Deep Learning for Detecting Multiple Space-Time Action Tubes in Videos

Suman Saha[1]
suman.saha-2014@brookes.ac.uk
Gurkirt Singh[1]
gurkirt.singh-2015@brookes.ac.uk
Michael Sapienza[2]
michael.sapienza@eng.ox.ac.uk
Philip H. S. Torr[2]
philip.torr@eng.ox.ac.uk
Fabio Cuzzolin[1]
fabio.cuzzolin@brookes.ac.uk

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

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

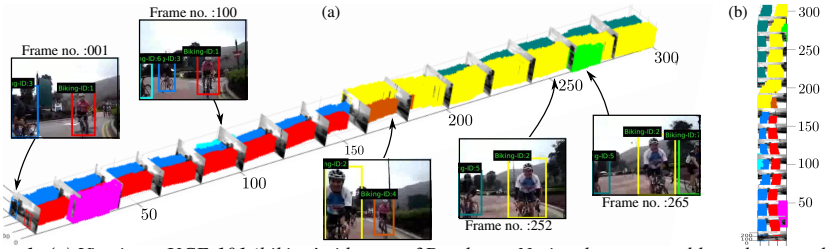

Figure 1: ***(a)*** *Viewing a UCF-101 'biking' video as a 3D volume. Notice that we are able to detect multiple action instances in both space and time.* ***(b)*** *Top-down view.*

In this work, we propose an approach to the spatiotemporal localisation (detection) and classification of multiple concurrent actions within temporally untrimmed videos. Our framework is composed of three stages. In stage 1, appearance and motion detection networks are employed to localise actions from colour images and optical flow. In stage 2, the appearance network detections are boosted by combining them with the motion detection scores, in proportion to their respective spatial overlap. In stage 3, sequences of detection boxes most likely to be associated with a single action instance, called action tubes, are constructed by solving two energy maximisation problems via dynamic programming. While in the first pass, action paths spanning the whole video are built by linking detection boxes over time using their class-specific scores and their spatial overlap, in the second pass, temporal trimming is performed by ensuring label consistency for all constituting detection boxes.

We demonstrate the performance of our algorithm on the challenging UCF101, J-HMDB-21 and LIRIS-HARL datasets, achieving new state-of-the-art results across the board and significantly increasing detection speed at test time. We achieve a huge leap forward in action detection performance when compared to the top competitor [2], and report a **20%** and **11%** gain in mAP on UCF-101 and J-HMDB-21 datasets respectively. The proposed $appearance + motion$ fusion strategy improves the mAPs by 9.4%, 3.6% and 2.5% on the UCF-101, J-HMDB-21 and LIRIS HARL datasets respectively. Further, our 2-pass energy maximisation algorithm contributes to a great extent to significantly boost the performance. Finally, we demonstrate that our action detection pipeline is relatively faster in training and test time detection speeds than the state-of-the-art [1, 2]. Sample qualitative results are provided in the supplementary video [1], and on the project web page [2], where the code and the pretrained models have also been made available.

[1] G Gkioxari and J Malik. Finding action tubes. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015.

[2] 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.

[1] https://www.youtube.com/embed/vBZsTgjhWaQ
[2] http://sahasuman.bitbucket.org/bmvc2016