Activities in Extended Video

Fan Ma, Peike Li, Linchao Zhu, Xuanyi Dong, Yanbin Liu, Yi Yang
UTS-CETC

Abstract

In this paper, we present a system based on detection, tracking and 3D convolution neural network dealing with Activities in Extended Video (ActEV) task in TRECVID 2018. In the proposed system, videos are first unfolded into frames for training detection network, then we use it to generate bounding box for tracking areas where target activities could be happen. The tracking clips are then classified using a 3D convulution network.

1 Introduction

Activities in Extended Video(ActEV) in TRECVID 2018[1] is an extension task of the annual TRECVID Surveillance Event Detection (SED) evaluation by adding a large collection of multi-camera video data, both of simple and complex activities. ActEV will address activity detection for both forensic applications and for real-time alerting. The VIRAT-V1 dataset is used for ActEV evluations. The ActEV leaderboad evalution contains following 20 activities: *Closing, Closing_trunk, Entering, Exiting, Loading, Open_Trunk, Opening, Transport_HeavyCarry, Unloading, Vehicle_turning_left, Vehicle_turning_right, Vehicle_u_turn, Interacts, Pull, Riding, Talking, activity_carrying, specialize_talking_phone, specialized_texting_phone*. The training data contains 64 videos, which include 1426 target activities with 267139 frames. The validation data contains 54 videos, which include 1223 target activities with 201944 frames. There are two tasks this year about this dataset. One is Activity Detection(AD) task, given a target activity, a system automatically detects and temporally localizes all instances of the activity. For a system-identified activity instance to be evaluated as correct, the type of activity must be correct and the temporal overlap must fall within a minimal requirement. The other is Activity and Object Detection task, given a target activity, a system detects and temporally localizes all instances of the activity and spatially detects/localizes the people and/or objects associated with the target activity. For a system-identified instance to be scored as correct, it must meet the temporal overlap criteria for the AD task and in addition meet the spatial overlap of the identified objects during the activity instance.

We designed a system for the AD task based on object detection, tracking and 3D feature extraction technologies. The detailed architecture for the proposed system are described in Section 2.

2 Retrospective System

The proposed system first detect and crop clips from the given videos where and when the target events could happen based on object detection and tracking algorithms. Then those selected clips are represented by features through 3D convolution networks which are pretrained on action recognition dataset. An classifier is later trained on the selected clips and given labels. In the testing stage, the classification result on selected clips are submitted for evaluation.

2.1 Object Detection and Tracking

Object detection algorithms have been well developed in recent years. And Faster-RCNN[6] is one of the most wildly applied algorithm, it employs two step detection mechanism which first propose interesting region and then classify it. Based on the annotated information of training VIRAT V1 dataset, we fine-tuned the Faster-RCNN with inception and residual network as backbone architecture. Since a few objects are related to target activities and many objects are too small to be recognized, we should only focus on the objects which are easy to be detected and are likely to be involved into a action. Therefore, only person, bicycle and car are detected in our system. We evaluation the detection performance on the validation data, and get map over 95.

Once detection for each frame has been done, the system will track the target with bounding box given in every frame. Here we employed [2] to track the object. And for each detected object, the tracking result with its bounding box are used to crop clips from the whole videos in spatial and temporal. A Almost all the target activities are contained in the cropped clips. And the cropped clips can be seen the proposal action cubes compared to the proposal interesting regions in detection task.

2.2 Feature Extraction

For action recognition task, the representation of videos are very important. Recent researches employed 3D convolution neural network to extract fetarues of given videos. Here we employed I3D[3] to extract features of selected clips. The I3D model is pretrained on kinetics dataset[4]. We also calculate the flow feature based one the flow using TVL1[5] optical flow algorithm.

2.3 Classification

SVM and lightGBM are the classifiers in the proposed system. Each classifier are trained on extracted RGB and Flow features separately, and in inference stage, the probabilities from each calssifer on different source are fused to generate the final predictions. We submit two results based on different classification settings. The "base" gets 0.9510 mean-p_miss@0.15rfa, and the "window 2" gets 0.9249 mean-p_miss@0.15rfa. w The "window 2" means we split cropped clips into several 2s segments, and adopt the split segments for classification.

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, J. Magalhaes, D. Semedo, and S. 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.

[2] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.

[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[5] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.