BUPT-MCPRL at TRECVID 2019: ActEV and INS*

Ya Li, Guanyu Chen, Xiangqian Cheng, Chong Chen, Shaoqiang Xu, Xinyu Li, Xuanlu Xiang, Yanyun Zhao, Zhicheng Zhao, Fei Su

> Multimedia Communication and Pattern Recognition Lab Beijing Key Laboratory of Network System and Network Culture Beijing University of Posts and Telecommunications, China

> > {zhaozc, zyy, sufei}@bupt.edu.cn

Abstract

In this paper, we describe BUPT-MCPRL systems and evaluation results for TRECVID 2019 [14]. We join two tasks: activities in extended video and instance search.

Activities in Extended Video (ActEV):

- p_baseline_2: the baseline of our method. In this submission, we use 3D-Faster-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 events (activity instances).
- p_baseline_14: add OTW dataset to train 3D-Faster-RCNN.
- p_baseline_22: split long events.
- p_baseline_23: fuse similar events.

Instance Search (INS):

- F_M_E_E_BUPT_MCPRL_1: retrieve persons and using pose-based action detection
- F_M_E_E_BUPT_MCPRL_2: retrieve persons and detect their actions
- F_P_E_E_BUPT_MCPRL_3: the same method as in F_M_E_E_BUPT_MCPRL_2
- I_M_E_E_BUPT_MCPRL_4: first retrieve target persons and detect their actions with partial correction manually

1. Activities in Extended Video

The ActEV task is more challenging than action detection tasks, as it requires accurate spatial and temporal localization. Our approach has inherited the solution at ActEV challenge, that is, we divide the spatiotemporal activity localization task into three subtasks: spatial activity localization, activity tube tracking, and temporal activity detection. For spatial activity localization, we generalize Faster-RCNN [1] to 3D form (3D-Faster-RCNN) to localize activity regions in a short video clip. For activity tube linking, we extend the object tracking algorithm to track activity tubes. For temporal activity localization, we build efficient temporal localization systems to detect activity instances.

In this year, our improvements mainly come from spatial localization and post-processing. Focal Loss [2] is used to improve our 3D-Faster-RCNN. Besides, we use OTW [3] dataset to extend training data, as we find activities of OTW are similar with ActEV. We find simple post-processing method can help the detection results. The reason may lie in the changed activity detection metrics, that is, the normalized partial Area Under the DET Curve (nAUDC) focus more on recall rather than precision, and the alignment is strictly for short activities and loosely for long activities.

1.1 Framework

According to the attributes of different activities, we divide 18 activities into 3 groups, as illustrated

^{*}This work is supported by Chinese National Natural Science Foundation (U1931202, 61532018).

in Table 1, i.e. vehicle-person, turning, person-centered.

Our system includes three stages, whose overall structure is illustrated in Figure 1. First of all, we localize possible activity regions of each clip. A sliding window is used to slice clips from continuous videos. We sample 10 frames from each clip and put them into 3D-Faster-RCNN to detect activity tubes with rough classification. Notably, we treat activity categories of vehicle-person group as one class, activity categories of turning group as another class when detecting activity regions. Adding seven person-centered activities, there are nine categories of detection targets. Next, given activity tubes of nine categories, we use the Hungarian algorithm [4] to associate tubes into activity tracks. Finally, we employ different temporal detection approach for tracks of each activity group. For tracks that belong to the person-centered group, we directly output a track as an activity instance. For tracks of the vehicle-person group, a two-stage temporal activity localization system is built to detect fine-grained activity instance on tracks. Moreover, a one-stage temporal activity detection model is employed to detect activity instances of turning group.

Activities	super-category	
Closing	vehicle-person	
Closing_trunk	vehicle-person	
Entering	vehicle-person	
Exiting	vehicle-person	
Loading	vehicle-person	
Open_Trunk	vehicle-person	
Opening	vehicle-person	
Transport_HeavyCarry	person-centered	
Unloading	vehicle-person	
Vehicle_turning_left	turning	
Vehicle_turning_right	turning	
Vehicle_u_turn	turning	
Pull	person-centered	
Riding	person-centered	
Talking	person-centered	
activity_carrying	person-centered	
specialized_talking_phone	person-centered	
specialized texting phone	person-centered	

Table 1. 18 activities are divided into 3 groups



Figure 1. The overall structure of our framework

1.2 Spatial localizationTo detect activities in untrimmed videos, it's imperative to localize activity in space at first. To this end, we extend the image object detector to localize activity regions in a short video clip. Inspired by the success in object detection [1] and action recognition [5], we adapt Faster-RCNN [1] into 3D form to utilize spatiotemporal feature, 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 Faster-RCNN with I3D [5]. FPN [6] is adopted to fuse low-level appearance features with high-level semantic features to facilitate small activity region detection, as most activities only occupy a small part of a frame.

Focal loss is originally proposed to solve the foreground-background class imbalance in onestage detector, which makes the detector putting more focus on hard, misclassified examples. Now we extend the Focal Loss to multi-class case and use it in the second stage of our 3D-Faster-RCNN, which is proved to improve the recall a lot with a little harm to precision and make the detector perform better.

1.3 Tracking

We extend the task of object box tracking to activity tube tracking, which can be formulated as a data association problem. Inspired by [7], the Hungarian algorithm [4] is adopted to link tubes of adjacent clips. This data association method is light and effective. Compared with other stages, the time cost can be ignored.

1.4 Temporal localization

Given activity tracks produced by the data association module, our next step is to perform temporal detection based on tracks. We employ different strategies for each activity group. For tracks that belong to the person-centered group, we have already known the specific category at the spatial activity localization stage, thus no further analysis required, we just recognize a track as an activity. For activity tracks of vehicle-person and turning groups, multiple activities of different subclasses may occur in one track.

Our vehicle-person activity detection system consists of two stages, and models of two stage are trained independently. The first stage is an RPN-like module to propose possible 3D-proposals, named Time Region Proposal Network(T-RPN). A 3D-anchor is sent through T-RPN to predict probability that the content of the anchor corresponds to a valid activity and regress its time boundaries. T-RPN return sparse 3D-proposals as we filter out most 3D-anchors that "activity-ness" lower than a specific threshold. We set different thresholds during training and testing. The second stage module classifies 3D-proposals into activities of interest or background and perform further time refinement, we call it Time Regression ECO [8] (TR-ECO). Both stages use ECO as the backbone, 12 frames evenly sampled from an anchor or proposal are sent to the backbone network. An extra branch is added to the final layer of ECO to predict temporal offsets relative to predefined anchors. Notably, our temporal regression of TR-ECO is independent of category.

Vehicle turning group consists of left turn, right turn, and U-turn activities. We build a simple onestage temporal activity detection system to detect turning activity instances. ECO is used to classify anchors into activity of interest or background. We evenly sample 16 frames for each anchor.

1.5 Post-processing

We use two simple post-process methods to improve the performance: split long events and fuse similar events.

Split long events. According to the definition the Time-based False Alarm(TFA), the value of TFA will not be changed when we split a long event. What's more, when we split long events, the new generated events will be likely to match the unmatched reference events, so the probably of missed events may decrease. There are two reasons about the new match. On the one hand, our tracking methods may link the two events. On the other hand, the reference events may be separately labeled.

Fuse similar events. As in Figure 2, We find the person-car events usually cut off at 0.05 or 0.1 TFA on the TFA based DET curve on the validation set, which illustrate the events we proposed is not enough. We also notice that an event will be divided into other events with some probability displayed according to the confusion matrix of the classification network. For example, an Opening event may be classified to an Entering events, so we fuse the Entering events to Opening events with a multiply factor of 0.05 to the confidence score.



Figure 2 TFA based DET curve on the validation set and the confusion matrix of the classification network

1.6 Conclusion

We proposed an effective three-stage framework for spatiotemporal activity localization in surveillance videos. In this work, we further improve the results by Focal Loss to 3D-Faster-RCNN, adding OTW dataset and post-processing. The improvement and the final DET curve are as follows.

Tabl	le 2.	Impro	vement	on	the	test	set
------	-------	-------	--------	----	-----	------	-----

SYSTEM_NAME	PARTIAL AUDC	MEAN-P_MISS@0.15TFA	MEAN-W_P_MISS@0.15RFA	Delta AUDC	method
p_baseline_2	0.56781	0.48641	0.70693		FL-3D-Faster-RCNN
p_baseline_14	0.55726	0.47596	0.69853	0.01055	Add OTW dataset
p_baseline_22	0.54409	0.46306	0.74914	0.01317	Split long events
p_baseline_23	0.52408	0.4328	0.74914	0.02001	Fuse similar events



Figure 3. TFA based DET curve on the test set (p_baseline_23)

2 Instance Search

This year, we propose a similar search framework for both automatic and interactive search tasks. We extract video key frames with a sample rate of 1 fps to improve efficiency. To retrieve specific persons doing specific action, we need to consider how to achieve satisfying results in person retrieval and action recognition respectively and fuse them in an appropriate manner. The final results are summarized in Table 3. More details will be given in the following sections.

Table 3. Results for each run				
Run ID	mAP			
F_M_E_E_BUPT_MCPRL_1	11.9			
F_M_E_E_BUPT_MCPRL_2	11.6			
I_M_E_E_BUPT_MCPRL_4	21.2			

2.1 Person retrieval

For person retrieval, we take the following steps: (1) Detect persons in each frame. (2) Describe persons being detected with designed features. (3) Select metric method to distinguish different persons. First, we adopt a multi-task CNN model [9], which has proved its accuracy and efficiency in face detection. Then, we extract face features for queries and persons in this datasets based on dlib [10]. Finally, we get 128-dim face representation and conduct cosine distance between queries and detected persons.

After obtaining five rank lists for five query images, we employ an adaptive fusion scheme [11] to fuse five rank lists for each target person.

To recall correct faces, we also conduct query expansion for face retrieval. Specifically, we conduct retrieval again for several former of the first retrieval results of query. It also works. Combine the above methods, we get an excellent result in persons retrieval and several results are shown in Figure 4.



Figure 4. Person retrieval results display. We select Bradley and Sean as examples. From the left column to the right column are the first, the 1000-th, the 3000-th, the 5000-th, the 10000-th retrieval results respectively.

To further improve retrieval efficiency, we only pay attention to the shots of a specific person doing actions. For this reason, we set a threshold to filter out the shots where the specific person do not appear. We sort the remaining shots by the probability of occurrence of a specific person and get an initial ranklist for subsequent processes.

2.2 Instance Retrieval

For instance retrieval, we divide instances into three categories: emotion-related actions, humanobject interactions and general actions.

2.2.1 Emotion-related action retrieval

We group crying, laughing and shouting together as the first category. We use emotion recognition to retrieve scenes in which someone is sad, happy or angry. We train emotion recognition models based on VGG19 networks while taking FER-2013 [12] and CK+ [13] datasets as the main training set. Data augment is also performed, including horizontal flip and random cropping etc. As shown in Figure 5, we make some progress.



Figure 5. Examples of someone who is laughing, crying or shouting. Row 1-3 show a few examples of Bradley laughing, Max crying and Sean shouting respectively. Besides, the last column of images is the failure examples.

2.2.2 Human-object interactions retrieval

To retrieve human-object interactions, we explore the dependences between semantic objects and

human keypoints by using object detection and pose estimation models. Specifically, we choose YOLO to detect key objects such as glass, bag, phone, person etc. Afterwards, we feed human bounding boxes into HRNet to estimate human poses. Then, we simply calculate the distance between object location and target person's interactive keypoint to measure the dependences of object-pose such as 'holding_glass', 'holding_phone' and 'carrying_bag'. Based on the dependences, we divide the initial ranklist into several groups.

2.2.3 General actions retrieval

Other instances, such as kissing, hugging, walking, involves human-human interactions and strenuous motions. These actions could be easily recognized by action detection models, and we use two methods to retrieve these general actions.

Firstly, we use spatio-temporal networks to extract video representation. In experiments, we choose ECO as the basic network for feature extraction. Furthermore, inspired by SlowFast networks, we parallelly feed videos in different frame rates into ECO network to extract video representation, and we could observe large improvement.

The other method is using pose-based action detection models to extract video features. Considering camera movement, we propose a new pose representation by using both absolute and relative positions of poses. Specifically, we encode human poses into two feature map using absolute positions and relative distances. Then we construct a light CNN trained on JHMDB datasets to classify above pose representations. Finally, we use pretrained models to extract video vectors for general actions retrieval.

After extracting video features for every shot, we could rerank initial ranklist to get the final ranklists.

2.4 Conclusion

This year, we propose the overall framework for completing Instance Search task. We combine multiple methods and use several fusion methods, which push us to make some progress. Next year, we plan to design a more elaborate structure to improve our accuracy. We will also attempt to add text or voice information in the future work.

References

- Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
- [2] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
- [3] Castanon G, Shnidman N, Anderson T, et al. Out the Window: A Crowd-Sourced Dataset for Activity Classification in Surveillance Video[J]. arXiv preprint arXiv:1908.10899, 2019.
- [4] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset..
- [6] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

- [7] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-andtrack: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages350–359, 2018.
- [8] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–712, 2018.
- [9] Zhang, Kaipeng et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." IEEE Signal Processing Letters 23.10(2016):1499-1503.
- [10] http://dlib.net/files/dlib-19.18.tar.bz2.
- [11] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query adaptive late fusion for image search and person re-identification. In Computer Vision and Pattern Recognition, 2015. 1
- [12] http://www.pitt.edu/~emotion/ck-spread.htm
- [13] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recogniti-onchallenge
- [14] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot}, TRECVID 2019: An evaluation campaign to benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & retrieval}, Proceedings of TRECVID 2019, NIST, USA.