APT: Action localization Proposals from dense Trajectories

Jan C. van Gemert[1,2]
j.c.vangemert@gmail.com
Mihir Jain[1]
m.Jain@uva.nl
Ella Gati[1]
ellagati@gmail.com
Cees G. M. Snoek[1,3]
cgmsnoek@uva.nl

[1] Intelligent Systems Lab Amsterdam
University of Amsterdam
Amsterdam, The Netherlands

[2] Computer Vision Lab
Technical University Delft
Delft, The Netherlands

[3] Qualcomm Research Netherlands
Amsterdam, The Netherlands

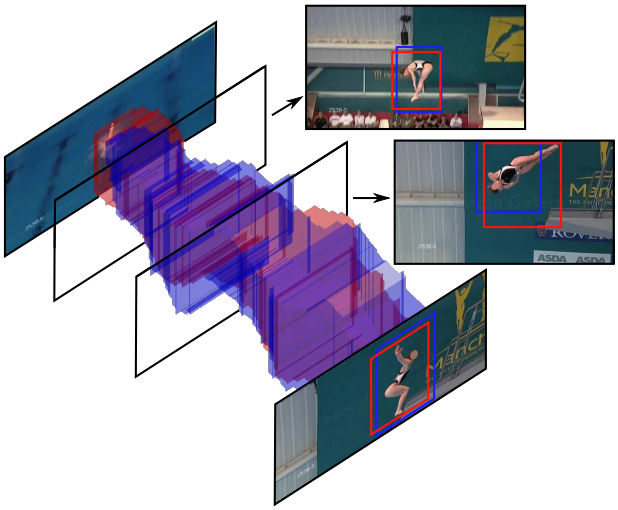

Figure 1: **We propose APT:** an efficient and effective spatio-temporal proposal algorithm for action localization. Different from existing work, which consider different representations for the localization and classification stage, we aptly re-use the dense trajectories as used in the classification representation for localization as well. This allows large-scale deployment with good localization and classification accuracy. For the action *Diving* our blue best proposal results in an overlap of 0.62 with the red ground truth.

	ABO	MABO	Recall	#Proposals
UCF Sports				
Brox & Malik, ECCV 2010 [1]	29.84	30.90	17.02	**4**
Jain et al., CVPR 2014 [3]	63.41	62.71	78.72	1,642
Oneata et al., ECCV 2014 [4]	56.49	55.58	68.09	3,000
Gkioxari & Malik, CVPR 2015 [2]	63.07	62.09	87.23	100
APT (ours)	**65.73**	**64.21**	**89.36**	1,449
UCF 101				
Brox & Malik, ECCV 2010 [1]	15.16	15.06	1.94	**3**
APT (ours)	**47.16**	**46.79**	**46.38**	2,299
MSR-II				
Brox & Malik, ECCV 2010 [1]	2.28	2.34	0	**6**
Jain et al., CVPR 2014 [3]	34.88	34.81	2.96	4,218
Yu & Yuan, CVPR 2015 [6]	n.a.	n.a.	0	37
APT (ours)	**48.02**	**47.87**	**44.33**	6,706

Table 1: **Comparing action proposals** for commonly used metrics against other methods, where *n.a.* denotes not reported values. *ABO* is the Average Best Overlap, *MABO* the Mean ABO over all classes, *Recall* is measured at an IoU overlap $\sigma \geq 0.5$ and the number of proposals is averaged per video. APT outperforms others with a modest number of proposals.

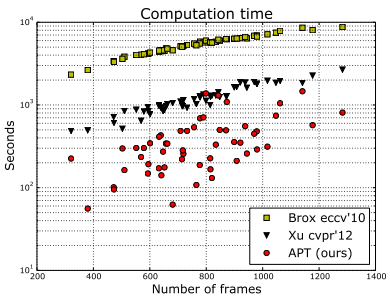

Figure 3: **Evaluating APT computation time** on all videos of MSR-II. APT is faster than the video segmentation of Xu & Corso [5] used in [3]. APT is an order of magnitude faster than the trajectory clustering of Brox & Malik [1].

This paper is on action localization in video with the aid of spatio-temporal proposals. To alleviate the computational expensive segmentation step of existing proposals, we propose bypassing the segmentations completely by generating proposals directly from the dense trajectories used to represent videos during classification. Our Action localization Proposals from dense Trajectories (APT) use an efficient proposal generation algorithm to handle the high number of trajectories in a video. Our spatio-temporal proposals are faster than current methods and outperform the localization and classification accuracy of current proposals on the UCF Sports, UCF 101, and MSR-II video datasets.

[1] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[2] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015.

[3] M. Jain, J. van Gemert, H. Jégou, P. Bouthemy, and C. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014.

[4] D. Oneata, J. Revaud, J. Verbeek, and C Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014.

[5] C. Xu and J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

[6] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015.

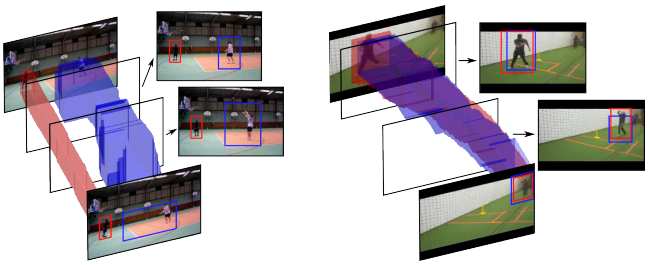

Figure 2: **Success and failure case of APT**. In the left video our method (blue tube) ignores the standing person (red tube) and tracks the moving actor. In the right video, there is ample variation in depth and position yet APT tracks the action well. Our method is intended for actions and thus works best when motion is present.

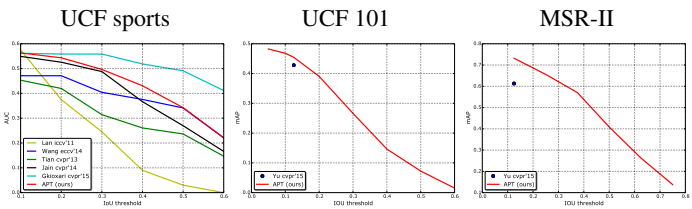

Figure 4: **Classifying action proposals** with a varying IoU threshold σ. Left: AUC on UCF Sports, the fully supervised method of Gkioxari & Malik [2] is best, we are competitive with the unsupervised state of the art. Middle: mAP for UCF 101, we outperform the recent scores on this dataset by Yu & Yuan [6]. Right: mAP for MSR-II, where we outperform the best results on this set by Yu & Yuan [6].