BUPT-MCPRL at TRECVID 2022 ActEV SRL Challenge

Hangyue Zhao[1],Zhihang Tong[1], Yuchao Xiao[1], Shuhao Qian[1], Song Li[1], Zihan tian[1], Yanyun Zhao[1,2], Fei Su[1,2], Zhicheng Zhao[1,2]
[1]Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture, China
{zhaohy21315, tongzh, ycxiao, qiansh, ls0577,Tianzh, zyy,sufei,zhaozc}@bupt.edu.cn

Abstract

1. Training data: MEVA, COCO (pretraining), Kinetics400 (pretraining).
2. Approach: (27305: 3D detectors, 5 different classification methods)
3. Difference: None.
4. Contribution: Our targeted 5 classification methods and targeted strategy greatly benefit the performance, reaching state-of-the-art on the Self-Reported Leaderboard (SRL) Challenge on the MEVA dataset. The PMiss@0.1rfa indicator is as high as 0.6309, far exceeding the second-place result.
5. Conclusion: Surveillance video scene activities recognition is complex, for different activities, specific classification methods should be specifically analyzed and designed.

1. Method

The MEVA dataset contains a total of 20 categories of activities, which we roughly divide into 5 categories (or activity groups) and process them separately:

1)vehicle-only: vehicle starts, vehicle stops, vehicle turns left, vehicle turns right;

2) person-object: person picks up, person puts down, person sits down, person stands up, person transfers object;

3) person-specific object: person interacts with laptop, person reads document, person texts on phone;

4)person-vehicle: person exits vehicle, person enters vehicle, person opens vehicle door, person closes vehicle door;

5) scene-related and person-person: person opens facility door, person enters scene through structure, person exits scene through structure, person talks to person.

Based on this, we propose a comprehensive surveillance video activity detection framework as shown in Figure 1. For each activity group, the video is split into clips, sent to the 3D detector, and then connected into trajectories. These trajectories are fed into the specially trained classifiers, and the results are obtained through different post-processing algorithms. Concretely, we develop different methods to detect and recognize the activities for the different activity groups respectively.

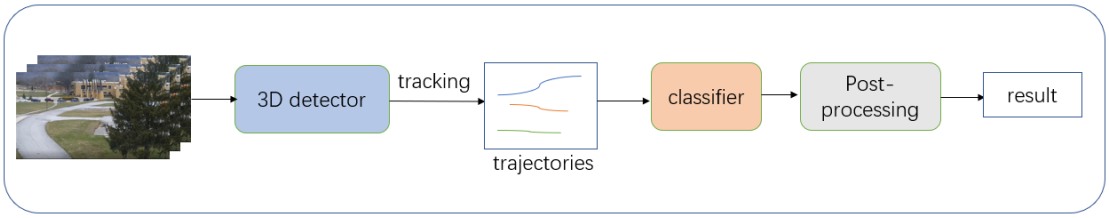

Figure 1. The framework of our activity detection.

An overview of our detection method for vehicle-only activities is shown in Figure 2. We adopt

Cascade R(2+1)D[9] as 3D detector of activity proposals and associate these proposals to form activity trajectories with the IoU tracker. And then, we generate dense temporal domain anchors for activity classification models from these trajectories. Here we adopt Swin Transformer [6] as 3D classifier for vehicle-only activity group. We train two models for strat/stop and left/right respectively. Considering that no activity is labelled in the ground truth of MEVA training dataset when reverse occurs, we add a reverse classification model to eliminate the reverse segment of activity trajectories before start/stop and left/right activity classification. In post processing, we sum up the scores of classification which are obtained through dense temporal domain anchors classification and re-determine the activity boundary based on the fusion score of each frame. In this way, we obtain the vehicle-only activity detection results.

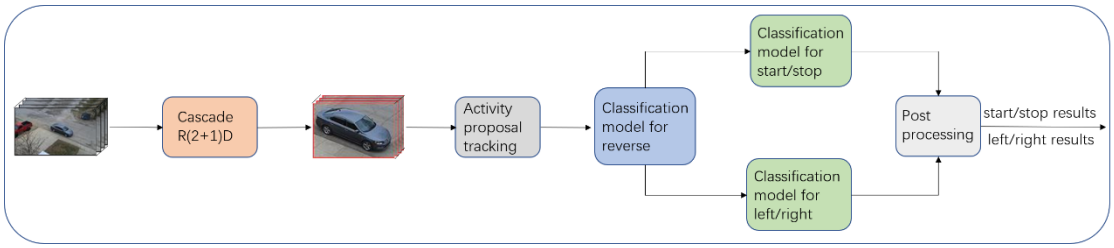

Figure 2. The overview of vehicle-only activity detection.

For person-object activity group detection, we first detect the activity proposals in time-space domain with Cascade RCNN 3D [1] detector from the video clips, then link the activity proposals according to their IoUs to form an activity trajectory. After that, the activity trajectory is transferred into the classifier trained for this activity group to be classified. Finally, we get the final activity tracks for each class with the t-NMS post processing. Note that, different from [5], which only uses one classifier, we exploit a classifier score merge strategy as presented in Figure 3, which could synthesize the results of different classifiers to obtain a more representative score results. Specifically, we use Video Swin Transformer [6] and ActionCLIP [7] as the classifiers, and merge the scores for each category by weighted average. And for post-processing, we use different parameters for each category to get better activity detection performance. In order to improve the performance of activity region detection, we combined 4 kinds of activity samples of person-object group and 3 kinds of activity samples of person-special object group to train activity proposal detector.

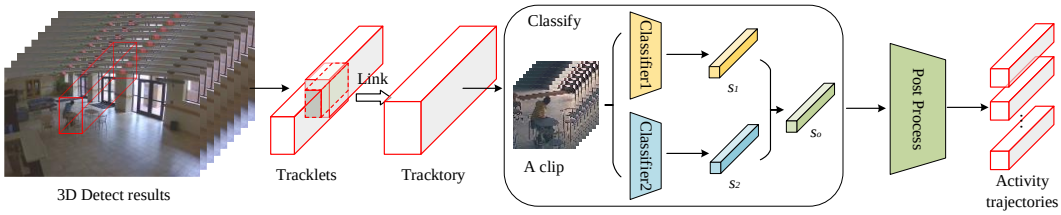

Figure 3. The overview of person-object activity detection.

For person-specific object activity group, we use the same activity proposal detector as person-object group. And we think that this group activities can be recognized by the posture of actors and the objects they interact with in single-frame, and temporal modeling is not necessary. Since CLIP[8] has the ability to achieve zero-shot, we adopt this model as the classifier of this activity group. We crop the activity proposals detected by Cascade RCNN 3D detector from the video clip and link them to from an activity trajectory by our IoU based tracker, and input into the image encoder for frame encoding. And the text prompts we use are shown in Figure 4. In addition to the seven

activities detected by our detector, we also add a background class text in the text prompts so as to recognize and eliminate the false activity region detection of our detector. Text features are obtained by text encoder. Each image of an activity trajectory is classified according to the similarity between the image features and text features. For each activity track, we determine its category and score with voting algorithm.

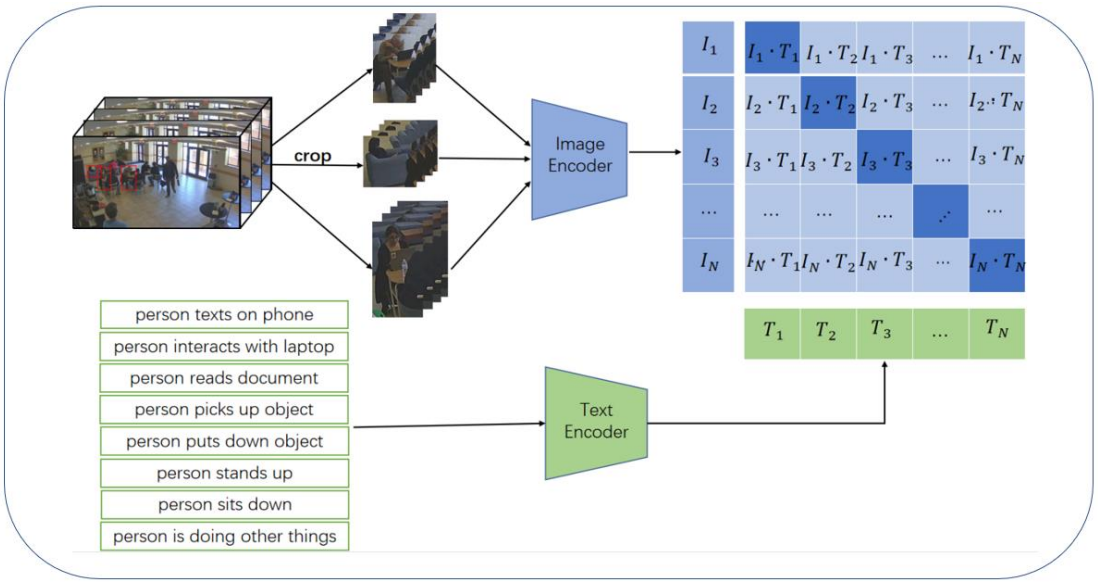

Figure 4. Person-specific object classifier.

For person-vehicle activity group, as shown in Figure 5, the activity proposal clips are extracted by the Cascade RCNN 3D detector and classified by the MViT network [10], and both network models are trained on MEVA training dataset. We get scores for four classes with our classifier. We constrain the results of the two categories of people opening and closing doors with the predicted results of people entering and leaving the vehicle in post-processing. For a trajectory, when the prediction of entering and exiting the vehicle is greater than the threshold we set, we improve the predicted probability of the open vehicle door in the first half of the trajectory, and improve the predicted probability of the close vehicle door in the second half of the trajectory.

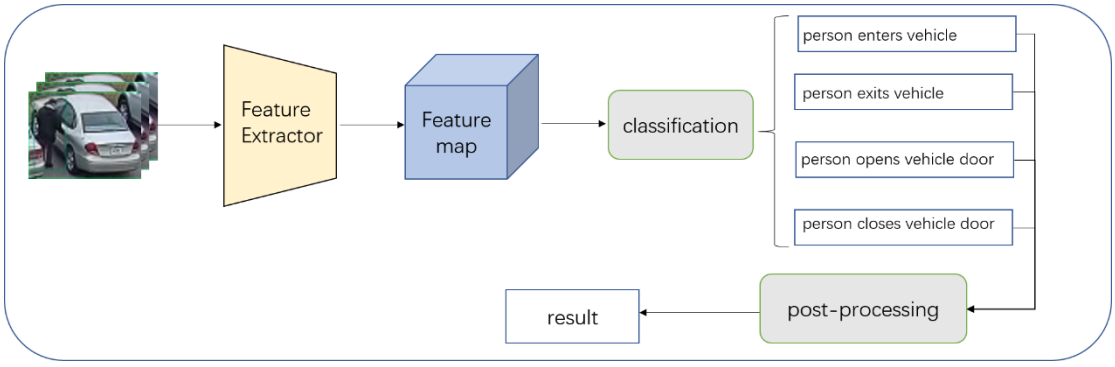

Figure 5. Person-vehicle classifier.

For scene-related activity group, we locate activity proposals with Cascade RCNN 3D detector and track these proposals as mentioned above. As Figure 6 shows, we firstly extend the bounding box of activity proposal clips and use different frame sampling rate to obtain activity proposal. Secondly, we extract features with Swin-tiny model from extended and sampled clips to classify and get classification scores. Finally, these scores are weighted averaged for each class. For "person_talks_to_person", as the bottom of Figure 6 shows, we just extract features with Swin-tiny

model from clips and get classification score.

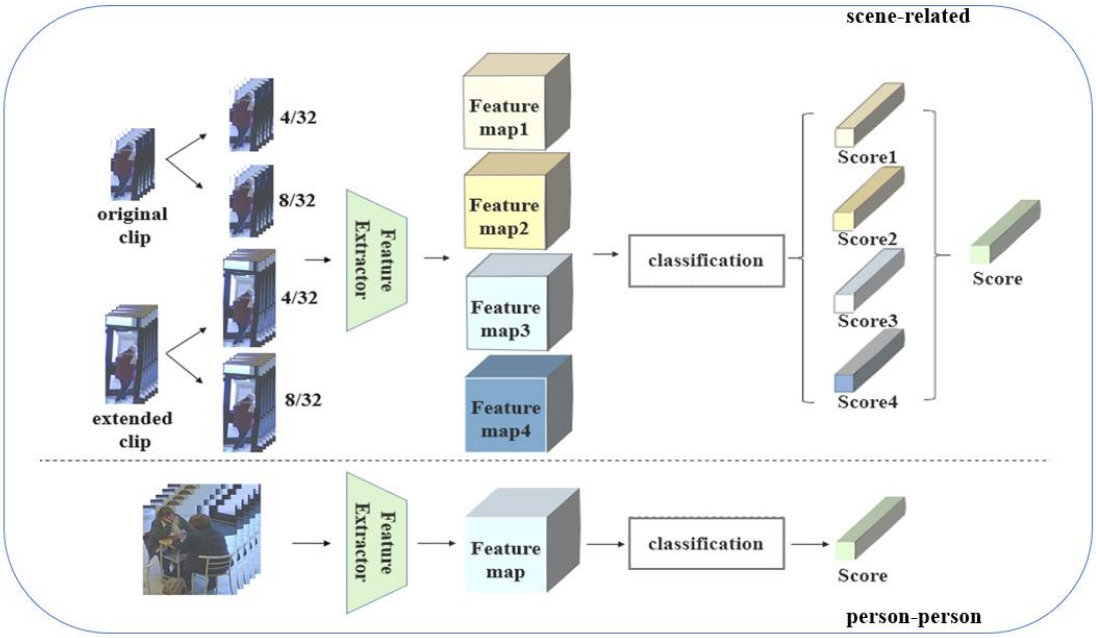

Figure 6. Scene-related and person-person classifier.

2. Results

The ActEV SRL challenge based on the MEVA dataset has strict requirements on the time and space of the system output, so we use different detectors and classifiers to meet the needs of different categories. Our system achieves PMiss@0.1rfa=0.6309 and won the first place on the MEVA dataset, which proves the effectiveness of our method.

Table 1. Results in TRECVID 2022 ActEV Self-Reported Leaderboard (SRL) Challenge

Team	PMiss
BUPT-MCPRL	0.6309
UMD	0.8131
mlvc_hdc	0.9921
WasedaMeiseiSoftbank	0.9961

Reference

[1]Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.

[2] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE international conference on image processing (ICIP). IEEE, 2017: 3645-3649.

[3] Zivkovic Z, Van Der Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction[J]. Pattern recognition letters, 2006, 27(7): 773-780.

[4] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers[J]. arXiv preprint arXiv:2104.11227, 2021.

[5] Yunhao Du1 , Junfeng Wan1 , Binyu Zhang, et al. BUPT-MCPRL at TRECVID 2021 ActEV: 215AD[C]. TRECVID2021.
[6] Liu Z, Ning J, Cao Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3202-3211.
[7] Wang M, Xing J, Liu Y. Actionclip: A new paradigm for video action recognition[J]. arXiv preprint arXiv:2109.08472, 2021..
[8] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. arXiv preprint arXiv:2103.15049.
[9] Tran, Du, et al. A closer look at spatiotemporal convolutions for action recognition. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.
[10] Fan, Haoqi, et al. Multiscale vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.