MCPRL-CMCC at TRECVID 2020: ActEV*

Xiyu Zhao[1], Qun Cao[1],Xiangqian Cheng[1], Junfeng Wan[1],Yunhao Du[1],
Binyu Zhang[1], Yanyun Zhao[1], Zhicheng Zhao[1], Fei Su[1], Hong Chen[2]

[1]Multimedia Communication and Pattern Recognition Labs,
Beijing Key Laboratory of Network System and Network Culture,
Beijing University of Posts and Telecommunications, Beijing, China
[2] China Mobile Research Institute, Beijing, China
{zhaozc, zyy, sufei}@bupt.edu.cn

Abstract

In this paper, we describe the BUPT-MCPRL ActEV system and evaluation results for TRECVID 2020[1]. Our proposed system framework for ActEV 2020 consists of three stages: spatial activity location, activity tracking, temporal activity detection and classification.

- **p_baseline_4** : the baseline of our method. In this submission, we use 3D-RetinaNet and 3D-Cascade-RCNN with Focal Loss to generate activity tubes, then we use tracking algorithm to link the tubes and build efficient temporal localization systems to detect activities (activity instances).
- **p_baseline_12**: replace TPN with r(2+1)d and adjust the threshold.

The results are shown in Table 1.

Table 1. Improvement on the test set

SYSTEM_NAME	PATIAL AUDC
p_baseline_4	0.625
p_baseline_12	0.55515

1. Introduction

The ActEV task is more challenging than online consumer video challenges task, as its videos of VIRAT dataset are untrimmed video sequences, and there may be multiple activities or no activities in one video. Moreover, the activity of interest only accounts for a small part in the space-time domain in surveillance video, and most of which is background interference. So the activity detection system must firstly locate the specified activities in space-time domains as accurately as possible, and then clip the candidate areas from videos and classify them to get the activity detection results. In ActEV 2020 evaluation , the categories of activities have increased to 35, which are more complex and diverse. And it brings about more difficulties of task research, but also makes more interesting.

In this year evolution, our approach has also inherited the solution of our team at ActEV-PC 2019 challenge[2], that is, we divide the spatiotemporal activity detection task into three modules:

*This work is supported by Chinese National Natural Science Foundation (62076033, U1931202), and MoE-CMCC Artifical Intelligence Project (MCM20190701)

spatial activity localization, activity tracking, and temporal activity detection. And we made improvements to the problems in related modules, including:

(1) For spatial activity location, we detect the candidate areas where activities may occur in a short video clip based on 3D-RetinaNet to improve recall and detect the activity of person_talks_to_person with 3D-Cascade-RCNN to improve accuracy, instead of 3D-Faster-RCNN[2];

(2) For spatial activity location, we raise NMS threshold appropriately and use Soft-NMS[3] instead of NMS;

(3) For activity classification of temporal activity detection module, the input of classifier doesn't take the way of time-domain anchor, but adopts the way of fixed time-domain window instead.

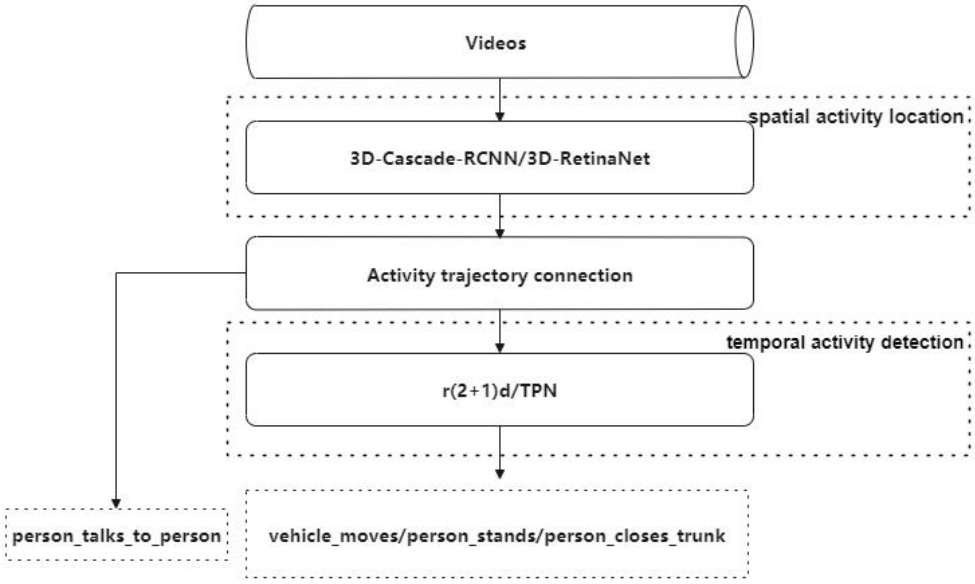

Figure1. The overall structure of our framework

2. Method

Our system consists of three stages: spatial activity localization, activity tracking, and temporal activity detection, which framework is illustrated with Figure 1. According to the characteristics of activities, we divide the activities of ActEV 2020 into four groups: person-centered, person-vehicle, vehicle-only, person_talks_to_person, as shown in Table 1.

Firstly, to detect the first three categories shown in Table 2, we locate the possible activity regions of each clip and get a classification of groups for each region based on 3D-RetinaNet (We replace the backbone of 3D-Faster-RCNN with RetinaNet, so we call it 3D-RetinaNet). And we detect person_talks_to_person activity with 3D-Cascade-RCNN (We replace the backbone of 3D-Faster-RCNN with Cascade-RCNN, so we call it 3D- Cascade-RCNN). Notably we have a total of four groups of detection targets. The input of these detectors includes 8 frames sampling from the sequence of 32 frames.

Next, activity proposals of four categories detected by first stage are associated into activity trajectories temporally with the Hungarian algorithm[4]. Finally, for a trajectory of person_talks_to_person obtained by the above stages, we directly output the trajectory as an activity instance. And r(2+1)d[5] model and TPN[6] model are used respectively to subdivide the rest of the groups. We observed that a target usually carries multiple actions, i.e. one person may carries a bag while walking, so we change single-label action positioning to multi-label action positioning in our method.

Table 2. 35 activities are divided into 4 groups

Type	Events/Activities
Person-centered	person_interacts_object,person_pulls_object,person_pushs_object,person_uses_tool person_rides_bicycle,person_sets_down_object,person_carries_heavy_object, person_carries_object,person_crouches,person_gestures,person_runs, person_sits,person_stands,person_walks,person_talks_on_phone, person_texts_on_phone
Person-Vehicle	person_closes_facility_or_vehicle_door,person_exits_facility_or_vehicle, person_closes_trunk,person_enters_facility_or_vehicle,person_loads_vehicle, person_opens_trunk,person_opens_facility_or_vehicle_door, vehicle_picks_up_person,person_unloads_vehicle
Vehicle-only	vehicle_starts,vehicle_stops,vehicle_turns_left,vehicle_turns_right,vehicle_makes_u_turn,vehicle_moves
Person-Person	person_talks_to_person

3. Spatial Localization

Inspired by the success in 3D-Faster-RCNN, we adapt RetinaNet[7] and Cascade-RCNN[8] into 3D form to utilize spatiotemporal features, which can significantly improve activity spatial localization performance compared with single-frame detector.

To capture the spatiotemporal information in video, we replace the 2D convolution backbone of RetinaNet and Cascade-RCNN with I3D [9]. Focal loss is originally proposed to solve the foreground-background class imbalance in one-stage detector, which makes the detector putting more focuses on hard, misclassified examples. Since Focal loss focused too much on hard samples, now we extend the GHM[10] to multi-class case and use it in the second stage of our 3D-RetinaNet, which is proved to improve the recall a little. Based on observations of the dataset, we raise NMS threshold appropriately and use Soft-NMS instead of NMS. We found that the operation to raise NMS threshold improved the recall a lot and a little improvement for precision. Since speaking events are direct outputs from detector, we use 3D-Cascade-RCNN improves accuracy.

4. Trajectories connection

We extend the task of object box tracking to activity trajectories connection, which can be

formulated as a data association problem. Inspired by[4], the Hungarian algorithm is adopted to link adjacent clips.

5. Activity classification

Given activity trajectories produced by the past step, our next step is to perform activity fine-grained classification. We employ different strategies for each activity group. When we training our classification models, we split the videos into 32-frame or 64-frame clips, then we extract 8 frames as inputs. And when testing, the clip's length is set 32 frames, and inputs also are 8 frames. As we mentioned before, one object may contain one more activities, so for the person-centered group and the vehicle-only group, BCEWithLogitsLoss is chose as classification loss function, and our clips may include one more labels. TPN as one of the best models we know of in action classification performs well in the person-centered group and the vehicle-only group. However, we found the person-vehicle group with more complex activities, was not match TPN, so we employ r(2+1)d model. For r(2+1)d, we found one object only contains an activity, so we choose CrossEntropy Loss as classification loss function. Also for all groups, the problem of class imbalance is seriously, then the class re-balance strategy is employed.

6. Post-processing

During the test, for the trajectories after activity classification, for the person-vehicle group, we assign one label to each frame, and for the person-centered group and the vehicle-only group, assign one or more labels to each frame. However, for background frames, we discard. Then, we divide the trajectories according to the labels, so we get several trajectories. Each trajectory only corresponds to one label, and outputs as one activity instance.

7. Conclusion

In this work, we propose an activity detection framework based on deep learning. The results show that our proposed framework is effective. First, spatial localization focuses on determining where events occur and improving recall rates while ensuring accuracy. Second, trajectories connection focuses on quickly and accurately connecting the detection boxes in the previous stage to provide long time tracks for subsequent classification. Last, activity classification focuses on accurately classifying all the activities on the trajectories based on short clips.

References

[1] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Adrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, Georges Quénot}, TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains, Proceedings of TRECVID 2020.

[2] Ya li, et al. An Effective Detection Framework for Activities in Surveillance Videos

[3] Bodla N , Singh B , Chellappa R , et al. Improving Object Detection With One Line of Code[J]. 2017.

[4] Harold W Kuhn. The hungarian method for the assignment problem. Naval research logistics

quarterly, 2(1-2):83–97, 1955.

[5] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450–6459, 2018

[6] C. Yang, Y. Xu, J. Shi, B. Dai and B. Zhou, "Temporal Pyramid Network for Action Recognition," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 588-597, doi: 10.1109/CVPR42600.2020.00067.

[7] Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017:2999-3007.

[8] Cai Z , Vasconcelos N . Cascade R-CNN: Delving Into High Quality Object Detection[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset..

[10] Li B , Liu Y , Wang X . Gradient Harmonized Single-stage Detector[J]. 2018.