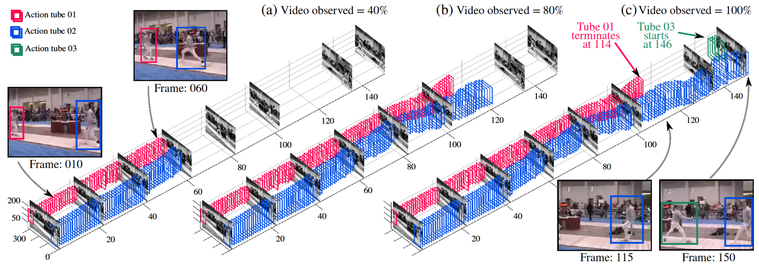
Online Real-time Multiple Spatiotemporal Action Localisation and Prediction

Gurkirt Singh[1] Suman Saha[1] Michael Sapienza[2*] Philip Torr[2] Fabio Cuzzolin[1]

[1]Oxford Brookes University [2]University of Oxford

Online Action Localisation and Prediction



Task is to determine what action is occurring in a video, as early as possible and localise it.

- **Action Localisation:** defined as a set of linked bounding boxes covering each individual action instance, called action tubes.
- **Online:** method designed to construct tubes incrementally.
- **Prediction:** to predict the label of the video at any given point of time, for e.g. when only 10% of the video has been observed.

Why?

- Real-time online action localisation is essential for many applications, for e.g. surveillance, human-robot interaction.
- Early action label prediction is crucial in interventional applications, for e.g. surgical robotics or autonomous driving.

Contributions

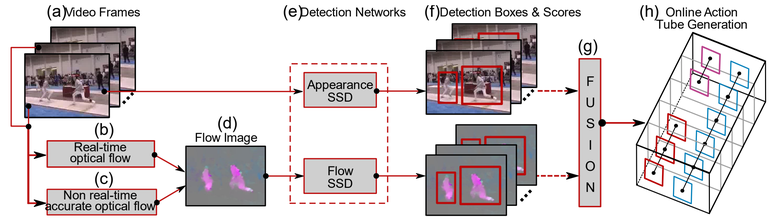
- Unlike previous methods [1,2,3], we construct multiple action tubes **simultaneously and incrementally**, getting rid of recursive calls to dynamic programming to generate multiple tubes.
- Slow optical flow [5] and the Faster RCNN detection network are replaced by real-time optical flow [6] and SSD [7] for **speed**.
- **First** method to perform online spatiotemporal action localisation in **real-time**, while still outperforming previous methods.
- Unlike [4], we perform early action localisation and prediction in **untrimmed** videos of the UCF101 dataset.

Early label prediction

- Early label prediction is a side product of our online tube generation algorithm.
- At any given point of time the video is assigned the label of the tube with highest score from the current set of tubes.

*Michael Sapienza is now with the Think Tank Team at Samsung Research America

Online and Real-Time Pipeline

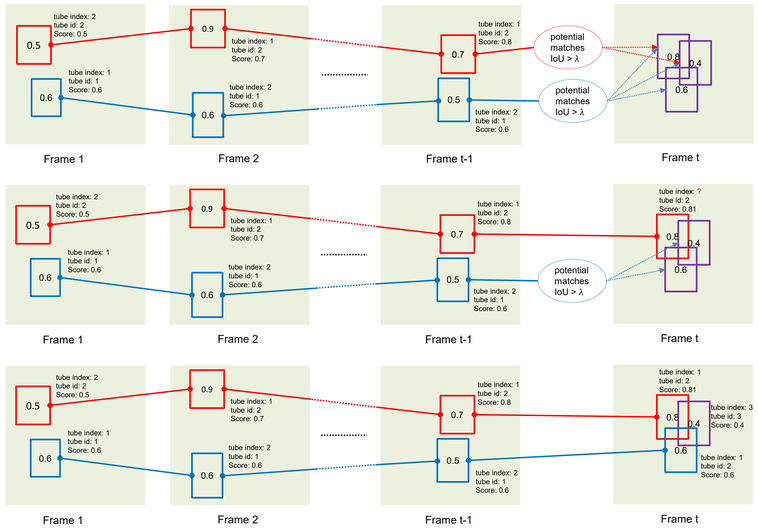


- We provide the option of using two different type of optical flows as shown in above figure in block (b) and (c).
- Two SSD [7] networks (e) are used for appearance & optical flow.
- We provide two options for fusing two sets of output detections (f) from appearance and flow networks (g).
  - i) union-set: take union of the two sets of detections .
  - ii) boost-fusion: boost the box score based on the IoU with other set of boxes.
- Tube generation (h) is performed for each class *c* in parallel and independently, frame by frame starting from first frame.

Step-1: Initialise the tubes using top *n* detections for class *c*.

Step-2: At any time *t* sort the tubes based on their score.

Step-3: For each tube find potential matches based on the IoU and the box scores for class *c* as illustrated below:



Step-4: Update temporal labelling of each tube using the new added box, by using a temporal label changing cost as shown by [8].

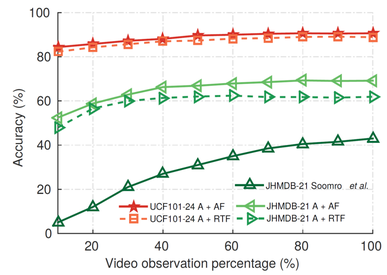
Step-5: Terminate tubes based on the labeling or if no match is found for the tube in the past *k* frames.

Step-6: Initialise new tubes using unassigned detections.

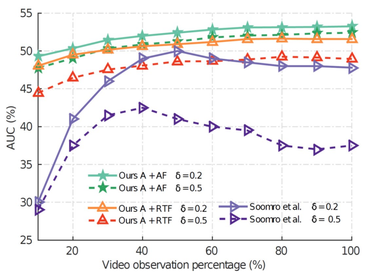
- We used n = 10, $\lambda = 0.1$ and $k = 5$ in all of our experiments

Results

Early Label Prediction



Online localisation-JHMDB21



Spatiotemporal localisation results on UCF101*

	Method \\ threshold δ	0.2	0.5	0.75	0.5:0.95
Faster RCNN	Peng et al. [5]	**73.5**	32.1	02.7	07.3
	Saha et al. [6]	66.6	36.4	07.9	14.4
SSD	Appearance (A) (Fastest)	69.8	40.9	**15.5**	18.7
	Real-time flow (RTF)	42.5	13.9	00.5	03.3
	Accurate Flow (AF)	63.7	30.8	02.8	11.0
	A + RTF(union-set) (Real-time & SOTA)	70.2	43.0	14.5	19.2
	A + AF (boost-fusion)	73.0	44.0	14.1	19.2
	A + AF (union-set) (Our best)	**73.5**	**46.3**	15.0	**20.4**
	A + AF (union-set) Saha et. al.[6]	71.7	43.3	13.2	18.6

* All the result are shown on revised spatiotemporal annotations, corrected manually by authors

Real-time analysis

Modules \\ Setup	A	A+RTF	A+AF	[6]
Flow computation time (ms)	--	07.0	110	110
Detection network time (ms)	21.8	21.8	21.8	145
Tube generation time (ms)	02.5	03.0	03.0	10.0
Overall speed (fps)	40.0	28.0	07.0	03.3

References

[1] G. Gkioxari, and J. Malik, Finding action tubes, CVPR, 2015.
[2] X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection", ECCV, 2016.
[3] S. Saha, et al, Deep learning for detecting multiple space-time action tubes in videos, BMVC 2016.
[4] K. Soomro, et al, Predicting the where and what of actors and actions through online action localization, CVPR, 2016.
[5] T. Brox, et al, High accuracy optical flow estimation based on a theory for warping, ECCV, 2004.
[6] T. Kroeger, et al, Fast optical flow using dense inverse search, ECCV, 2016.
[7] W. Liu, et al, SSD: Single shot multibox detector, ECCV, 2016.
[8] G. Evangelidis et al, Continuous gesture recognition from articulated poses", ECCVW, 2014.

Paper

Code

Code: https://github.com/gurkirt/realtime-action-detection

Paper: https://arxiv.org/pdf/1611.08563.pdf

Email: Gurkirt.Singh-2015@brookes.ac.uk

Revised annotations: https://github.com/gurkirt/corrected-UCF101-Annots