Fast Action Proposals for Human Action Detection and Search

Gang Yu, Junsong Yuan
Nanyang Technological University, School of EEE, Singapore.

Previously, action proposals [1, 2] are generated based on video segmentation [3] which itself is a challenging problem. Moreover, it is difficult to efficiently and accurately segment the human action from the clutter video sequences. In this paper, we present an approach to efficiently propose action candidates of generic type in unconstrained videos without video segmentation. Each proposed action candidate corresponds to a temporal series of spatial bounding boxes, i.e., a spatio-temporal video tube, which locates the potential action in the video.

For many video analytics tasks, e.g., action detection [4] and action search [5], we argue that a quick generation of action proposals is of great importance, because sophisticated action recognition can focus on the action proposals rather than the whole video to save computational cost and improve the performance, similar to the benefits of using object proposals for object detection and recognition.

Despite the success of object proposals, generating action proposals in videos is however a more challenging problem due to two reasons. First, different from objectness measure that relies on visual appearance only, action proposals need to take both appearance and motion cues into consideration. For example, actions should be coupled with human with meaningful motion. However, due to the diversity and variations of human actions, it is difficult to learn the actionness measure that can well differentiate human actions from the background clutters and other dynamic motions, which are quite common in unconstrained videos. Second, the candidate number of action proposals can be much larger than that of the object proposals. Given a video of size $M \times N \times T$, even with the fixed size bounding box, the candidate number of action proposals can be as large as $O(MNTk^T)$ [6], where k is the number of spatial neighbors a bounding box will consider to link in the next frame, which controls the smoothness of the action proposal tube. As the spatial extent of the action can vary across frames, if we consider a flexible bounding box size, it becomes an even much larger size of $O(M^2N^2Tk^T)$. As a result, it is computationally infeasible to explore the full candidate set to pick action proposals.

To address the above two challenges when generating the action proposals, denoted as $\mathbf{P}$, we formulate the problem based on the maximum set coverage problem [7]. Each action candidate (or path) $\mathbf{p}^{(i)}$ can be considered as a set with the bounding box $b^{(i)}$ as its element, i.e., $\mathbf{p}^{(i)} = \{b_{t_s}^{(i)}, b_{t_s+1}^{(i)}, \cdots, b_{t_e}^{(i)}\}$ where t_s and t_e are the start frame and end frame of the path, respectively. We want to maximize the following function:

$$\max_{\mathbf{P} \subset \mathcal{S}} \sum_{b_t \in \cup \mathbf{p}^{(i)}} w(b_t) \tag{1}$$

$$\text{s.t.} \quad |\mathbf{P}| \leq K, \tag{2}$$

$$\mathbf{O}(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) \leq \delta_p, \ \ \forall \mathbf{p}^{(i)}, \mathbf{p}^{(j)} \in \mathbf{P}, i \neq j. \tag{3}$$

The first constraint (Eq. 2) is to set the maximum number of action proposals as K while the second constraint (Eq. 3) is to avoid generating redundant action proposals that are highly overlapped. More specifically, we first perform human and motion detection to generate candidate bounding boxes (b_t) that may cover the human action in each frame. After picking up the bounding boxes of high "actionness" scores $w(b_t)$, we utilize the max sub-path search algorithm to locate the top-N maximal spatio-temporal paths based on "actionness" score. Due to the spatio-temporal redundancy in the video, many high quality paths may largely overlap with each other as the example shown in Fig. 1. The red paths illustrate three detected action proposals. But the green paths and blue path which significantly overlap with path 1 are redundant paths and should be removed. To pick the action proposals, a greedy based search algorithm is performed to select a set of action proposals $\mathbf{P}$ that can maximize the overall actionness score in Eq. 1.

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

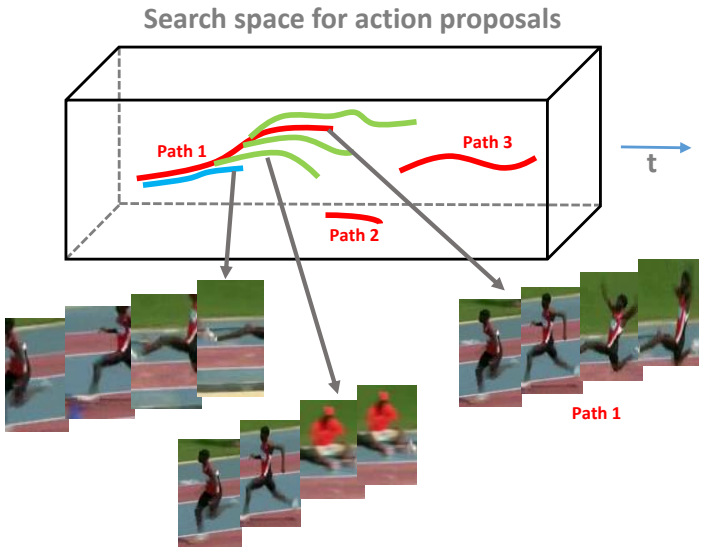

Figure 1: An illustration of action proposals. The red paths in the upper figure represent three detected action proposals, where each action proposal corresponds to a series of bounding boxes in the video space. The green and blue paths, which have large spatial-temporal overlap with the red paths, should be removed for the path diversity.

To evaluate the performance of our action proposals, we test two benchmark datasets, MSR II and UCF 101. We notice that a small number of action proposals, e.g., 2000 proposals for all the 54 video clips in MSRI-I dataset, can already provide promising recall rate. Also, based on our action proposals, we can obtain state-of-the-art action detection and action search results in MSRII dataset compared with existing results. Moreover, the competitive result on UCF 101 dataset validates that our action proposal can well track the actions in unconstrained videos. Last but not the least, compared with existing action proposal approaches, our action proposals do not rely on video segmentation and can be generated in nearly real-time on a normal desktop PC.

Acknowledgement

This work is supported in part by Singapore Ministry of Education Tier-1 Grant M4011272.040.

[1] M. Jain, J. Gemert, H. Jegou, P. Bouthemy, C. Snoek, "Action localization by tubelets from motion," in *CVPR*, 2014.

[2] D. Oneata, J. Revaud, J. Verbeek, C. Schmid, "Spatio-Temporal Object Detection Proposals," *ECCV*, 2014.

[3] C. Xu, J.J. Corso, "Evaluation of super-voxel methods for early video processing," *CVPR*, 2012.

[4] J. Yuan, Z. , Y. Wu, "Discriminative Video Pattern Search for Efficient Action Detection," in *TPAMI*, pp. 1728 - 1743, Vol. 33, 2011.

[5] G. Yu, J. Yuan, Z. Liu, "Unsupervised Random Forest Indexing for Fast Action Search," in *CVPR*, 2011.

[6] D. Tran, J. Yuan, D. Forsyth, "Video Event Detection: from Subvolume Localization to Spatio-Temporal Path Search," *PAMI*, 2014.

[7] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, "An analysis of approximations for maximizing submodular set functions I," *Mathematical Programming*, Vol. 14, pp. 265-294, 1978.