NTT_CQUPT@TRECVID2019 ActEV：Activities in Extended Video

Yongqing Sun[1], Xu Chen[2], Chaoyu Li[2], Kiyohito Sawada[3], Takashi Hosono[1],

Jun Zhu[2], Chengjuan Xie[2], Sixiang Huang[2], Lan Wang[2], Kai Hu[2], Qingsong Zhou[2], Chenqiang Gao[2], Jun Shimamura[1], Atsushi Sagata[1]

1 NTT Media Intelligent Laboratories, Japan

2 School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China

3 National Police Academy, Japan

Abstract

In this notebook paper, we present our activity detection system, which aims to temporally localize activities in surveillance videos. Our pipeline composed of five modules, object detection, activity proposal generation, feature extraction, classification and post-processing. We input RGB and optical flow into this pipeline separately and obtain frame level predictions by late fusion. The final detections are generated by greedily merging these predictions and filtering invalid results.

1. System Description

Activity detection in surveillance videos is a challenging task due to the low resolution, occlusion of objects and similarity between activities. In order to get reliable results, most previous participants used the method of decomposing the task into multiple subtask [1, 2, 3]. Our system for activity detection in extended videos (ActEV) in TRECVID2019[4] is composed of five modules: object detection, activity proposal generation, feature extraction, classification and post-processing. The diagram of the five modules in our system is shown in Figure 1. We evaluate and analysis each module separately in the following.

Object detection: It locates and classifies objects and activities.

Activity proposal generation: It generates candidate tubes by temporally tracking bounding boxes for activities. These tubes are called activity proposals.

Feature extraction: It finetunes the backbone network and extracts features for activity proposals.

Classification: It trains classifier to classify activity proposals.

Post-processing: It merges the activity proposals for activity localization.

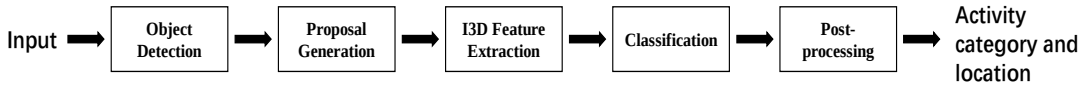


Figure 1: System Overview.

2 Object detection

All the 18 activities handled in this task are people-centered or vehicle-centered. Thus, obtaining person and vehicle bounding boxes for each frame could be an effective way. We use the Mask R-CNN [5] with the feature pyramid network on ResNet-101[6] as the backbone for object detection, and use different labeled data for training. Due to the limitation of official annotation (only activities related objects are labeled), we labeled ~3000 images for four main objects in the training and validation set by ourselves, including persons, vehicles, bikes, and boxes. We apply object detection on every 8 frames from the videos. Full resolution images are input to the model and we train model using the full 4 object class annotation labeled by ourselves. Experiments are conducted to test the effectiveness of our detection network. We first use original Mask R-CNN without finetuning to directly detect objects, which outputs a poor result. This is because the scenes in VIRAT is different with COCO and some objects in VIRAT is difficult and dim. Then, we finetune the Mask R-CNN with our annotations, which reaches a better performance. The results are shown in Table 1.

Table 1: Object Detection Results.

Method	mAP
Original Mask R-CNN	19.6
Mask R-CNN finetuned using our annotations	44.1

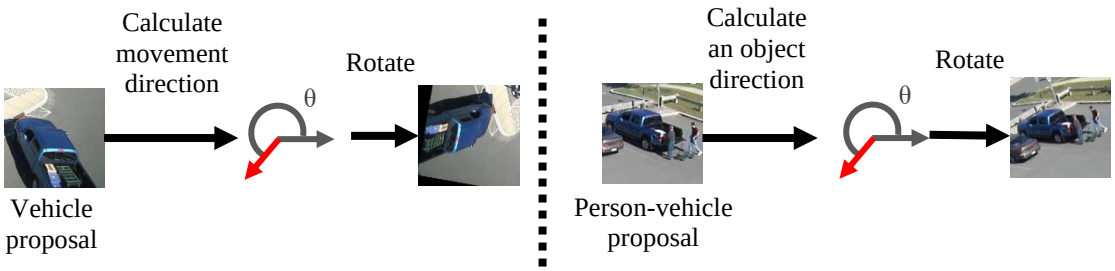
3 Activity proposal generation

Tracking and removing background We first link the person/vehicle bounding boxes by Deep SORT [7] and pad them spatially. Since the objects contain a huge number of still persons or vehicles, we remove tracks with small movements using the background modeling method, which greatly decreases the irrelevant background objects and reduces the cost of computation for subsequent steps. Tracks generated by tracking can be regarded as the raw proposals, but they lack interaction activities. Therefore, we compare the spatial distance between all the tracks and combine the person-centered and vehicle-centered tracks as the proposal of the interactive activities if they are close enough to each other. This step produces total 4,151 proposals on the validation set and achieves a recall rate of 85.6%.

In this work, we do not use the sliding window to further generate more refined proposals. We label frames of each proposal at intervals of 8 frames for subsequent classification. This strategy can reduce duplicate feature extraction, and we can link individual predictions to produce the final activity prediction results.

Alignment moving/objects direction each activity has diversity of appearances since even in the same activity, there are various movement/object direction. This diversity makes learning and inferring activity recognition difficult especially when there are few training data. Therefore, before extracting features, we rotate input proposals to align the direction in order to reduce the diversity and improve recognition accuracy. Figure 2 shows two types of processes for alignment directions. We calculate movement/object direction for each proposal. Then we rotate proposals so that the direction is zero degrees. In case of a proposal containing only vehicle (vehicle proposal), it is rotated based on the movement direction since activity of vehicle proposals often have large movement (Fig.2 (a)). In case of a proposal containing person and vehicle (person-vehicle proposal), it is rotated based on the vehicle direction since activity of person-vehicle proposals often change

the vehicle direction (Fig.2 (b)). The details of calculating the movement/object direction are described below.



(a) Vehicle proposal alignment (b) Person-vehicle proposal alignment

Figure 2: Overview of alignment movement/object direction.

a) Calculate Movement direction regarding vehicle proposal, the direction of movement is calculated from the optical flow. A flow histogram is calculated by voting on the angle calculated from the optical flow of each pixel. Here, we use only the first half of the proposal since the movement direction of the vehicle changes greatly between the first half and the second half. Then, the mode value of the histogram is selected as movement direction.

b) Calculate Object direction regarding person-vehicle proposal, the direction of a vehicle is calculated from the image gradient obtained by a Sobel filter. A gradient histogram is created by voting on the angle calculated from the gradient of each pixel in the first half of the video frame. Then, the mode value of the histogram is selected as the object direction.

We conducted experiment to confirm the effectiveness of the alignment module with the validation data. In this experiment, we used the classification module, which will be described in Section 4. Proposals are generated by ground truth data. The results of mAP for classification with/without the alignment module are shown in Table 2. As can be seen from the results, the alignment module improve mAP. Note that the alignment module not integrated into the submitted system, since we could not confirm the effectiveness when testing with our activity proposals.

Table 2: mAP for classification with/wo the alignment module.

	mAP
Without alignment module	0.46
With alignment module	0.50

4 Classification

Feature Extraction We utilize I3D network [8] for feature extraction. We consequently crop 64 frames from each ground truth activity tube or activity proposal with an interval of 32 frames as the input of I3D and output the feature map of 5-th layer. Both rgb and optical flow frames are used for training and evaluation. To get a better feature extractor, we finetune I3D using the officially annotated samples on the training set. The results on the validation set are shown in Table 3.

Classifier We first use the SVM classifier, and it flattens feature maps of the 5-th layer and loses the temporal information. However, temporal information is extremely important when modeling activities. Thus, we continue to investigate the LSTM framework to preserve the temporal

information. Specifically, we use the Conv-LSTM network [9], which can process the spatial parts sequentially, and enable the information to go through in a convolution manner. Therefore, the spatial and contextual information can be modeled efficiently by convolution operations.

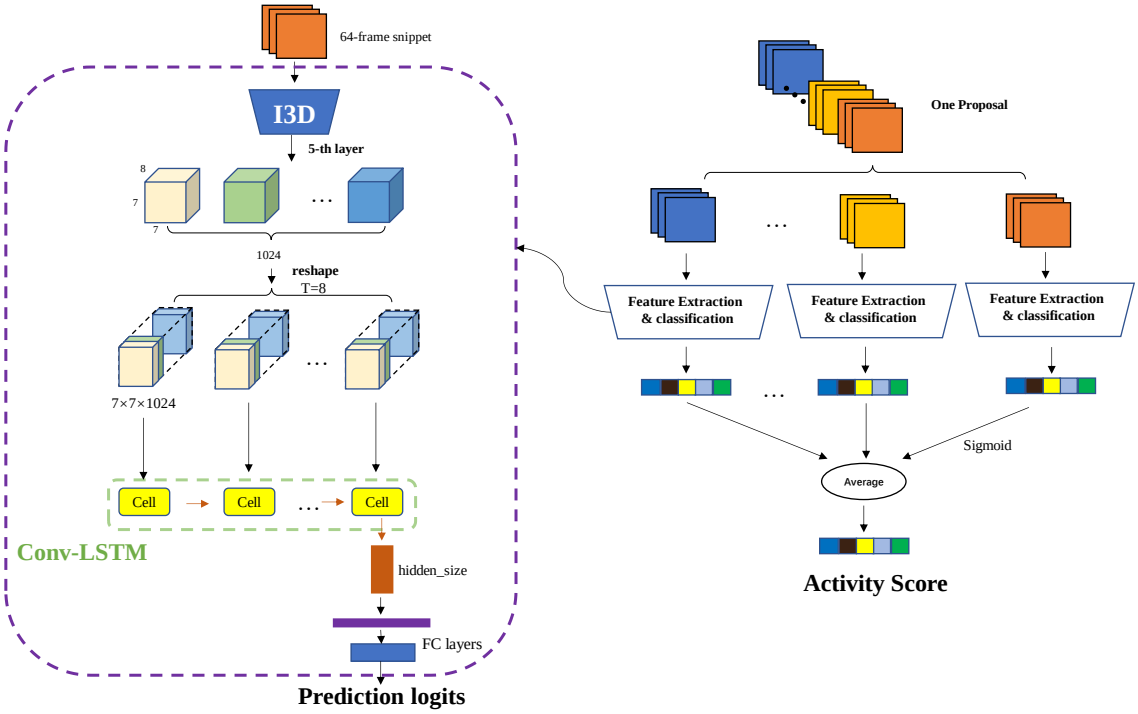


Figure 3：overview of feature extraction and classier.

Figure 3 shows the process of Conv-LSTM. After obtaining feature map of the 5-th layer, we reshape it to 7×7×1024×T (T=8) to separate the temporal dimension. Thus, the time step of Conv-LSTM is set as T. Table 2 presents the classification results of Conv-LSTM.

Table 3: Results of Conv-LSTM classifier.

	mAP
rgb	0.1315
rgb(finetune)	0.1671
flow	0.1276
flow(finetune)	0.1308

5 Post-processing

After classification, we fuse the scores of two modalities and average the scores in three different epochs for each activity. Since we do not use temporal boundary regression, a simple boundary refinement is implemented. We compute the mean score of 18 activities and the Euclidean distance between this value and score of boundary frame. Then we remove it if the distance is too large.

The output activity predictions need to be further merged to localize the temporal position of the activity. We propose several steps for post-processing. Firstly, we filter out predictions with low scores using a set threshold $\{t_c\}$. Each activity category is filtered separately. If a proposal exceeds the threshold in multiple activity category scores, we copy it multiple and mark the

corresponding activity labels. After that, predictions are merged to long trajectories. Our merging principles are simple: 1) If there is an overlap in time for two predictions, we merge them. 2) predictions of each category are merged independently. 3) The score of merged proposal is the average of predicted scores. Finally, we remove the result that the object type does not match the activity type, and apply NMS to them.

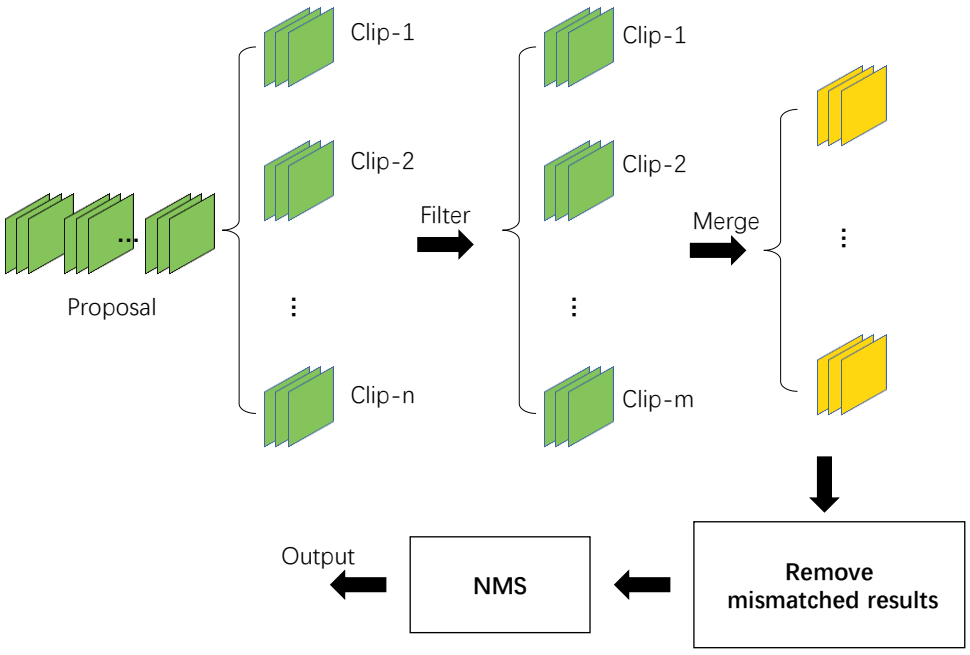


Figure 4: An illustration of Post-processing.

We validate our system using AD metrics of TRECVID[10]. When inference on test set, the classifier is trained on the training and validation sets. We submit three systems on the test set, the only difference between them is threshold set $\{t_c\}$ for 18 activities. The final submission results are shown in Table 4.

Table 4: system results on test set.

System	Partial AUDC	Mean-Pmiss@0.15TFA	Mean W-Pmiss@0.15RFA
p-NTT-CQUPT	0.60058	0.51122	0.87254
p2_NTT_CQUPT	0.60396	0.51677	0.87168
system2	0.60524	0.51755	0.87381

6 Conclusion

We designed a multi-stage pipeline for activity detection. In this work, many strategies including ensemble, two-stream and specific rules are used to boost the final performance. The classifier gets a low mAP due to the lack of training samples. We will explore more effective training methods and better network architecture in future works, especially focus on the case of small sample learning.

References

[1] Gleason, J., Ranjan, R., Schwarcz, S., Castillo, C., Chen, J. C., & Chellappa, R. (2019, January). A Proposal-Based Solution to Spatio-Temporal Action Detection in Untrimmed Videos. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 141-150). IEEE.

[2] Chen, J., Liu, J., Liang, J., Hu, T. Y., Ke, W., Barrios, W., ... & Hauptmann, A. G. (2019, January). Minding the Gaps in a Video Action Analysis Pipeline. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) (pp. 41-46). IEEE.

[3] Long, F., Cai, Q., Qiu, Z., Hou, Z., Pan, Y., Yao, T., & Ngo, C. W. (2019). vireoJD-MM at Activity Detection in Extended Videos. arXiv preprint arXiv:1906.08547.

[4] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Qunot, "Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search retrieval," in Proceedings of TRECVID 2019. NIST, USA, 2019.

[5] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 3645-3649). IEEE.

[8] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[9] Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802-810).

[10] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, New York, NY, USA, 2006, pp.321–330, ACM Press.