Object-Centric Spatio-Temporal Activity Detection and Recognition

Mandis Beigi, Lisa M. Brown, Quanfu Fan, John Henning, Chung-Ching Lin,
Honghui Shi, Chiao-fe Shu, Rogerio Feris
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

ActEV Submission Summary

We only used the data provided by NIST. Below are the Run IDs, submission names, AD and AOD scores followed by a brief description of the entry.

Run-ID (AD/AOD) Submission AD AOD Note
283/284 IBM_E2E 0.8396338981 0.8433932966 S1: Baseline entries of IBM system for eval-1a
330/331 IBM_E2E_ALL (TRN3-TRN3 filtered) 0.7716112433 0.7928942585 S2: Baseline entries of IBM system for eval-1a, for all 19 activities.
351/352 IBM_E2E_ALL (unfiltered) 0.7706395834 0.7932181452 S2, but with unfiltered results
355/356 IBM_E2E_ALL (filtered & merged) 0.7762367494 0.7958497241 S2 with a merging algorithm
371/- IBM_E2E_ALL (TRN16) 0.7953969555 - S3: TRN 16 instead of TRN 3.
397/398 IBM_E2E_ALL_New (unfiltered) 0.7589621185 0.7785331614 S4: S2 + new turning algorithm
436/438 IBM_E2E_ALL_New (unfiltered + Ensemble) 0.7189424498 0.7629197006 S5: S2 + newly trained activity classifiers
453/440 IBM_E2E_ALL_New (unfiltered + Ensemble + LR-turns) 0.7087156227 0.7520320714 S6: Our winning entries, S4 + S5.

A significant difference between the final runs was achieved from a post merging process to combine proposals. Other improvements were achieved in subsequent runs but it is difficult to analyze the specific contributions of components. For our system, the validation set provided a good measure of generalization to the test set.

Introduction

Our ActEV (Activities in Extended Video) experiments from TRECVID 2018 [5] utilized a feature pyramid network (FPN) combined with a deformable convolutional network (DCN) to perform very accurate and fine-grain object detection. This approach provides a strong baseline for our subsequent action detection and leverages IBMs pioneering work in multi-scale CNNs [1]. Object detection is followed by tracking and action proposals; the latter are performed separately for the three classes of actions: vehicle-turns, vehicle-person-interactions, and person-object-interactions. Proposals are generated

analogously to a region proposal network in object detection, but on activity tubes cropped out from the original video. Our final action classification is based on an ensemble of temporal relational networks.

Background

Unlike the vast majority of action detection systems, the ActEV challenge requires both spatial and temporal localization. Three major approaches can be taken to address this multi-scale localization, i.e,

1) activity detection (AD) such as R-C3D [2] followed by object localization (OL), herein referred to as AD->OL;
2) joint activity detection and object localization, i.e., Action Tubelets [3], herein referred to as AD+OL
3) object localization followed by activity detection, herein referred to as OL->AD

While AD+OL can parallelize activity detection, only OL->AD can address issues of low resolution, occlusion, clutter and multiple objects, overlapping activities and activities of varied length. Most importantly for the ActEV challenge, OL->AD can leverage state of the art object detectors and trackers to handle low resolution objects and provide stronger cues about where and when an activity occurs.

Figure 1 shows an overview of our OL->AD approach. We start with object detection and tracking and proceed with activity tube generation. The latter is primarily focused on spatial localization and does not attempt to separate related actions. This is followed by proposal generation which samples the temporal extent around specific actions. Lastly, we perform activity classification using a temporal relational network [5].

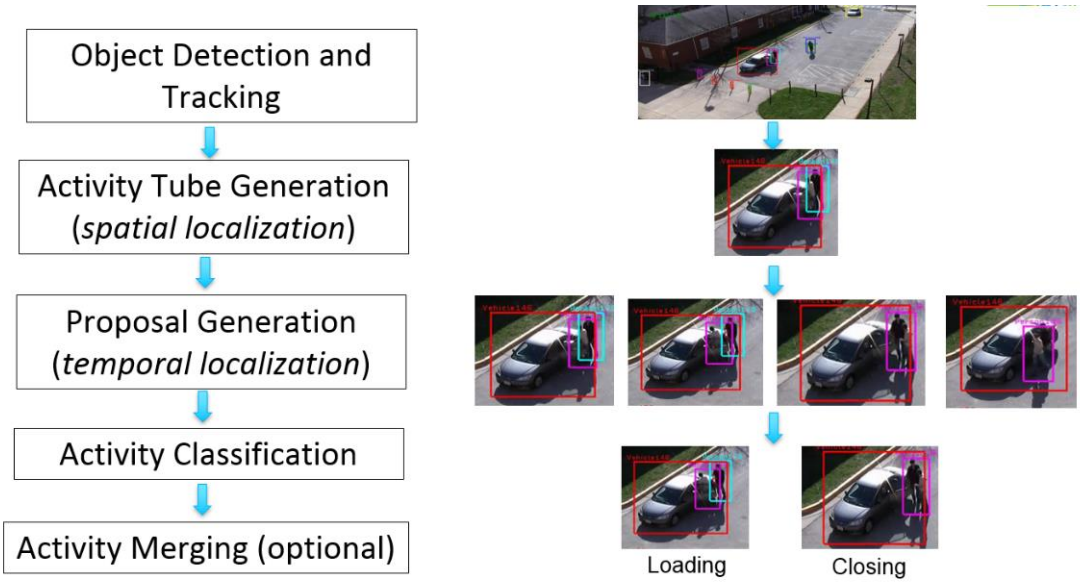

FIGURE 1. Overview of Our Approach

Object Detection & Tracking

Object detection for the ActEV challenge must address small objects, grouping, clutter, occlusion and very limited training samples. We have explored a few state-of-the-art object detection system choices for vehicle and people detection on the VIRAT dataset, including YOLO, SSD, Faster RCNN, FPN with Deformable ConvNets. We selected FPN with Deformable ConvNets in the end due to its capacity in accurately detecting small scaled vehicles and people. We trained a few models with different learning rate schedules and pretrained networks, and we have achieved overall high mean average precision and recall. Our final detection model was trained with all the DIVA V1 training and validation data and achieved 97% and 99% mAPs on training set for people and vehicle respectively. Results on the object types available in the VIRAT data are shown in Figure 2.

Object detection is followed by detection-based tracking [5]. Detected bounding boxes are used to update the target states where the velocity components are solved optimally via a Kalman filter framework. The assignment cost matrix is computed as the intersection over union distance between the detection and the predicted bounding boxes from existing targets. Assignment is solved using the Hungarian algorithm.

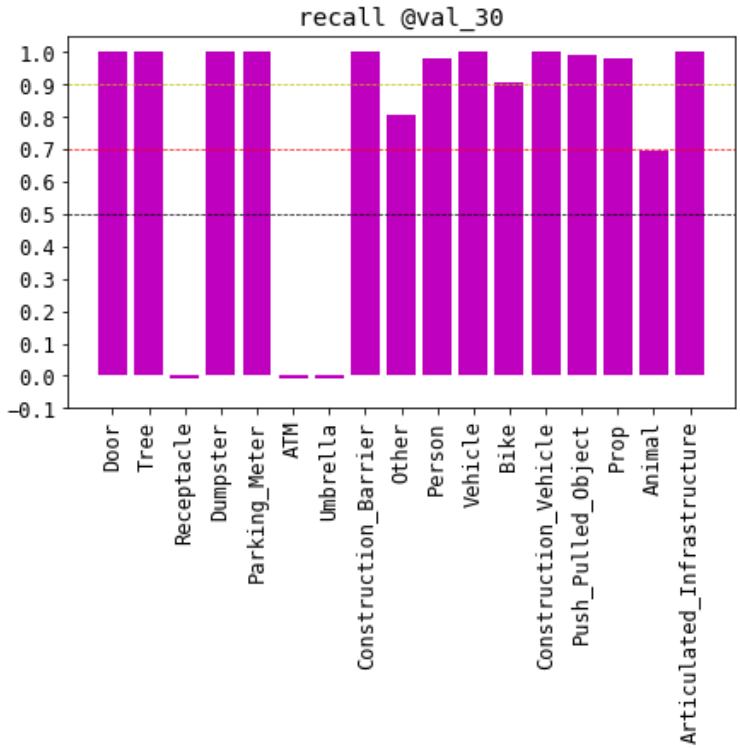

Figure 2. Results of Object Detection on VIRAT data

Action Detection: Activity Tube Generation, Temporal Proposal Generation, Action Classification

Following object detection and tracking, spatial-temporal activity localization is performed in three steps. Initial spatial and coarse temporal segmentation are performed using Activity Tube Generation. An activity tube is a cropped-out video containing one or multiple activities centered at the objects of interest in the activities. Secondly, object-centric activity tubes are extracted through analysis of person-object interactions as well as individual object trajectories. These tubes indicate the spatial locations of where activities of interest can possibly occur. Lastly, temporal proposals are generated from the tubes,

analogous to region proposal network in object detection, and the final action classification is based on an ensemble of temporal relational networks.

In the second stage of proposal generation, we construct temporal proposals utilizing the type of action class. Specifically, for vehicle centric actions the vehicle is the primary object and we localize using the proximity of the person and the vehicle. For person-centric actions the person is the primary object and we localize using the proximity of the person to the relevant objects (prop, pulled object, another person or bike). Lastly for non-interactive actions either the vehicle or person track is used.

Proposal generation uses a temporal relation network [4] with 16 frames to "detect" the optimal temporal localization of the action. Figure 3 shows the results of this proposal generation on the validation set. Notice, we can achieve 89% retrieval of actions using this approach at a threshold of 0.1. This was based on a non-maximum suppression of 0.7 and minimum detection confidence of 0.2.

In the last stage of action detection, we use Temporal Relation Networks. Temporal relation networks are efficient and straightforward to train. They handle varied lengths of activities and can recognize many of the ActEV actions with only a few frames. In the TrecVid evaluation, we use a combination of TRN-3, TRN-4 and TRN-8 action classification models. The final results, for which we obtain top performance, are shown in Figure 4.

Threshold	0.01	0.05	0.10	0.20	0.30	0.40	0.50
Entering	0.873	0.873	0.873	0.859	0.789	0.704	0.620
Exiting	0.846	0.846	0.846	0.831	0.754	0.585	0.431
Opening	0.898	0.898	0.898	0.843	0.661	0.512	0.331
Closing	0.932	0.932	0.932	0.833	0.583	0.356	0.174
Loading	1.000	1.000	1.000	1.000	1.000	0.892	0.811
Unloading	0.875	0.875	0.875	0.875	0.844	0.812	0.812
Open_Trunk	0.955	0.955	0.955	0.955	0.818	0.636	0.500
Closing_Trunk	1.000	1.000	1.000	0.952	0.810	0.571	0.429
talking_phone	1.000	0.941	0.706	0.706	0.706	0.647	0.647
Talking	0.780	0.780	0.756	0.683	0.561	0.439	0.366
Interacts	0.916	0.874	0.853	0.747	0.632	0.579	0.453
texting_phone	1.000	0.750	0.750	0.500	0.500	0.500	0.500
Transport_HeavyCarry	0.968	0.968	0.968	0.903	0.839	0.677	0.516
Pull	1.000	1.000	1.000	1.000	0.957	0.870	0.870
Riding	0.955	0.955	0.955	0.955	0.909	0.864	0.818
activity_carrying	0.922	0.893	0.878	0.815	0.722	0.634	0.566
Average	0.914	0.902	0.891	0.836	0.717	0.594	0.480

Figure 3. Results of Proposal Generation on Validation Set

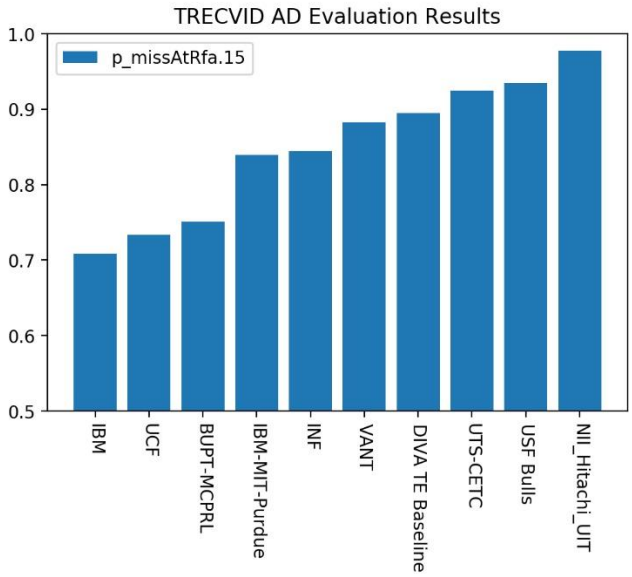

Figure 4. Results on TrecVID ActEV Challenge AD Track

Conclusions

We proposed and developed an effective object-centric approach for multi-scale spatio-temporal activity localization. We demonstrated that accurate object localization is critical and provides an important cue for both spatial and temporal localization. We plan to explore object-guided attention to enhance activity detection and to integrate activity proposing and classification into a single end-to-end system. We also intend to apply sequence modeling to explore temporal dependencies between actions.

References

[1] Zhaowei Cai, Quanfu Fan, Rogerio Feris and Nuno Vasconcelos, "A Unified Multi-scale Deep Convolutional Neural Network For Fast Object Detection," CVPR 2016.
[2] Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: ICCV (2017)
[3] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, Cordelia Schmid. Action Tubelet Detector for Spatio-Temporal Action Localization. In: ICCV (2017)
[4] Bolei Zhou et al., Temporal Relation Networks, ECCV 2018
[5] Bewley et al., Simple online and realtime tracking, ICIP 2016
[6] George Awad and Asad Butt and Keith Curtis and Jonathan Fiscus and Afzal Godil and Alan F. Smeaton and Yvette Graham and Wessel Kraaij and Georges Quénot and Joao Magalhaes and David Semedo and Saverio Blasi, "TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search," Proceedings of TRECVID 2018, NIST 2018 {http://www-nlpir.nist.gov/projects/tvpubs/tv18.papers/tv18overview.pdf}.