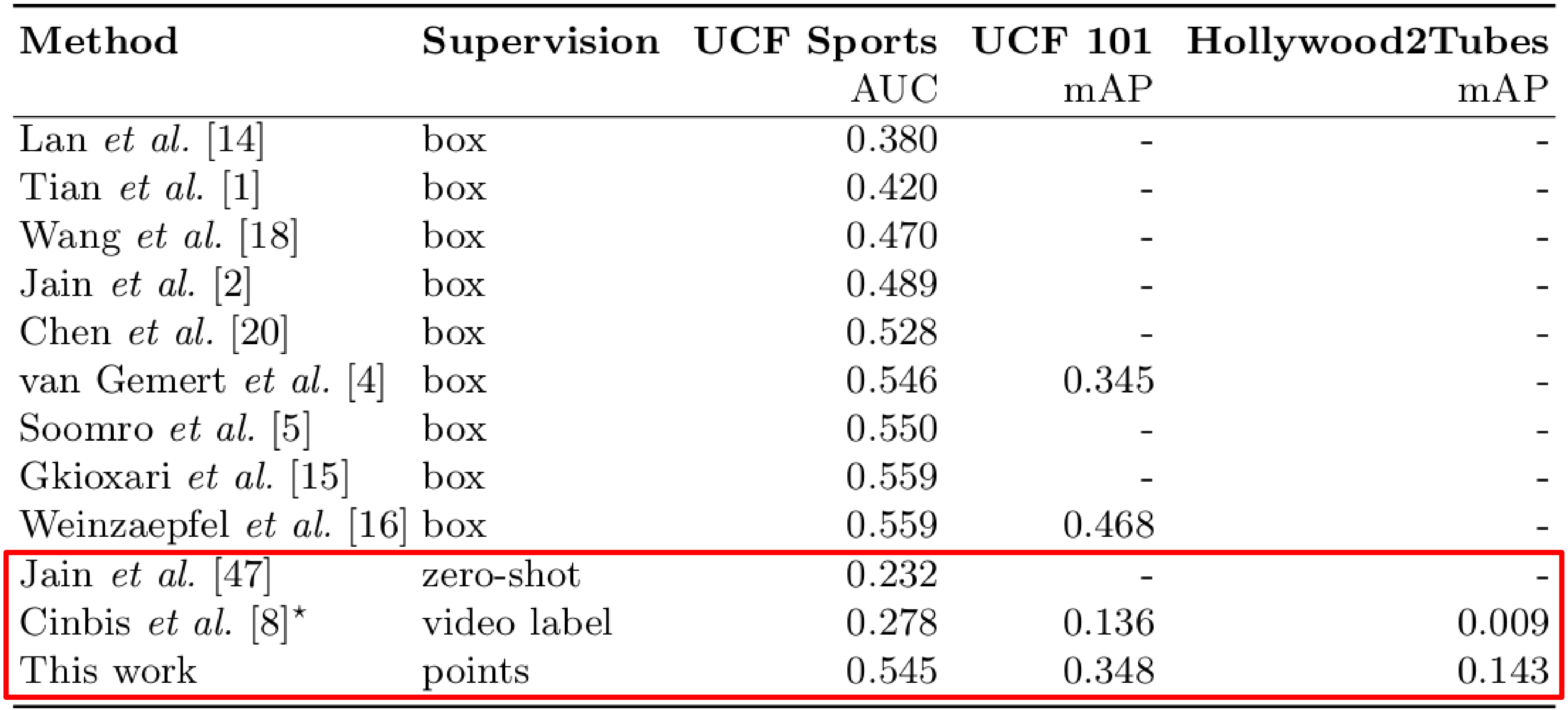
We train on unsupervised proposals, guided by point annotations.

We introduce Hollywood2Tubes.

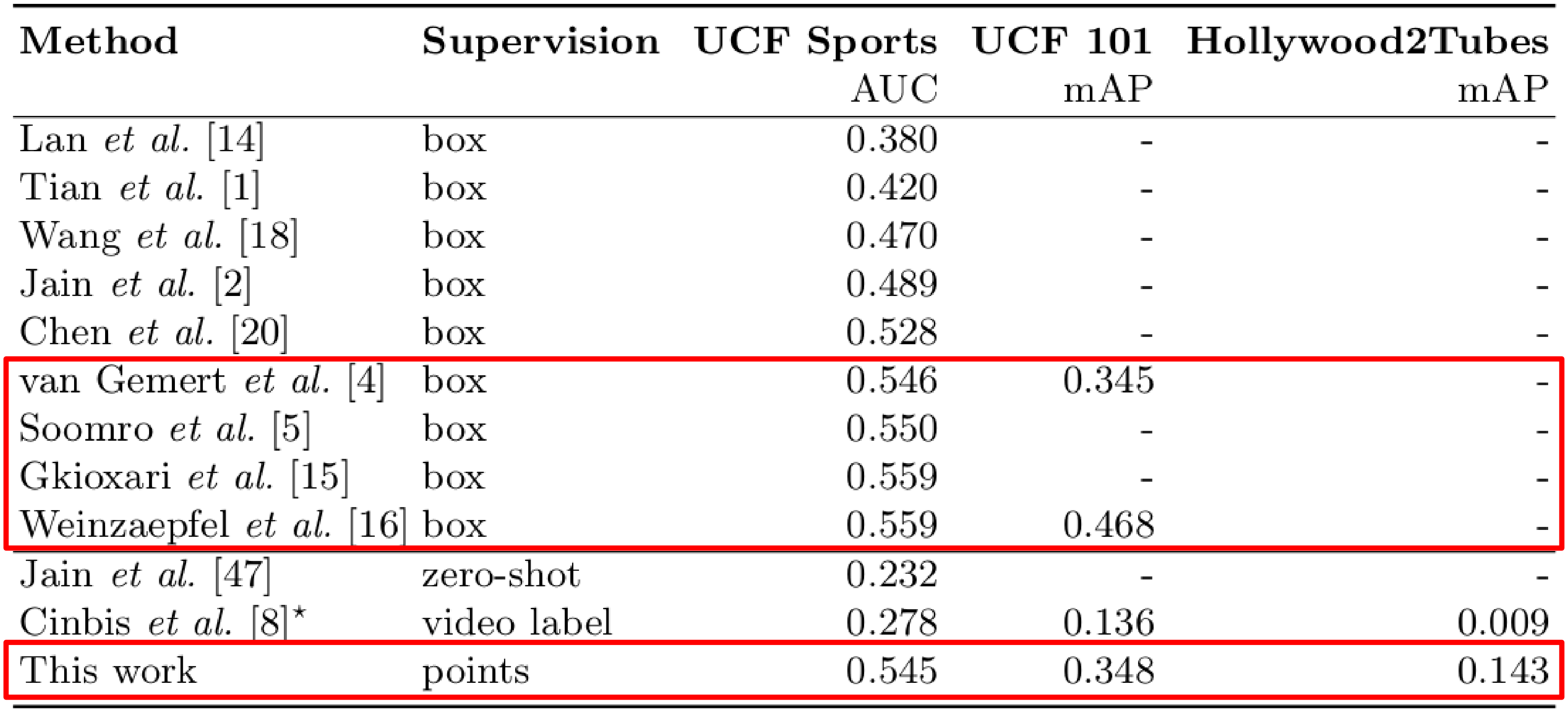
Mail: P.S.M.Mettes@uva.nl

Poster: **O-4B-01** (roof garden)

Method	**Supervision**	**UCF Sports**	**UCF 101**	**Hollywood2Tubes**
		AUC	mAP	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

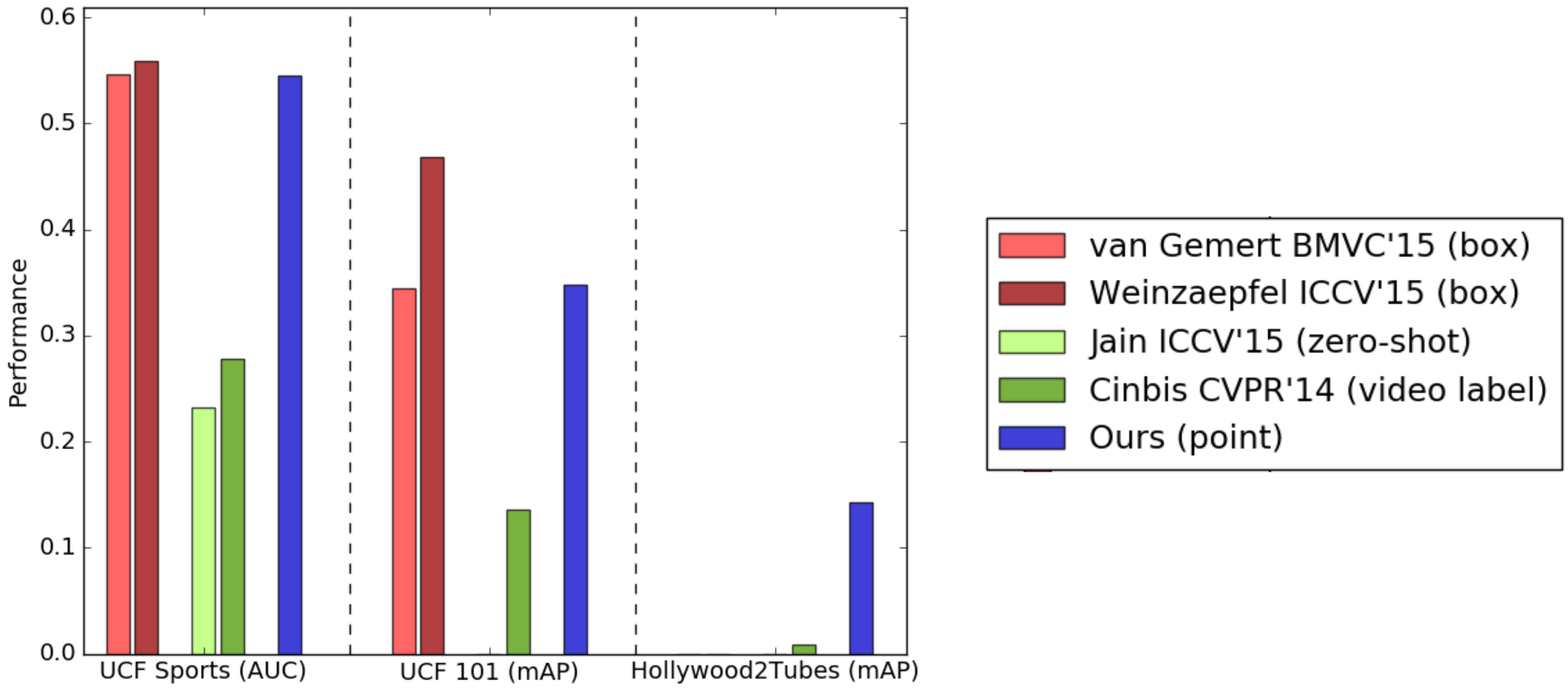


Method	**Supervision**	**UCF Sports**	**UCF 101**	**Hollywood2Tubes**
		AUC	mAP	mAP
Lan *et al.* [14]	box	0.380	-	-
Tian *et al.* [1]	box	0.420	-	-
Wang *et al.* [18]	box	0.470	-	-
Jain *et al.* [2]	box	0.489	-	-
Chen *et al.* [20]	box	0.528	-	-
van Gemert *et al.* [4]	box	0.546	0.345	-
Soomro *et al.* [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



Method	**Supervision**	**UCF Sports**	**UCF 101**	**Hollywood2Tubes**
		AUC	mAP	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

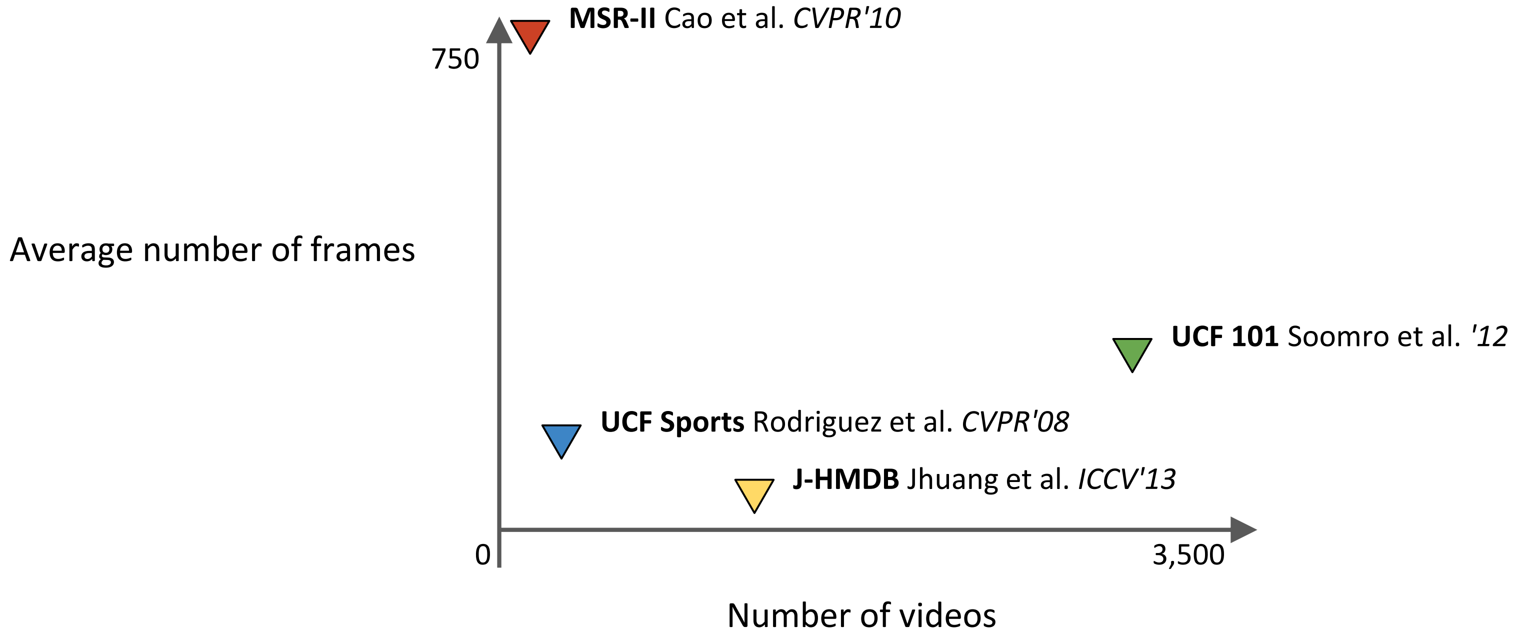
4- Comparison to state-of-the-art



Competitive to boxes, better than other weak supervision.

Related work: Training from box supervision

Annotate boxes for *each* frame of *each* video.


MSR-II Cao et al. *CVPR'10*
Average number of frames
UCF 101 Soomro et al. *'12*
UCF Sports Rodriguez et al. *CVPR'08*
J-HMDB Jhuang et al. *ICCV'13*
Number of videos