Supplementary Material: Incremental Tube Construction for Human Action Detection

Harkirat Singh Behl[1]
harkirat@robots.ox.ac.uk
Michael Sapienza[2]
m.sapienza@samsung.com
Gurkirt Singh[3]
gurkirt.singh-2015@brookes.ac.uk
Suman Saha[3]
suman.saha-2014@brookes.ac.uk
Fabio Cuzzolin[3]
fabio.cuzzolin@brookes.ac.uk
Philip H. S. Torr[1]
phst@robots.ox.ac.uk

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] Think Tank Team
Samsung Research America
Mountain View, CA

[3] Dept. of Computing and
Communication Technologies
Oxford Brookes University
Oxford, UK

1 Improving Computational Performance

It is interesting to note that on varying initiation threshold in the range 0.1 to 0.9, the variance in mAP @ $\delta = 0.75$ on UCF-101 is only 0.3%, where the mean is 32.1%, which proves that our algorithm is robust to spurious tube initiations. The threshold for terminating tubes is 5 consecutive missed detections, which takes care of spurious mis-detections by the network and short occlusions. On varying this in range 5 to 20, the variance is less than 1%. The constant c_o in Eqn. 3 in main paper is fixed to 10, but the results remain the same for all values of c_0 above 2.

2 Results on JHMDB-21

On J-HMDB-21, only one action category and tube are present in each video, which is why it is the easiest of the three datasets that we evaluate on. Also, it only has trimmed videos, thus there is no concept of temporal localization, and it is limited to spatial localization only. [2] outperforms our method on JHMDB. It is interesting to note that despite falling behind [2, 3] on this subproblem, our method performs at par with them on overall results in UCF-101 and LIRIS-HARL; this is because our algorithm tackles the larger problem of action-detection in real-life scenarios, and is not overfitted to perform well on any subproblem. [2, 3] remove the labelling and temporal trimming terms for JHMDB. Whereas, our algorithm does not make any such assumptions for any specific dataset.

Table 1: Quantitative action detection results on J-HMDB-21.

mAP @ space-time overlap threshold δ	.5	.6	.7	.5:.95
STMH [8]	60.7	–	–	–
Saha *et al.* [2]	71.50	**68.73**	**56.57**	–
Singh *et al.* *et al.* [3]†	72.00	–	–	41.60
Kalogeiton *et al.* [4]†	**73.70**	–	–	**44.80**
Ours (OJLA)†	60.27	57.82	49.77	34.59
Ours (OJLA with multiple labels)†	67.29	61.48	49.09	36.08

† Online

3 More Results on UCF-101

Table. 2 shows the comparison of our method to [2, 3] in terms of speed. Fig. 1 shows another example where the method of [2] predicts multiple overlapping tubes in a region where only one action is happening. In contrast, our method predicts tubes(non-overlapping) with a single label.

Table 2: Test-time speed comparison.

linking speed in fps on UCF-101	speed	
Saha *et al.* [2]	65 fps	Batch processing and Offline
Kalogeiton *et al.* [4]	300fps	Incremental and Online
Singh *et al.* [3]	400fps	Incremental and Online
OJLA	**550 fps**	**Incremental and Online**

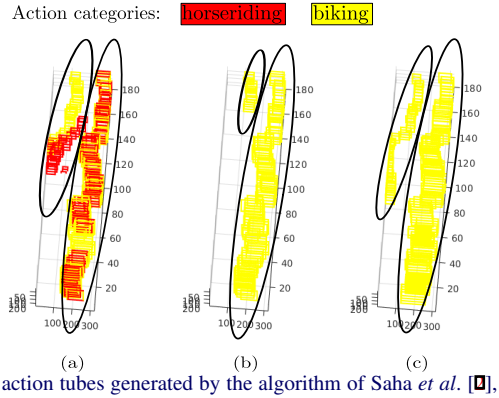

Figure 1: **(a)** The quality of action tubes generated by the algorithm of Saha *et al.* [2], compared to **(b)** the action tubes generated by our method with an improved cost function. **(c)** The ground truth action tubes.

4 Ablation Studies on LIRIS-HARL

In Table 3, we show the per-class video AP at threshold $\delta = 0.4$. Table 4 shows the contribution of various components to the results. It can be seen that fusing the results of the flow network with the detection network helps a lot towards the video mAP. Fig.2 shows the qualitative results of our method on yet another video from LIRIS-HARL

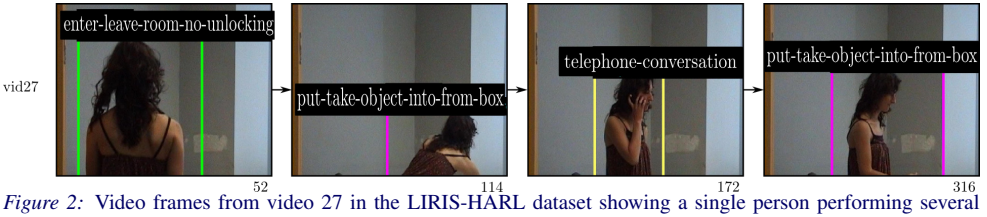

Figure 2: Video frames from video 27 in the LIRIS-HARL dataset showing a single person performing several actions. From left, the person first 'enters the room' (frame 52), then has a 'telephone conversation' (frame 172). Before and after the telephone conversation, the human is detected as 'putting/taking an object from a box'. The full video sequence is shown in the supplementary video.

Table 3: Per class video-AP comparison on LIRIS-HARL at $\delta = 0.4$.

Action	discussion	give_object_to_person	put_take_obj_into_from_box_desk	enter_leave_room_no_unlocking	try_enter_room_unsuccessfully
Ours (with OJLA)	48.67	0	29.83	14.01	57.14
Ours (with OJLA multilabel)	44.74	8.33	9.05	3.16	77.42
Action	**unlock_enter_leave_room**	**leave_baggage_unattended**	**handshaking**	**typing_on_keyboard**	**telephone_conversation**
Ours (with OJLA)	49.21	14.69	25.33	44.47	8.33
Ours (with OJLA multilabel)	94.64	11.14	0	25.44	0

Table 4: Quantitative action detection results on LIRIS-HARL.

mAP @ space-time overlap threshold δ	.4	.5	.6	.75	.5:.95
Ours (only app with OJLA)	20.14	16.42	11.02	2.51	5.86
Ours (only flow with OJLA)	23.99	15.92	7.78	**3.09**	5.05
Ours (with OJLA)	**29.17**	**21.77**	**9.59**	2.48	**6.16**

References

[1] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017.

[2] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *Proc. British Machine Vision Conference*, 2016.

[3] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, 2017.

[4] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, June 2015.