Finding Action Tubes

Xuewen Yang

June 6 2018

In object recognition, there are two traditional problems: whole image classification,"is there a chair in the image?", and object detection, "is there a chair and where is it in the image?". The two problems have been quantified by the PASCAL Visual Object Challenge [2] and more recently the ImageNet Challenge [1]. The focus has been on the object detection task due to its direct relationship to practical, real world applications. When we turn to the field of action recognition in videos, we find that most work is focused on video classification,is there an action present in the video, with leading approaches [5] [6] [7] trying to classify the video as a whole. In this work, we address the problem of action detection, is there an action and where is it in the video.

Our goal is to build models which can localize and classify actions in video. Figure 1 outlines our approach. Inspired by the recent advances in the field of object detection in images [3], 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. Figure 2 shows the design of our action models. Given a region, appearance and motion cues are used with the aid of convolutional neural networks to make a prediction. Our experiments indicate that appearance and motion are complementary sources of information and using both leads to significant improvement in performance. 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.

Our detection pipeline is inspired by the human vision system and, in particular, the two-streams hypothesis. The ventral pathway (what pathway) in the visual cortex responds to shape, color and texture while the dorsal pathway (where pathway) responds to spatial transformations and movement. We use convolutional neural networks to computationally simulate the two pathways. The first network, spatial-CNN, operates on static cues and captures the appearance of the actor and the environment. The second network, motion-CNN, operates on motion cues and captures patterns of movement of the actor and the object (if any) involved in the action. Both networks are trained to discriminate between the actors and the background as well as between actors performing different actions.

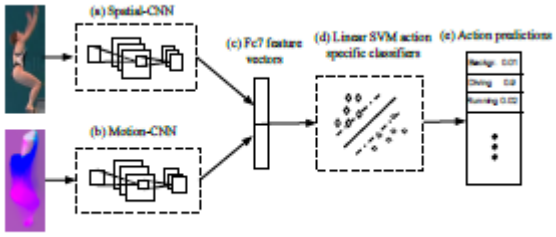

Figure 2. We use action specific SVM classifiers on spatiotemporal features. The features are extracted from the fc7 layer of two CNNs, spatial-CNN and motion-CNN, which were trained to detect actions using static and motion cues, respectively.

We show results on the task of action detection on two publicly available datasets, that contain actions in real world scenarios, UCF Sports [4] and J-HMDB. These are the only datasets suitable for this task, unlike the task of action classification, where more datasets and of bigger size (up to 1M videos) exist. Our approach outperforms all other approaches 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. On the larger J-HMDB, we present an ablation study and show the effect of each component when considered separately. Unfortunately, no other approaches report numbers on this dataset. 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 [6] and 56.5% achieved by a whole frame video classification technique with CNNs.

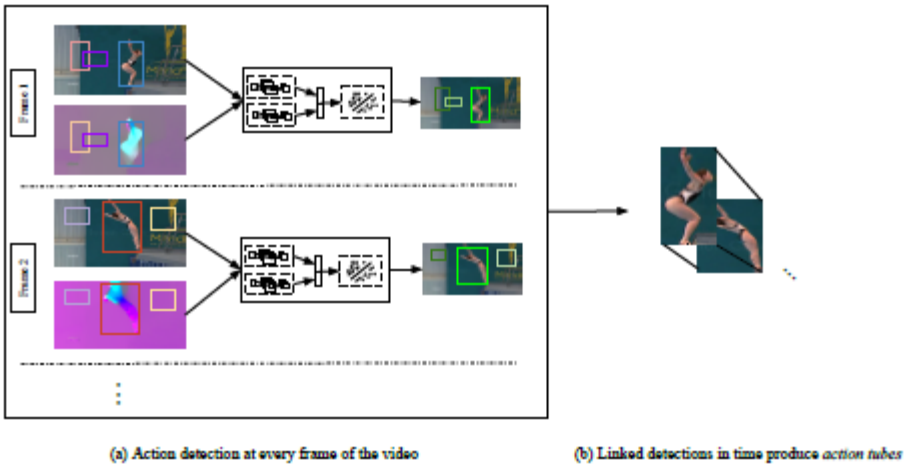

Figure 1. An outline of our approach. (a) Candidate regions are fed into action specific classifiers, which make predictions using static and motion cues. (b) The regions are linked across frames based on the action predictions and their spatial overlap. Action tubes are produced for each action and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

References

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *IJCV*, 2010. 1

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

[4] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 1

[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[6] H. Wang, A. Klaser, C. Schmid, and C. L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 1

[7] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2014. 1