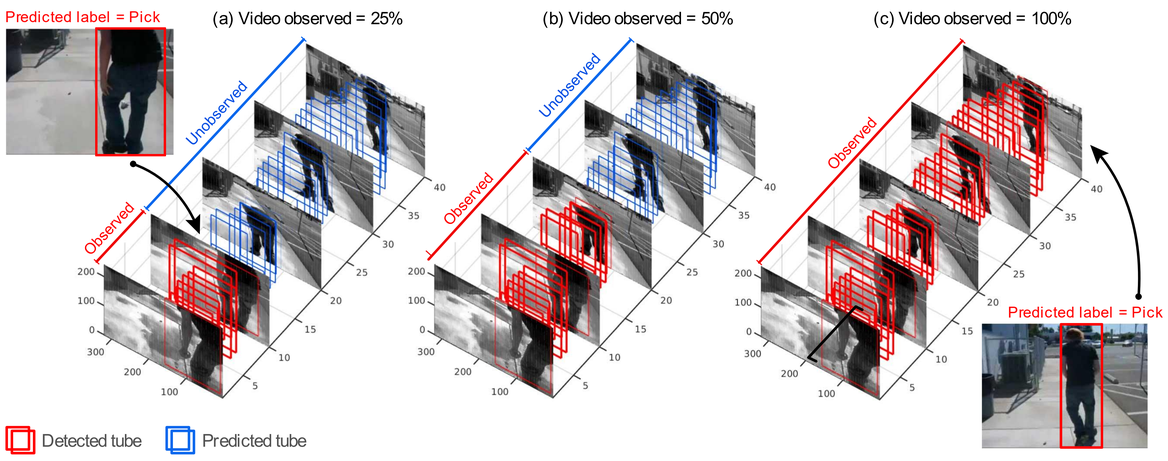
Predicting Action Tubes

Gurkirt Singh Suman Saha Fabio Cuzzolin
Visual Artificial Intelligence Laboratory (VAIL), Oxford Brookes University

European Commission | Horizon 2020 European Union funding for Research & Innovation

ECCV 2018

Online action localisation and future location prediction



The task is to determine/predict what action is occurring in a video, as early as possible using the observed part of the video, localise it (in red above), and predict its future locations (in blue above).

- **Action Tube Localisation:** a set of linked bounding boxes covering each individual action instance.
- **Online:** the action tube should be constructed incrementally.
- **Label Prediction:** to predict the label of the action tube at any given point in time.
- **Location Prediction:** to predict the future locations of the action tube at any given point in time.

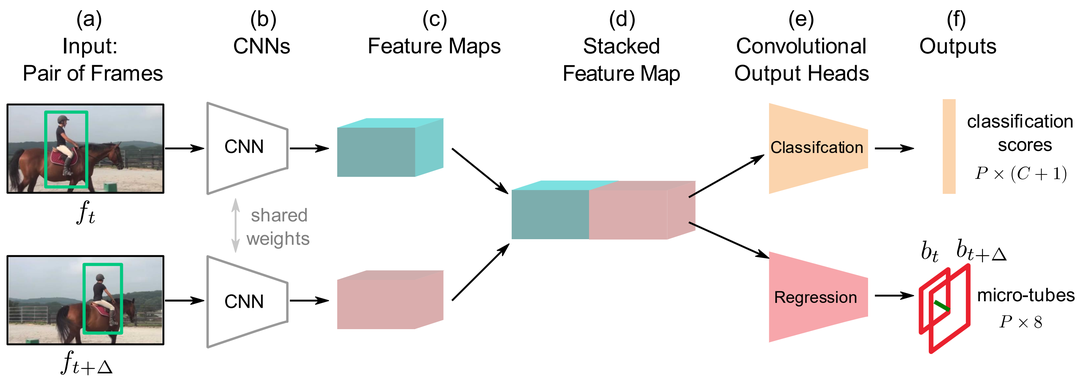
Why?

- Real-time online action localisation and future prediction are essential for many applications, e.g. surveillance, human-robot interaction, autonomous driving, robotic surgery etc.
- Future location prediction is essential to design reactive system, e.g. autonomous driving, robotic surgery.

Contributions

- Unlike other action label prediction [1,2], or trajectory prediction methods [3,4], for the **first** time, we solve the action prediction and future location prediction problem **simultaneously and incrementally;**
- Training a network to make predictions also helps improv action detection performance;
- We demonstrate that the feature-based fusion of optical flow [5] based feature with appearance based features works better than late fusion in the context of action detection.

Action Micro-tubes and Action Detection



- Action micro-tube detection based on two frames separated by Δ frames, Saha *et al.* [7].
- The Linking of micro-tubes to create whole action-tube is based on Singh *et al* [6]

Results: Action detection on JHMDB-21

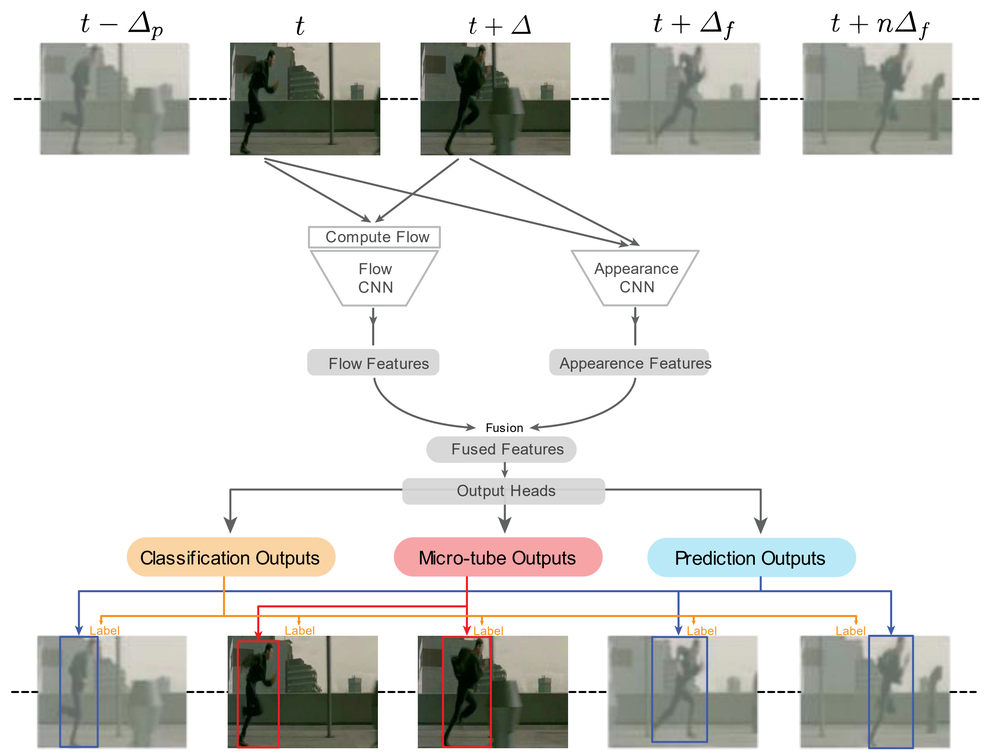
Method \\ threshold δ	0.2	0.5	0.75	0.5:0.95	Accuracy %
Online-SSD Singh *et al.* [6]	73.8	72.0	44.5	41.6	--
AMTnet Saha *et al.* [7] rgb-only	57.7	55.3	--	--	--
ACT Kalegoton *et al.* [8]	**74.2**	73..7	**52.1**	44.8	61.7
T-CNN Hou *et al.* [9]	78.4	**76.9**	--	--	67.2
AMTnet-LateFusion	71.7	71.2	49.7	42.5	65.8
AMTnet-FeatFusion-Concat	73.1	72.6	59.8	48.3	68.4
AMTnet-FeatFusion-Sum	73.5	72.8	59.7	48.1	69.6
Ours TPnet – 053	72.6	71.2	58.0	46.7	67.5
Ours TPnet – 453	73.8	73.0	59.1	47.3	68.4
Ours TPnet – 051	74.6	73.1	60.5	49.0	**69.8**
Ours TPnet – 451	74.8	74.1	**61.3**	**49.1**	68.9

TPnet - abc represents our TPnet, where a = Δp, b = Δf and c = n.

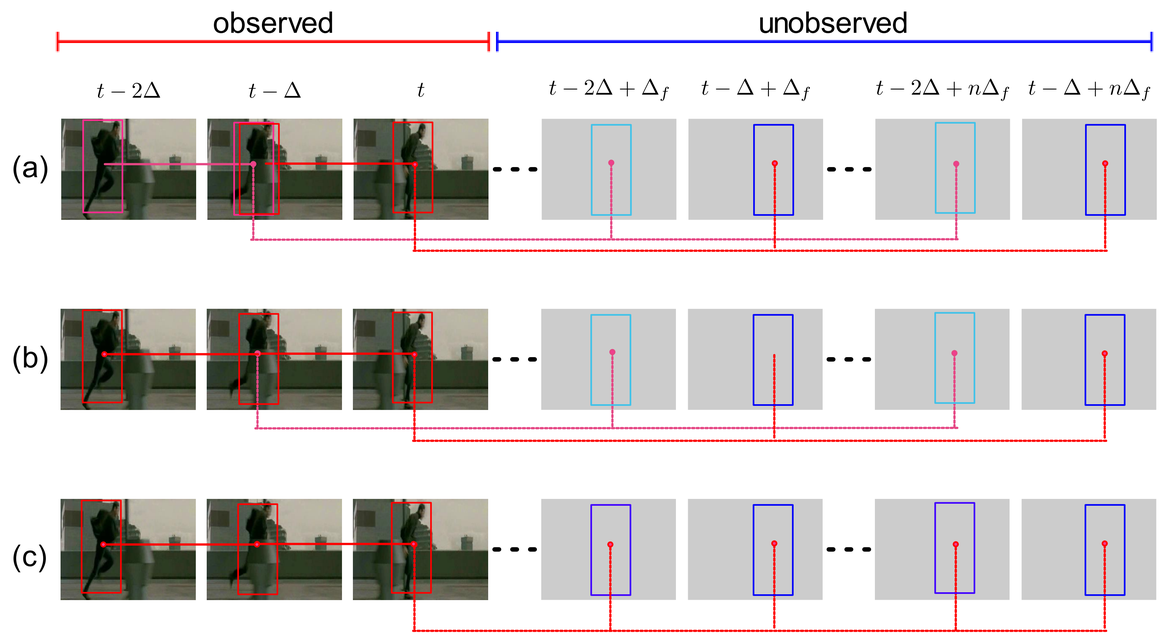
References

[1] Y. Kong et al., Deep sequential context networks for action prediction CVPR 2016. **[2]** M.S. Ryoo Humanactivityprediction:Earlyrecognitionofongoingactivitiesfromstreaming videos, ICCV 2011. **[3]** K.M. Kitani, Activity forecasting, ECCV 2012. **[4]** A. Alahi, Social lstm: Human trajectory prediction in crowded spaces, CVPR 2016. **[5]** T. Brox, et al, High accuracy optical flow estimation based on a theory for warping, ECCV, 2004. **[6]** G. Singh, et al, Online Real-time Multiple Spatiotemporal Action Localisation and Prediction, ICCV 2017. **[7]** S. Saha et al. AMTnet : Action-micro-tube regression by end-to-end trainable deep architecture, ICCV 2017. **[8]** V. Kalogeiton et al., Action tubelet detector for spatiotemporal action localization ICCV2017 **[9]** R. Hou,Tube convolutional neural network (t-cnn) for action detection in videos, ICCV 2017.

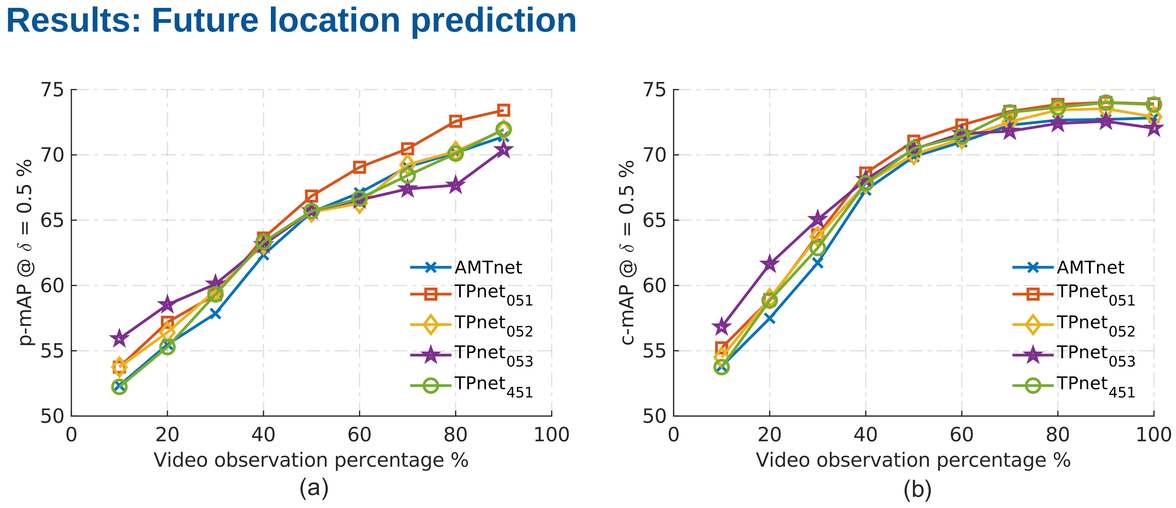
Predicting future and past locations along with micro-tubes



Linking micro-tubes to predicting future of action tube



Results: Future location prediction



$TPnet_{abc}$ represents our TPnet, where a = Δp, b = Δf and c = n.

Results: Action label prediction and online action detection