Finding Action Tubes

Georgia Gkioxari, Jitendra Malik
University of California at Berkeley

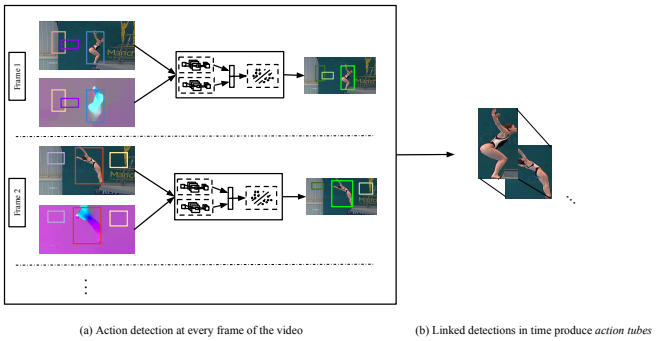

Figure 1: An outline of our approach. (a) Candidate regions are fed into action specific classifiers, which make predictions using static and motion cues. (b) The regions are linked across frames based on the action predictions and their spatial overlap. *Action tubes* are produced for each action and each video.

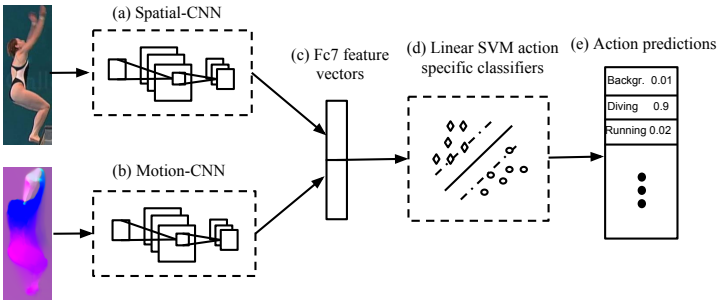

Figure 2: We use action specific SVM classifiers on spatio-temporal features. The features are extracted from the fc7 layer of two CNNs, *spatial-CNN* and *motion-CNN*, which were trained to detect actions using static and motion cues, respectively.

Most work on action recognition in video [4, 6, 7] is aimed at **action classification** *"Name the action being performed in the video"*. Instead, our goal is **action detection** *"Is there an action being performed in the video, and where and when is it happening"*.

Inspired by the recent advances in the field of object detection in images [1], we start by selecting candidate regions and use convolutional networks (CNNs) to classify them. Motion is a valuable cue for action recognition and we utilize it in two ways. We use motion saliency to eliminate regions that are not likely to contain the action. This leads to a big reduction in the number of regions being processed and subsequently in compute time. Additionally, we incorporate kinematic cues to build powerful models for action detection. Given a region, appearance and motion cues are used with the aid of convolutional neural networks to make a prediction. Predictions from all the frames of the video are linked to produce consistent detections in time. We call the linked predictions in time *action tubes*. Figure 1 outlines our approach. Figure 2 shows in detail the design of our action models. We use two CNNs which operate on the image and flow signal respectively. Their feature representations are combined into a spatio-temporal feature vector which is subsequently used to classify the region into an action or background. Our experiments indicate that appearance and motion are complementary sources of information and using both leads to significant improvement in performance.

Action tubes outperform all other approaches ([2, 3, 5, 8]) on UCF sports, with the biggest gain observed for high overlap thresholds. In particular, for an overlap threshold of 0.6 our approach shows a relative improvement of 87.3%, achieving mean AUC of 41.2% compared to 22.0% reported by [8]. Figure 3 shows the average AUC for different values of intersection-over-union threshold. Additionally, we show that action tubes yield improved results on action classification on J-HMDB. Using our action detections we are able to achieve an accuracy of 62.5% on J-HMDB, compared to 56.6% reported by [7] and 56.5% achieved by a whole frame video classification technique with CNNs.

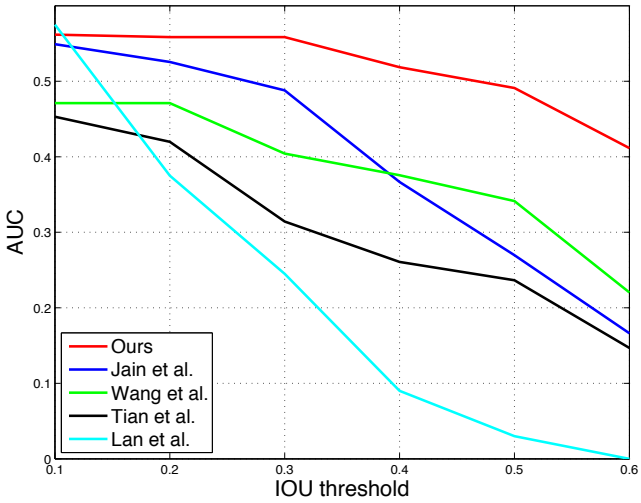

Figure 3: AUC on UCF Sports for various values of intersection-over-union threshold of σ (x-axis). Red shows our approach. We consistently outperform other approaches, with the biggest improvement being achieved at high values of overlap ($\sigma \geq 0.4$).

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] M. Jain, J.v. Gemert, H. Jegou, P. Bouthemy, and C. G. M. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014.

[3] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011.

[4] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[5] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013.

[6] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[7] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[8] L. Wang, Y. Qiao, and X. Tang. Video action detection with relational dynamic-poselets. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.