ITI-CERTH participation in TRECVID 2019

Konstantinos Gkountakos, Konstantinos Ioannidis, Stefanos Vrochidis, Ioannis Kompatsiaris

Information Technologies Institute/Centre for Research and Technology Hellas,
6th Km. Charilaou - Thermi Road, 57001 Thermi-Thessaloniki, Greece
{gountakos, kioannid, stefanos, ikom}@iti.gr

Abstract

In this work, an overview of the submitted run to TRECVID 2019 by ITI-CERTH is presented and more specifically, for the task of Activities in Extended Video (ActEV). Towards this objective, we deployed a state-of-the-art architecture for the human action recognition problem and with the application of an encoder-decoder model, we extract a threshold for every activity in order for the framework not only to recognize activities but also to identify them in extended videos.

1 Introduction

This paper describes the recent work of ITI-CERTH[1] in the area of video analysis. TRECVID [1] has always been a target initiative for ITI-CERTH given that is is one of the major evaluation activities in the domain of video analysis. In the past, ITI-CERTH participated in the Search task under the research network COST292 (TRECVID 2006, 2007 and 2008) and in the Semantic Indexing (SIN) task (also known as high-level feature extraction task - HLFE) under the MESH (TRECVID 2008) and K-SPACE (TRECVID 2007 and 2008) EU-funded research projects. In 2009 ITI-CERTH participated as a stand-alone organization in the SIN and Search tasks, in 2010 and 2011 in the KIS, INS, SIN and MED tasks, in 2012, 2013, 2014 and 2015 in the INS, SIN, MED and MER tasks ([2], [3], [4], [5]), in 2016 and 2017 in the AVS, MED, INS and SED tasks ([6], [7]) of TRECVID and in 2018 in the AVS, INS and ActEV [8]. Based on the acquired experience from previous submissions to TRECVID, our aim is to evaluate our algorithms and systems to extend their operational efficiency and the corresponding accuracy. This year, ITI-CERTH participated in one task, namely the ActEV initiative. In the following sections we are presenting in detail the employed algorithms and the evaluation for the runs that were performed.

2 Activities in Extended Video

Regarding the ActEV (Activities in Extended Video) challenge, firstly, the state-of-the-art approach proposed by Hara et al.[9] has been deployed, and subsequently, the implementation of a linear-layered-based encoder-decoder was performed for the learning of classed-depended thresholds. The rest of the section is organized as follows: Subsection 2.1 describes the objective of the submission while Subsection 2.2 provides the details of the action recognition module. In addition, Subsection 2.3 describes the activity detection module while in subsection 2.4 the details of the submitted runs are described. This report is concluded in subsection 2.5 where the results from our evaluation runs are commented and analyzed.

[1]Information Technologies Institute - Centre for Research and Technology, Hellas

Type of dataset	Number of videos	Number of activities
Train	64	1338
Validate	54	1128
Test	246	-

Table 1: TRECVID ActEV dataset

Target activities			
Closing	Closing trunk	Entering	Exiting
Loading	Open trunk	Opening	Transport heavyCarry
Pull	Riding	Talking	Unloading
Activity carrying	Vehicle turning left	Vehicle turning right	Vehicle u-turn
Specialized talking phone	Specialized texting phone		

Table 2: Target activities in TRECVID ActEV

2.1 Objective of the Submission

The main goal of this submission is to detect and recognize activities in multi-camera extended videos. The data that are used for this purpose were acquired from VIRAT-V1 and VIRAT-V2 that is consisted of 455 and 150 videos, respectively. The dataset was divided into three separate subsets: training, validation and test set. The number of the videos for each set are depicted in Table 1. Furthermore, the extracted activities for the training and the validation sets are presented as for the test is the target. Finally, in Table 2, the corresponding 18 target activities are presented.

2.2 Activity recognition module

For the training process of our system, we utilized the provided videos and the corresponding activities that were extracted using the annotated data. More specific for the training and the validation sets, we have extracted all the frames for every activity that was active during one video period, by sampling every 4 frames. Furthermore, we have trained the proposed system with a sampling number set equal to 1 in order to evaluate the impact of the sampling factor. Finally, the number of samples for each activity that used for both training and validation of the proposed framework are presented in figure 1.

A supervised learning framework for activity recognition that employs a deep neural network architecture, namely the 3D-ResNet neural network was adopted. The latter comprises a 3D-convolutional-based architecture that achieves faster processing and can thus perform activity recognition in near and/or real-time state, while using simultaneously (batch) frame processing. In particular, the architecture with 50 layers as described in [9] has been implemented. Specifically, the architecture consists of bottleneck blocks, where each block is consisted of three 3D-convolution layers followed by batch normalization and ReLU activation layers, with the convolution kernels being 1x1x1 for the first and the third convolution layers while the intermediate layer a size of 3x3x3 is applied. Finally, it should be highlighted that for all the architectures, the weights of the Kinetics dataset [10] were pre-loaded. The Kinetics dataset was selected for the application as it covers a large number of human activity classes (400 classes).

We have evaluated the activity recognition framework using the validation data of the ActEV dataset. It should be noted that the modified dataset that we have created for the training and the validation of our activity recognition module is a highly unbalanced dataset as can be observed in 1. Specifically, we have achieved more than 0.28 on top-1 retrieved results and more than 0.55 on top-3 as indicated in Figure 2.

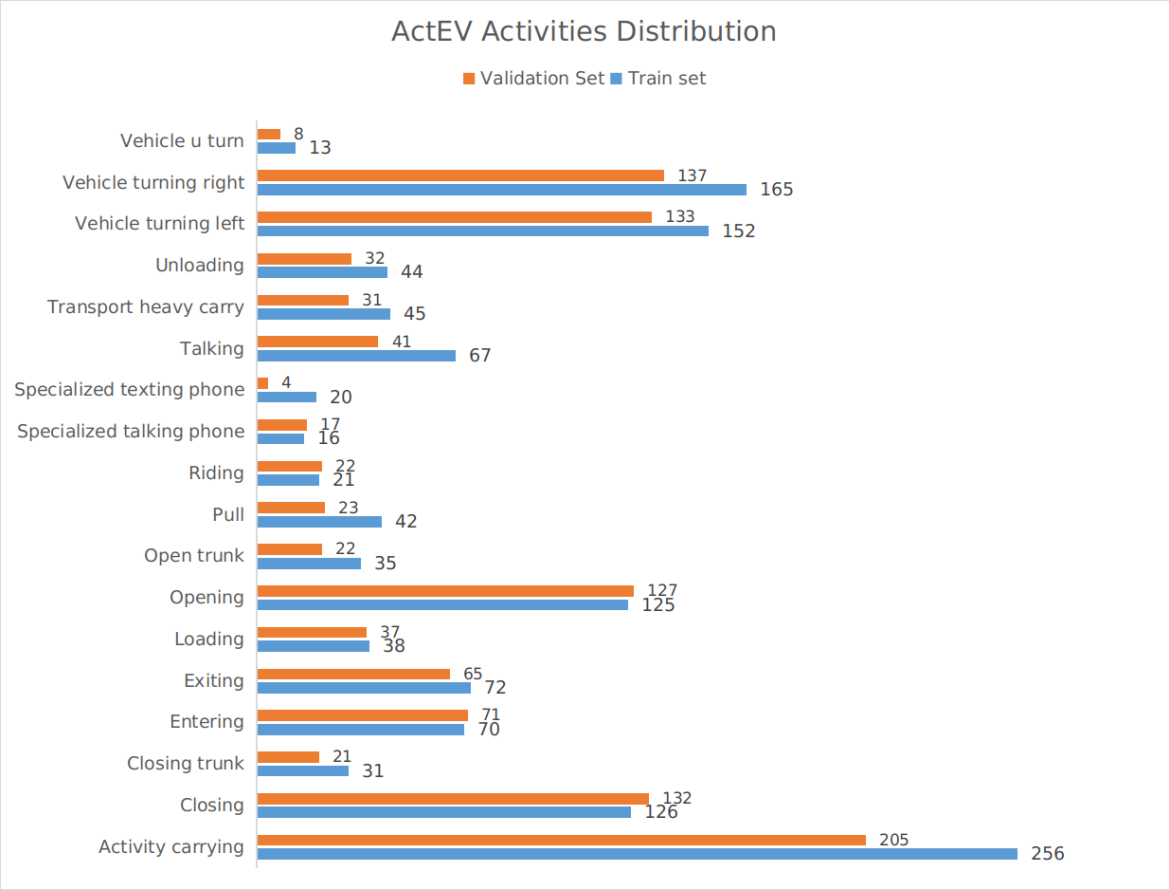

Figure 1: ActEV activities distribution for train and validation sets

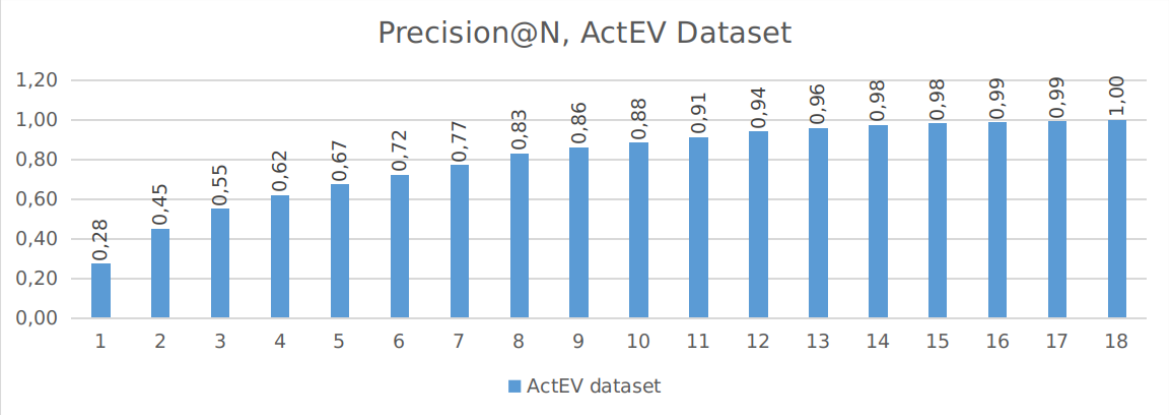

Figure 2: Precision@N, using validation data of ActEV dataset

2.3 Activity detection module

After the training and evaluation processes of the module, we were focused on developing approaches to detect the required activities in extended videos. We have inference the activity recognition module sequentially for all videos of the validation set that provided by ActEV. Specifically, for each video of the validation set, we have created a prediction vector for each activity across the total duration of the video. A graphical representation of the aforementioned technique is presented in figure 3. Each point

of x axis of figure 3 declares the probability of the prediction during the evaluation of 16 sequential frames using the activity recognition module. This 16-frame-based representation was produced due to the 3D-ResNet-50 architecture that was implemented and process 16 frames simultaneously while it returns one prediction for each activity.

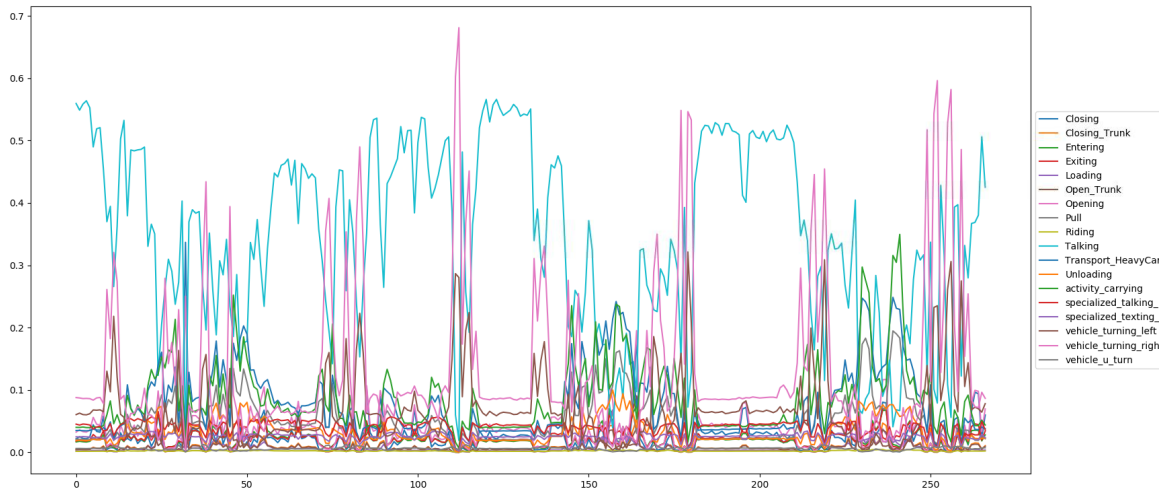

Figure 3: Graphical representation of predicted activities during time(frames) of a video from the validation set.

In order to identify an activity-based threshold, we have implemented an encoder-decoder neural network. Specifically, we have developed a simple architecture which consists of 6 encoded layers and another 6 comprises the decoder. More specifically, the architecture of the deployed encoder-decoder is presented below:

- Encoder:
  - Linear(in=1849, out=462, ReLU)
  - Linear(in=462, out=64, ReLU)
  - Linear(in=64, out=8, ReLU)
  - Linear(in=8, out=4)
- Decoder:
  - Linear(in=4, out=8, ReLU)
  - Linear(in=8, out=64, ReLU)
  - Linear(in=64, out=462, ReLU)
  - Linear(in=462, out=1849, Tanh)

The training process of the above network was carried out using the validation data of ActEV. Specifically, we utilized as input the real-valued predicted vector for one activity of all videos and the target was the binary vector (ground truth) that we have generated using the validation data. Due to the fact that the videos have unequal number of frames, we have declared the largest video length as input/output and zero-padded both input and output vectors when smaller videos were processed. It also should be noted that this process is applied using the test data of the TRECVID [11] before uploading our prediction.

2.4 Submitted runs

For the evaluation requirements, the predictions using the provided test set have been uploaded four times with different configurations for each attempt.

1. Threshold was set to 0.15 for all activities, sampling set to 4 frames.
2. Threshold was set to 0.15 for all activities, sampling set to 1 frame.
3. Threshold was set to 0.30 for all activities, sampling set to 4 frames.
4. Dynamic threshold value, sampling was set to 4 frames.

Initial experiments were examining the impact of the sampling factor using as a threshold a number set equal to 0.15. The rest two are examine the impact of the dynamic threshold approach in contrast to use the same number of threshold for all predicted activities.

2.5 Experimental results

Regarding the first two experiments, the sampling factor equal to 4 in contrast to 1 has been proved more consistent with the problem as can be depicted in Figure 4 (left) and in Figure 4 (right) correspondingly.

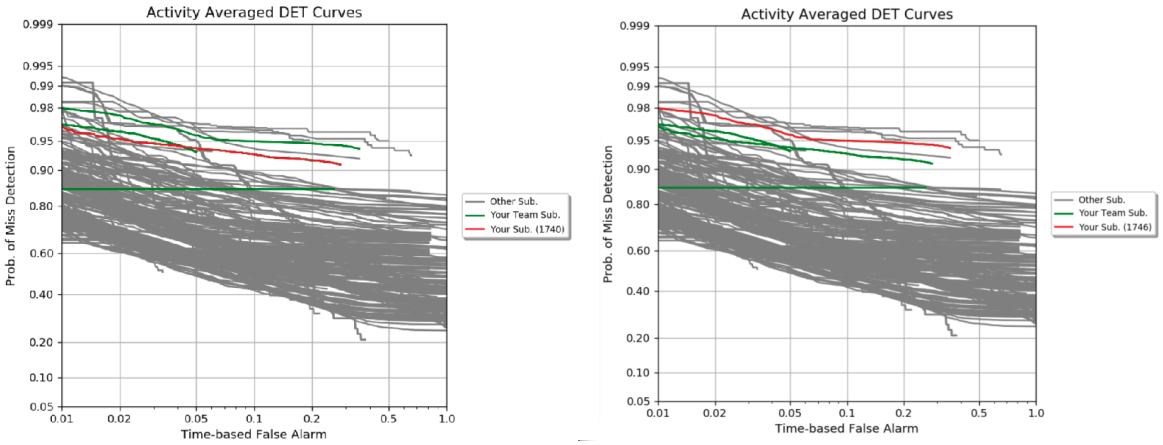

Figure 4: TRECVID submissions 1 and 2, on the left, depicted the Time-based False Alarm using a sampling number equal to 4 and on the right when the sampling factor is equal to 1.

From a detailed examination of the two first experiments, the impact of the sampling factor is defined. Specifically, sampling every 4 frames give a small advantage in contrast to sampling every 1 frame. We consider that this factor is not responsible for the low performance of the proposed framework. Hence, it is not a field for further exploration since it is common to apply a number equal to 4 or 8 sampling frames in the recent bibliography.

Regarding the evaluation of the proposed approach, a comparison between a single value as a threshold and dynamic value was examined with the sampling factor set equal to 4. The results of both experiments 3 and 4, single-threshold and class-based threshold respectively, are presented in Figure 5. It can be observed that the proposed approach seems to improve the overall detection but unfortunately, the existing results remain insufficient.

The analysis of the quantitative and quantitative results proves that the presented method produces inadequate results for detecting activities in extended videos in contrast to the recognition of activities in sort clips. From our perspective and analysis, the processing of the entire frames is the main purpose of the resulted outcomes and not only the region of the image that the action is performed. For this objective, we have investigated these predictions using the validation data. We observed that in many cases the performed actions were located in a small region at the bottom corner of the entire captured video. This results in adding severe noise ratios to our activity detection and recognition framework.

Nonetheless, it should be considered that the presented framework has been developed to predict and detect actions receiving as input the entire visual content.

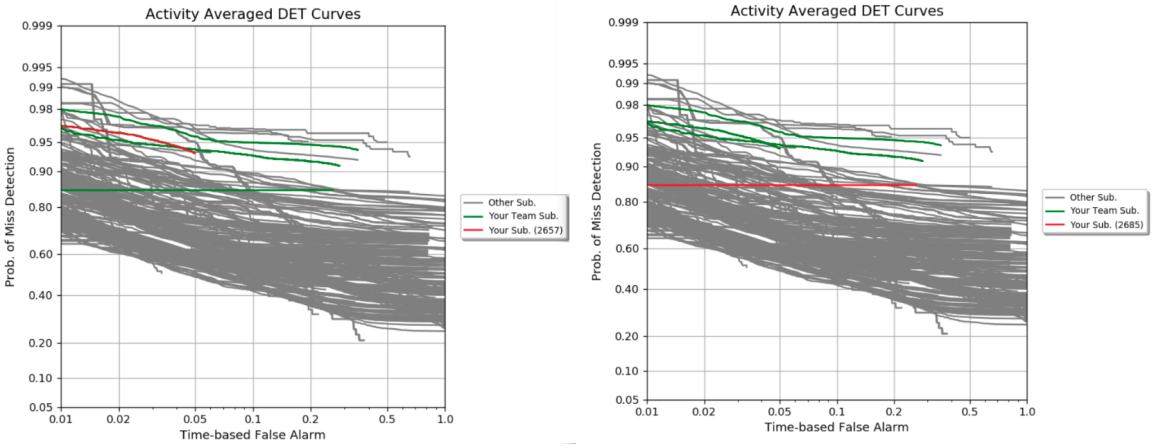

Figure 5: TRECVID submissions 3 and 4, on the left depicted the Time-based False Alarm using a single-threshold and on the right when diffident threshold values for each activity were used.

3 Conclusions

In this paper, we reported the ITI-CERTH framework for the TRECVID 2019 evaluation [12]. ITI-CERTH participated in activity detection in extended videos (ActEV) task in order to evaluate new techniques and algorithms. Specifically, a method based on deep neural networks was proposed. More specifically, a method that on the first step of the process classifies activities of short clips while in the second step, the framework is trained to identify a dynamic threshold value for each activity in order to be applied in extended videos. The results are not still promising but we plan to re-design and re-evaluate the proposed architecture in next challenges. Our future plans include an architecture similar to the proposed one, but with respect to the objects (vehicle, person) that describe each action in order to take the advantage of this information modality and restrict the processing of the whole visual content.

4 Acknowledgements

This work was partially supported by the European Commission under contracts H2020-779962 V4Design, H2020-700475 beAWARE, H2020-740593 ROBORDER, H2020-786731 CONNEXIONs

References

[1] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] A. Moumtzidou, N. Gkalelis, and P. Sidiropoulos et al. ITI-CERTH participation to TRECVID 2012. In *TRECVID 2012 Workshop*, Gaithersburg, MD, USA, 2012.

[3] F. Markatopoulou, A. Moumtzidou, and C. Tzelepis et al. ITI-CERTH participation to TRECVID 2013. In *TRECVID 2013 Workshop*, Gaithersburg, MD, USA, 2013.

[4] N. Gkalelis, F. Markatopoulou, and A. Moumtzidou et al. ITI-CERTH participation to TRECVID 2014. In *TRECVID 2014 Workshop*, Gaithersburg, MD, USA, 2014.

[5] F. Markatopoulou, A. Ioannidou, and C. Tzelepis et al. ITI-CERTH participation to TRECVID 2015. In *TRECVID 2015 Workshop*, Gaithersburg, MD, USA, 2015.

[6] F. Markatopoulou, A. Moumtzidou, and D. Galanopoulos et al. ITI-CERTH participation in TRECVID 2016. In *TRECVID 2016 Workshop*, Gaithersburg, MD, USA, 2016.

[7] F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, and K. Avgerinakis et al. ITI-CERTH participation in TRECVID 2017. In *TRECVID 2017 Workshop*. NIST, USA, 2017.

[8] Konstantinos Avgerinakis, Anastasia Moumtzidou, Damianos Galanopoulos, Georgios Orfanidis, Stelios Andreadis, Foteini Markatopoulou, Elissavet Batziou, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, et al. Iti-certh participation in trecvid 2018. *International Journal of Multimedia Information Retrieval*, 2018.

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[11] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quénot. 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.

[12] G. Awad, A. Butt, K. Curtis, J. Fiscus, et al. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.