VANT at TRECVID 2018

Nakamasa Inoue, Chihiro Shiraishi,
Aleksandr Drozd, Koichi Shinoda
Tokyo Institute of Technology.
inoue@ks.cs.titech.ac.jp, siraisi@ks.c.titech.ac.jp
alex@smg.is.titech.ac.jp, shinoda@c.titech.ac.jp

Activities in Extended Video (ActEV)

We propose a system for activity detection, which utilizes the Action Tubelet (ACT) Detector [1] to localize activities in video data. Our network is trained for all of activities in the ActEV dataset with a backbone convolutional neural network pre-trained on the ImageNet dataset. We inserted a thresholding module to the original ACT framework to adapt detector to the ActEV task [2], since activities in this task appear more sparsely distributed than those in the action detection task in [1]. Our result was 0.882 in mean-p_miss@0.15rfa [*1] at the AD Leaderboard Evaluation.

1 Method

The overview of our system for the ActEV task is shown in Figure 1. A thresholding module is inserted to the original ACT detector framework [1], which consists of a tubelet-extraction module and a tube-creation module. Our system consists of three modules as a result: extracting tubelets, thresholding and creating tubes.

The tubelet-extraction module takes image frames as input and outputs spatio-temporal regions of objects associated with the target activity. These regions are extracted from image frames in the following two steps as in [1]. First, deep features are extracted using a backbone CNN from each image frame in a video clip. Second, the Single-Shot MultiBox Detector framework [3] extended to multi-frames is applied to them to detect spatio-temporal regions, namely tubelets, of objects. In this step, an appearance probability of the target activity is also calculated for each tubelet.

The thresholding module has a function to reject detected tubelets based on the probabilities. This module is introduced to adapt the detectors to the ActEV task, which includes many objects not associated with the target activity. Thresholding parameters are optimized by grid search on the validation data of ActEV dataset.

The tube-creation module concatenates tubelets by computing the overlap between tubelets as in [1] to output activity tubes.

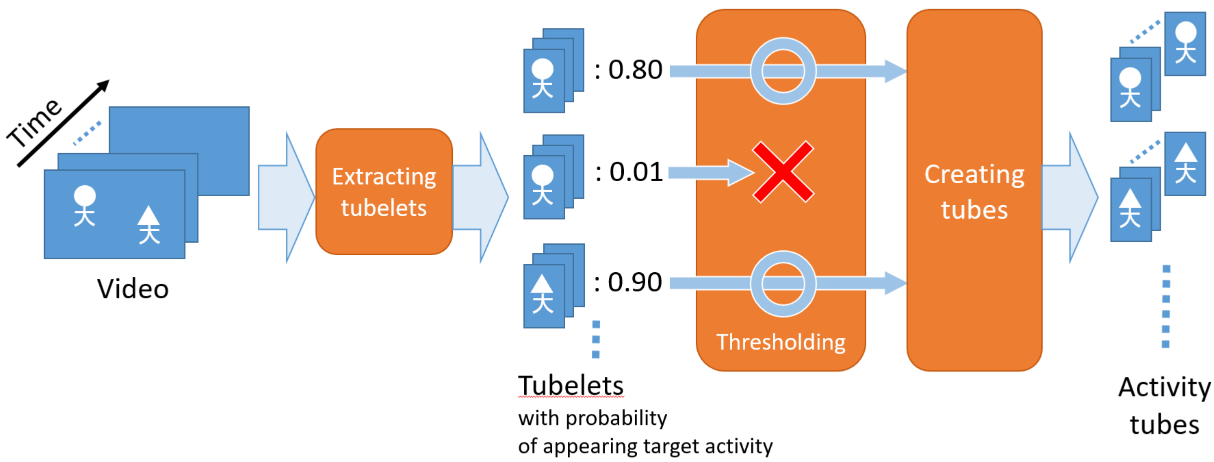

Figure 1: Overview of our system

[*1] probability of missed detections at condition which allow false alarm 0.15 times a minute

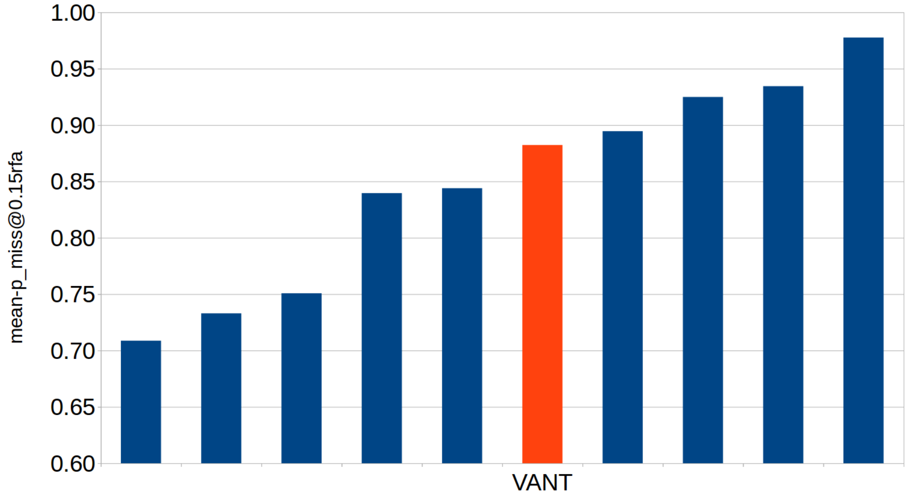

Figure 2: Overview results of the AD Leaderboard Evaluation. Runs are sorted with mean-p_miss@0.15rfa (lower is better). Our run is labelled as red.

	mean-p_miss@0.15rfa
w/o thresholding	0.885
w/ thresholding 0.05 (submitted)	**0.856**

Table 1: Comparison of baseline(original ACT-detector) and ours on validate dataset.

2 Experiments

We submitted one run to the leaderboard evaluation, which uses the best threshold (0.05) on the validation data. Our result was 0.882 in mean-p_miss@0.15rfa, which is ranked 6th among 10 teams as shown in Figure 2. Table 1 shows results with and without the thresholding module on validation data. We see that thresholding helps to reduce errors by 2.9% p_miss value.

3 Conclusion

We participated in Activities in Extended Video (ActEV) and evaluated our system using ACT detectors. Our run achieved 0.882 in mean-p_miss@0.15rfa at AD Leaderboard Evaluation. Future work will be focusing on temporal modeling of activities.

References

[1] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action Tubelet Detector for Spatio-Temporal Action Localization. In *ICCV 2017 - IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[2] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Qunot, Joao Magalhaes, David Semedo, and Saverio Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018. `http://www-nlpir.nist.gov/projects/tvpubs/tv18.papers/tv18overview.pdf`.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.