A proposal-based solution to spatio-temporal action detection in untrimmed videos

Ketul Shah[1], Joshua Gleason[2], Rama Chellappa[1]
[1]Johns Hopkins University [2]University of Maryland, College Park

Here we briefly describe our overall approach used in our Submission ID 27264 on the ActEV SRL Leaderboard [1]. Our method uses the MEVA dataset [2] for training. There are two main differences of our current approach from previous submissions (based on [3], [4]): (1) Proposal generation and (2) Post-Processing. We use a proposal generation network at inference time and perform filtering of our predictions as detailed below. Our overall pipeline is shown in Fig 1.

1 Method

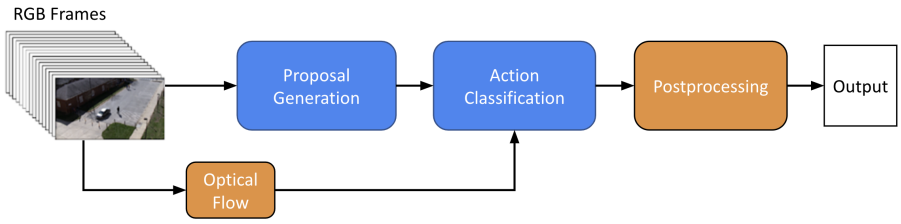

Figure 1: Overall pipeline.

Proposal Generation: Here we briefly describe our proposal generation method shown in Fig 2. At test time, we use data-driven proposals obtained using our proposal generation network. The design of this proposal generation network is inspired by recent progress in 3D semantic segmentation [5]. The input to this network are uncropped strided frames, and it is trained to predict whether each voxel in the 3D XYT volume is part of an activity or not. We use a U-Net like architecture and is trained using a combination of the Tversky [6] and BCE loss. The final proposals are obtained from this binary occupancy volume by finding the connected components and converting them to axis-aligned cuboids. For training proposals, we first perform object detections on the video frames, and perform hierarchical clustering with each detection represented as (x, y, f), where x, y is the center of the bounding box and f is the frame number. We use this method of proposal generation for training due to more uniform representation of data and better selection of negatives compared to the data-driven method.

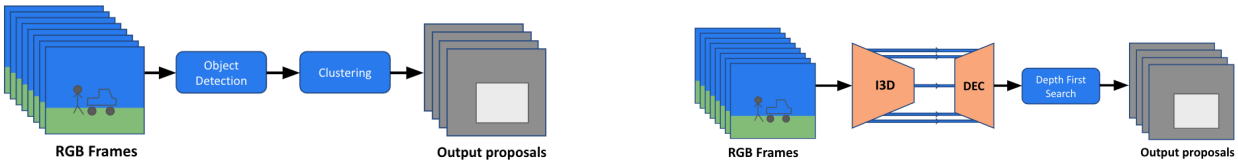

Figure 2: **(Left)** Proposal generation method at training time. **(Right)** Proposal generation method at inference time.

Action Classification: Our action classification model uses optical flow input. We use an Inception-v1 I3D [7] as the backbone, with 32-frame clips from the proposals as input to our model. For this multi-label classification task, we employ the BCE loss for training. Please refer to [3], [4] for more details.

Post-processing: At inference, we perform object-conditional and camera-conditional filtering. We filter activities during post-processing based on the location of the camera (indoor vs. outdoor). For indoor cameras, we suppress all vehicle activities. The can be some cases such as indoor parking lots where such filtering can fail. To determine where the camera is located, we directly make use of the camera meta-data provided for videos. We also filter our predictions in the post-processing step based on consistency with object detections. All activities can be divided into person-only, vehicle-only or person-vehicle activities. For each input cuboid proposal, we have objects detected in that cuboid, and the predicted activity for that cuboid. We perform object-conditional filtering to ensure the following: (1) All person-only activities have at least one person detected (2) All vehicle-only activities have at least one vehicle detected and (3) All person-vehicle activities have at least one person and one vehicle detected. The final predictions to evaluate are obtained after applying these two filtering steps.

2 Results

Below are the results of our submission 27264 from the SRL Leaderboard.

AD Task: Our approach results in AD mean p-miss@0.1rfa of 0.7789, and AD mean nAUDC@0.2rfa of 0.7995.

AOD Task: Our approach results in AOD mean p-miss@0.1rfa of 0.8131, and AOD mean nMODE@0.1rfa of 0.1620.

References

[1] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, , and G. Quénot, "An overview on the evaluated video retrieval tasks at trecvid 2022," in *Proceedings of TRECVID 2022*. NIST, USA, 2022.

[2] "The multiview extended video with activities (meva) dataset," https://mevadata.org/.

[3] J. Gleason, R. Ranjan, S. Schwarcz, C. Castillo, J.-C. Chen, and R. Chellappa, "A proposal-based solution to spatio-temporal action detection in untrimmed videos," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 141–150.

[4] J. Gleason, C. D. Castillo, and R. Chellappa, "Real-time detection of activities in untrimmed videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2020, pp. 117–125.

[5] Y. He, H. Yu, X. Liu, Z. Yang, W. Sun, Y. Wang, Q. Fu, Y. Zou, and A. Mian, "Deep learning based 3d segmentation: A survey," *arXiv preprint arXiv:2103.05423*, 2021.

[6] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International workshop on machine learning in medical imaging*. Springer, 2017, pp. 379–387.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.