論文 / 著書情報
Article / Book Information

Title	Tokyo Tech at TRECVID 2020: Relation Modeling for Video Action Detection
Authors	Ronaldo Prata Amorim, Nakamasa Inoue, Koichi Shinoda
Citation	TRECVID 2020 Notebook Papers, , ,
Pub. date	2020, 12

Tokyo Tech at TRECVID 2020: Relation Modeling for Video Action Detection

Ronaldo Prata Amorim
ronaldo@ks.c.titech.ac.jp

Nakamasa Inoue
inoue@ks.c.titech.ac.jp

Koichi Shinoda
shinoda@ks.c.titech.ac.jp

Tokyo Institute of Technology

Abstract

We propose an action detection system for detecting human and vehicle actions in long untrimmed videos, submitted for the TRECVID Activities in Extended Video (ActEV) 2020 challenge [1]. It utilizes an object detection and tracking stage to divide the initial video into object tracks for all possible actors, followed by action localization to temporally localize and classify all actions within these tracks. Finally, we conduct several experiments into spatial and temporal relation modeling, both showing limited performance improvement, but demonstrating the possibility of similar approaches for future video action detection research.

Besides the VIRAT dataset utilized for the challenge, we utilize networks pretrained on the ImageNet and ActivityNet datasets. Summaries of the different submitted runs are as follows:

- 22342 - TTA-baseline: Standard two-stage system without any relation modeling
- 22442 - TTA-SRM: Same as baseline, but utilizing spatial relation modeling post-processing
- 22658 - TTA-SF2: System using multiple sampling rates for temporal action localization
- 22657 - TTA-SF: Same as SF2, but utilizing spatial relation modeling

From the run results, we can see that utilizing the multi-sampling rate action localization slightly improves performance, while the relation modeling decreases performance, contrary to our validation experiments. This seems to indicate that our relation modeling is still premature.

1 Introduction

Over the past few years, video understanding has been receiving much attention within the research community, as it may represent a culmination of previous breakthroughs in related areas such as image understanding and object tracking. But it is still far behind these in performance and reliability, partially due to the many difficulties inherently present in video analysis, such as the larger degree of object variability and the larger computing power required to process entire videos in a unified manner. As such, in spite of the advent of ever bigger and denser video datasets annotated for such tasks, we currently lack the ability to process these videos taking into account both spatial and temporal features simultaneously.

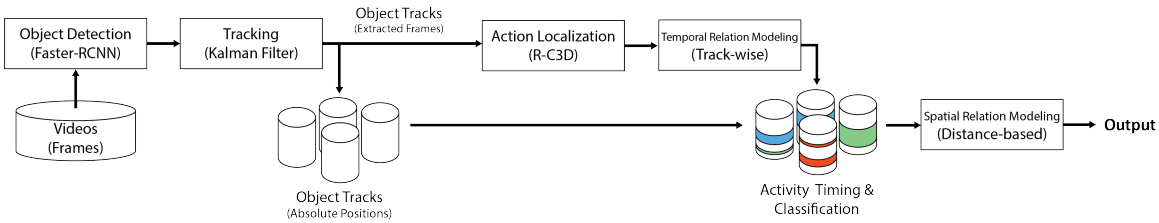

Figure 1: System overview

In view of this, we choose a method which attempts to isolate the spatial and temporal characteristics of a video, first spatially identifying actors which are expected to perform identifiable actions, and then temporally localizing actions for each actor, reassembling from these the full spatio-temporal actions. Finally, we utilize action relation modeling, in which we analyze both the spatial distance relation between actors and the temporal relation between actions, in order to refine detections over long videos.

2 System

Our system is based on a framework with two stages, a spatial stage and a temporal stage, as shown in Figure 1. First, an object detection network is run for each frame of the input video, localizing all possible objects present in the video. These detections are then concatenated into object tracks utilizing a heuristic-based object tracking system. These object tracks are then used as input for an action localization network which temporally localizes actions performed by each object. These detected actions are fused with the previously found spatial information to identify the full spatial-temporal actions.

Finally, in order to rescore and remove erroneous proposals, we analyze modeling relations between action proposals, both from the spatial and temporal aspects. Both take the form of post-processing, taking as input the list of previously detected action results and refining them by modifying action classes, tweaking scores and removing proposals where necessary.

2.1 Object Detection and Tracking

Given a video V composed of T frames, described as $V = \{I_t \in \mathbb{R}^{W \times H \times 3}\}_{t=1}^{T}$ with I_k being the k_{th} frame of the video with width W height H, we run an object detection network, e.g. Faster R-CNN [7], for every k_{th} frame, spatially localizing and classifying all objects on each frame, with special attention to those that are likely to perform the actions we wish to detect, which we call actors. In the case of the VIRAT dataset [6] utilized in this challenge, these actors are composed exclusively of persons and vehicles, so we focus on these two object classes for object detection.

This frame-wise actor detection information is then concatenated into object tracks, utilizing a Kalman Filter [8] based object tracking system. This system receives as inputs the spatial localization and class of every actor detected in the previous stage, and outputs for every actor a an object track $A_a = \{i_t \in \mathbb{R}^{w \times h \times 3}\}_{t=t_0}^{t_f}$. Here, i_k is the k_{th} frame of the video trimmed and centered around actor a, of width w and height h, t_0 and t_f are the first and last frames where actor A is present in the video respectively.

2.2 Temporal Action Localization

With object tracks for each actor in a video, we then input them to a temporal action localization network, e.g., the R-C3D network [9], which aims to temporally localize and classify all actions present in a given track. The input for this stage is the previously described object track A_a,

Table 1: Frame-wise object detection results on the VIRAT validation subset

person	vehicle	bike	parking meter	door	tree	dumpster	prop	push/pull object	mAP
31.7	69.6	2.1	73.2	57.0	94.0	89.4	5.0	17.8	48.9

with the output being a list of temporal action proposals, each proposal p consisting of a detected class c, starting and ending frame indexes t_i and t_j, and a confidence score s. This is done separately for every object track of a given input video.

Combining this information with the previously described spatial information for each respective object track, we produce full spatio-temporal action detection results. We then realize post-processing in the form of two different types of relation modeling, temporal and spatial, on top of these proposals. This is done with the primary intent of correcting wrongly detected actions, merging duplicates present in different object tracks and removing those that are detected in impossible conditions, such as actions involving interactions that are detected where there are no other objects nearby.

2.3 Temporal Relation Modeling

For temporal relations, two approaches were tested, one heuristic and one neural network-based, both analyzing the relations between proposals independently within each object track. These are based on the idea that many actions are often performed in succession to each other, such as a person loading a vehicle expected to have carried an object and opened the vehicle trunk before the loading action and closed the vehicle afterwards, with certain expected sequences having a higher probability of appearing and therefore we can complete these patterns where they're expected or remove them where they're impossible to occur. As such, for both approaches, the input is the list of proposals described as the output of the previous section, and the output is the same list with modified classes and scores as necessary.

In the case of the heuristic approach, these sequence probabilities are directly estimated from the training dataset, in the form of the percentage of times an action of class X is preceded or succeeded by action of a different class Y within a certain short time span, indicated as $p_b(X, Y)$ for the probability of Y occurring before X and $p_a(X, Y)$ for the probability of Y ocurring after X. Note that in this definition in general $p_a(X, Y) \neq p_b(Y, X)$, as they are calculated based on the occurrence rates of X and Y respectively, such that

$$p_a(X, Y) = \frac{n_{X \to Y}}{n_X} \neq \frac{n_{X \to Y}}{n_Y} = p_b(Y, X) \quad (1)$$

with $n_{X \to Y}$ being the number of times the pattern of "X followed by Y" appears, and n_X and n_Y being the number of occurrences of actions of classes X and Y respectively. With this information, we create a set of temporal relations with probabilities above an arbitrarily high threshold τ_p, and we rescore the confidence score for those proposals without relations by multiplying α $(0 < \alpha < 1)$.

On the other hand, the neural network-based temporal relation model utilizes a graph neural network in order to model temporally close predictions. With its input being the same list of temporal action proposals output by the temporal action localization stage (Section 2.2), a relation graph is built for each object track, with nodes representing each prediction, and edges set between predictions that are temporally close. Using features extracted from each proposal by a feature extractor, convolution is performed on top of this graph in order to leverage information from surrounding actions to correct mistakes in both the classification and temporal localization of each prediction.

Table 2: Temporal action localization results on the VIRAT validation subset

	closing_trunk	talking	talking_phone	entering	vehicle_moving	interacts	loading	unloading	vehicle_turning_left	unloading	activity_walking	vehicle_starting	opening	vehicle_stopping	vehicle_u_turn	pull	activity_standing	open_trunk	activity_carrying	exiting	transport_heavycarry	vehicle_turning_right	texting_phone	closing	mAP
Baseline	0.036	0.039	0.000	0.263	0.748	0.009	0.036	0.591	0.233	0.038	0.576	0.092	0.091	0.097	0.280	0.104	0.292	0.043	0.174	0.223	0.103	0.273	0.000	0.039	0.183
SF-like	0.076	0.038	0.000	0.290	0.781	0.013	0.061	0.557	0.281	0.074	0.588	0.153	0.104	0.197	0.331	0.232	0.269	0.045	0.152	0.251	0.136	0.353	0.000	0.115	0.212

2.4 Spatial Relation Modeling

For spatial relations, we model the spatial distance between action proposals in order to correct mistakes in the proposal results for actions that involve interactions between actors, such as two people talking to each other, or a person loading an object into a car, where the system is expected to individually identify the action for each actor involved in it.

Specifically, given a proposal p starting in frame t_0 and ending in frame t_f, we calculate the spatial distance between the center of its detected bounding box and the equivalent center for all other intersecting proposals in every frame between t_0 and t_f, taking the proposal with the lowest average distance over this time period as its adjacent pair. After this step of identifying paired proposals, we merge duplicated proposals identified for different actors and synchronize their temporal localization by averaging their start and end times, as well as remove actions that are expected to involve interactions but have no valid pairings.

3 Experiments

Experiments were conducted for each stage separately, testing the performance for object detection and temporal action localization, as well as for the entire video action detection task, training with the VIRAT dataset's training subset and testing on its validation subset.

For object detection we utilize a Faster-RCNN model [7] with a ResNet-50 [4] backbone, pretrained on ImageNet [2] and refined on VIRAT's own training subset for frame-wise object detection. Object detection is realized every 5 video frames for the 13 most common object classes in the VIRAT dataset, although only person and vehicle detection results are utilized for the rest of the framework. The results of these tests can be seen on Table 1.

For temporal action localization we utilize the R-C3D network [9], pretrained on ActivityNet [5] and trained on the canon tracks for each actor provided by VIRAT's training subset, aiming to temporally localize and classify all actions in which that actor participates. We experiment with two settings for video sampling, one baseline which samples every 5 frames, and another partially inspired by SlowFast networks [3], with a "fast" track sampled once every 5 frames and a "slow" track sampled every other frame, focusing on longer and shorter actions respectively. The results of these tests can be seen on Table 2.

Finally, experiments were conducted on heuristic approaches to relation modeling post-processing, refining the results given by the previous models. For spatial relation modeling, we calculate the distance between pairs of actors in order to find actors that are performing the same action, taking into account actions that involve interaction between different actors according to

their definitions. Using this information we then enforce proposal interactions, synchronizing actions between actor pairs and removing proposals for which no valid pairing is possible.

As for temporal relation modeling, we use a probability-based heuristic approach, wherein the probability of two action classes appearing in quick succession is calculated from the training subset for every class pair, and this information is used to correct mistakes in proposals, rewarding and penalizing them according to their surrounding actions and their respective probabilities.

The results of the submitted runs on the VIRAT testing set can be seen on Table 4. From it we can see that, although the multiple sampling rate action localization does indeed result in a performance improvement, that isn't the case with the usage of relation modeling, which consistently results in worse performances, contrary to the results on the validation set, which resulted on slight increases in action localization performance. This could indicate the failure of the spatial relation model implemented in correctly identifying proposal pairs with real object detection results.

4 Conclusion

We presented our framework for video action detection in the context of the ActEv challenge, and our related experiments in action relation modeling in hopes of tackling the challenges inherent in this task. We showed the results of our experiments in both the open VIRAT validation subset for individual framework stages, as well as the closed VIRAT testing subset for the full framework, where although small performance improvements were possible with adjustments to the action localization network, the relation modeling experiments did not equally translate to improvements to the full spatio-temporal action detection task. While the results of our experiments in relation modeling were not as successful as expected, we still have hope for relation modeling for action detection tasks, and hope to continue our research into such methods.

Table 3: Full spatio-temporal activity detection results on the VIRAT testing subset

	Partial AUDC	Mean p-miss
TTA-SF2	0.79753	0.75502
TTA-baseline	0.81868	0.78228
TTA-SF	0.83456	0.80451
TTA-SRM	0.85508	0.83174

References

[1] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA, 2020.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[5] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[6] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, 2011.

[7] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[8] G. Welch, G. Bishop, et al. An introduction to the kalman filter, 1995.

[9] H. Xu, A. Das, and K. Saenko. R-C3D: region convolutional 3d network for temporal activity detection. *CoRR*, abs/1703.07814, 2017.