ITI-CERTH participation in ActEV and AVS Tracks of TRECVID 2022

Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

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

Abstract

This report presents the overview of the runs related to Ad-hoc Video Search (AVS) and Activities in Extended Video (ActEV) tasks on behalf of the ITI-CERTH team. Our participation in the AVS task is based on a cross-modal deep network architecture utilizing several textual and visual features. As part of the retrieval stage, a dual-softmax approach is utilized to revise the calculated text-video similarities. For the ActEV task, we adapt our framework to fit the new dataset and overcome the challenges of detecting and recognizing activities in a multi-label manner while experimenting with two separate activity classifiers.

1 Introduction

In this work, the work carried out in the context of TRECVID 2022 by the ITI-CERTH[1] team in the area of video analysis, retrieval and understanding is presented. ITI-CERTH has participated in TRECVID [1] for many years as it is one of the most popular video understanding challenges. Especially, ITI-CERTH has participated in Search and Semantic Indexing (SIN) tasks under the research network COST292 (TRECVID 2006-2008) and the MESH and K-SPACE (TRECVID 2007-2008) EU-Funded research projects, correspondingly. From 2009 to 2015 [2, 3, 4, 5, 6, 7, 8] ITI-CERTH team has participated as a stand-alone organization in a significant number of tasks including but not limited to SIN, KIS, INS, and MED. In both 2016 [9] and 2017 [10], ITI-CERTH participated in the AVS, MED, INS and SED tasks. In 2018 [11], ITI-CERTH participated in the AVS, INS and ActEV; in 2019 [12], the participation was limited to the ActEV task. In 2020 [13] ITI-CERTH participated in the AVS, DSDI and ActEV tasks. Lastly, in 2021 [14] ITI-CERTH participated in the AVS and ActEV tasks. Considering the submissions mentioned above, we aim to evaluate improved algorithms and systems. This year, ITI-CERTH participated again in AVS and ActEV tasks. The following sections will present the employed algorithms and the evaluation of the runs during the AVS and ActEV tasks, respectively.

2 Ad-hoc Video Search

The TRECVID 2022 [15] Ad-hoc Video Search (AVS) task aims to develop a system for retrieving a ranked list of 1000 video shots for each ad-hoc textual query, ranked from the most relevant to the least relevant shot for the query. Firstly, we utilize a new cross-modal network that combines different textual and visual features and develops multiple joint latent feature spaces. Secondly, we examine a dual-softmax operation for revising text-video similarities using this year's queries or queries from previous years.

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

2.1 Approach

In our AVS 2022 participation, we utilize the $T \times V$ cross-modal network presented in [16] as our baseline network. This network utilizes multiple textual and visual features along with multiple textual encoders to eventually build multiple cross-modal joint latent feature spaces.

Our network consists of two key sub-networks, one for the textual and one for the visual stream. The textual sub-network inputs a free-text query s and vectorizes it into textual features. These features are used as input in a set of K textual encoders that encode the input sentence. Each of these encoders can be either a trainable network or simply an identity function forwarding its input. Similarly to the textual one, the visual sub-network inputs a video shot v consisting of a sequence of keyframes and we use L trained DNNs to extract the initial frame representations. To obtain video-shot level representations we follow the mean-pooling strategy. Subsequently, we create all the possible textual encodings-visual feature pairs and a joint embedding space is created for each pair, using to this end two fully connected layers. Thus, $K \times L$ different joint spaces are created. The objective of our network is to learn a similarity function $sim(s, v)$ that will consider every individual similarity in each joint latent space utilizing multi-loss-based training.

The second aspect of our study is to study a query-video similarities revision approach based on a dual softmax operation as presented in [16]. At the retrieval stage, and for a given query s, we calculate the similarities with the videos from the evaluation dataset, resulting in a vector $\mathbf{y}(s) = [sim(s, v_1), sim(s, v_2), \ldots, sim(s, v_D)]^T$, where D is the number of evaluation videos. To revise these similarities, we utilize a set of background textual queries (queries that are individual from the examined one) and calculate their similarities with the available videos within the dataset resulting in a similarity matrix $\mathbf{X} \in \mathcal{R}^{C \times D}$, where C is the number of background queries. A matrix $\mathbf{Z}(s) = concat(\mathbf{y}(s); \mathbf{X})$ is constructed, and a dual softmax operation revises the similarities as follows:

$$\mathbf{Z}^*(s) = Softmax(\mathbf{Z}(s),\ dim = 0) \odot Softmax(\mathbf{Z}(s),\ dim = 1)$$

where $\odot$ denotes the Hadamard product.

2.2 Submission

Our network is trained using a combination of four other large-scale video captioning datasets: MSR-VTTT [17], TGIF [18], ActivityNet [19] and Vatex [20]. The V3C2 [21] dataset is utilized to evaluate the networks' performance. The evaluation measure we use is the mean extended inferred average precision (MxinfAP). As initial textual features, we utilize four models: i) Bag-of-Words (bow), ii) Word2Vec model [22] iii) Bert [23] and iv) Clip model [24]. Also, we utilize two textual encoders that input the textual features and encode the text further; i) the textual sub-network (ATT) presented in [25] and ii) a Clip encoder that simply feedforwards the corresponding features through an identity layer. As video feature extractors, we use three trained networks: i) a ResNet-152 [26] network, trained on the ImageNet-11k dataset, ii) a ResNeXt-101 network, pre-trained by weakly supervised learning on web images followed and fine-tuned on ImageNet [27], and iii) the ViT-B/32 Clip model [24].

Similarly to previous works [28] [25] where the combination of multiple models leads to improved performance, we utilize different model configurations to train multiple models using three learning rates and two optimizers (i.e., Adam and RMSprop).

This year we submitted two runs on the AVS 2022 main task and two additional runs for the AVS progress subtask. Overall, we evaluate our methods on 50 different ad-hoc queries (30 from the main task and 20 from the progress subtask). The submitted runs are briefly described below:

- ITI_CERTH.22_run_1: The $T \times V$ model, using two textual encoders and three visual features. Late fusion of six different trained models derived from six different model configurations. And finally, query-video shot similarities revision through the dual softmax operation using all AVS 2022 queries as background queries.

- ITI_CERTH.22_run_2: Similar to run 1, but as background queries, the AVS 2019, 2020, and 2021 queries are used.

2.3 Experimental Results

Table 1 summarizes the evaluation results of our runs for the main AVS task. The ITI_CERTH.22 run_1 where the AVS 2022 queries were used as background queries at the similarity revision step slightly outperforms the ITI_CERTH.22 run_2 where we do not utilize this year's queries and the retrieval performance does not rely on a priori knowledge of all queries that are being evaluated. The results indicate that the utilization of all query knowledge is beneficial for the system's performance. But on the other hand, having access to such knowledge is an assumption we made and is not always applicable, especially regarding real-world applications. So, the ITI_CERTH.22 run_2 results show that knowing all evaluated queries with a small trade-off regarding the overall performance is not a prerequisite.

Figure 1 illustrates the performance of all submitted runs at the AVS 2022 competition. Our runs achieved 9th and 10th place across all submitted runs and 3rd place among all team participations.

Table 1: Mean Extended Inferred Average Precision (MXinfAP) for all submitted runs for the fully-automatic AVS task.

Run id:	Main
ITI_CERTH.22 run_1	**0.210**
ITI_CERTH.22 run_2	0.206

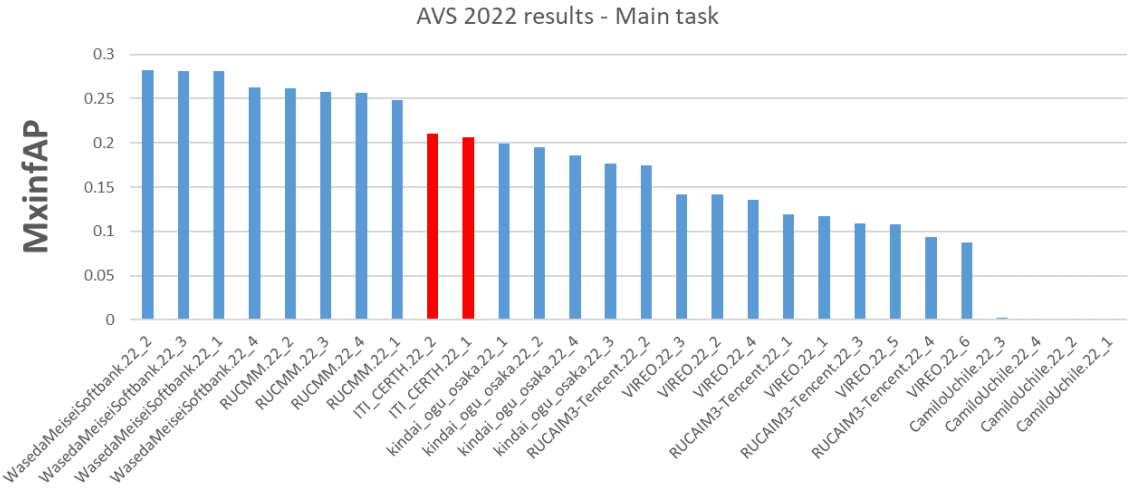

Figure 1: AVS 2022 ranking list of all submitted runs regarding the main task in MXinfAP terms. Red bars indicate our submitted runs.

3 Activities in Extended Video

In activity recognition systems, multimedia resources taken from cameras in indoor and outdoor environments are analyzed to identify the activity instances of the depicted objects. Due to its practical relevance, the activity recognition field of research has received a great deal of attention considering the numerous related applications, i.e. surveillance and traffic control systems. Activity recognition in those systems often deals with plenty of challenges, such as the untrimmed nature of the video resources, the multiple activities executed by the same object, and the interaction among objects. The challenges mentioned above and the need for a real-time response make human processing and analysis difficult. Thereby, automated methods are needed to accomplish the task of activity recognition.

Towards this goal, the Activities in Extended Videos challenge (ActEV) encourages the research of real-time activity detection methods in surveillance scenarios. In our work, we address the problem of activity recognition through a three-step pipeline using: object detection, object tracking, and activity classification. In terms of object detection, Yolov4 [29] identifies persons and vehicles in video resources. Then, the DeepSort [30] tracking method is employed to track the detected objects, which are then used as spatiotemporal activity proposals from the activity classifiers. Finally, two activity classifiers are developed based on the 3D-ResNet [31] model; one is dedicated to Person-Related (PR) activities, and the other to Vehicle-Related (VR) as well as Person-Vehicle-Related (PVR) activities. The MEVA dataset [32] was used to train and validate the activity classifiers using the official Kitware annotations[1], aiming at the 20 activity classes depicted in Table 2.

3.1 Approach

The task of activity recognition and localization is applied to a set of videos $V = \{v_i\}$ in order to identify the set of activities $A = \{a_i\}$ for all the Objects of Interest (OoI). An activity a_i is described by a type t_i according to a set of predefined classes and a temporal location $l = (t_{start}, t_{end})$ that indicates the start and end of it within the video it occurs.

Table 2: Activity classes in ActEV challenge 2022.

Activity Classes	
PR Classes	**VR & PVR Classes**
person_reads_document	person_closes_vehicle_door
person_enters_scene_through_structure	person_enters_vehicle
person_exits_scene_through_structure	person_exits_vehicle
person_stands_up	person_opens_vehicle_door
person_sits_down	vehicle_starts
person_talks_to_person	vehicle_stops
person_picks_up_object	vehicle_turns_left
person_puts_down_object	vehicle_turns_right
person_opens_facility_door	
person_texts_on_phone	
person_interacts_with_laptop	
person_transfers_object	

Considering the challenges of detected activities in surveillance videos, this work focuses on effectively detecting and tracking the OoI, persons and vehicles, and naming their activities. Unlike our previous submissions [13, 14] in the activity detection task of the ActEV challenge, this year we have introduced three major improvements: (1) we dismissed the step of spatio-temporal tubelets creation, named Extended Activity Bounding Boxes (EABBoxes) [33] by keeping only the information of the Bounding Boxes (BBoxes) for every object to exclude redundant information, (2) we utilize two activity classifiers instead of one by grouping the given activity classes into two groups to enhance learning, and (3) we incorporate a deep learning model in the object tracking task by replacing the Euclidean distance algorithm. In the following subsections, further details of the pipeline are provided.

3.1.1 Object Detection

The task of object detection is needed in order to extract frame-wisely objects, which are considered performers of potential activities. Given its fast and accurate performance, YOLOv4 [29] is incorporated as an object detector in our pipeline. More specifically, YOLOv4 [29] is an advanced version of YOLOv3 [34], combining faster operating speed along with greater accuracy, reaching 43.5% Average Precision (AP) for the Microsoft COCO [35] dataset at a real-time speed of approximately 65 Frames Per Second (FPS) on Tesla V100 GPU. For the experiments, we employed a pre-trained YOLOv4 [29] model that was trained on the Microsoft COCO [35] dataset to detect items included into two categories: "person" and vehicle ("car", "bus", and "truck").

[1]https://gitlab.kitware.com/meva/meva-data-repo/-/tree/master/annotation/DIVA-phase-2/MEVA

3.1.2 Object Tracking

Given a set of object detections for every video frame, the task of object tracking is to link those detections over time, yielding object trajectories. The Deep Simple Online Realtime Tracking (DeepSORT) [30] is used as a tracking algorithm, which assigns a unique ID to every object that tracks within a video. DeepSORT [30] is an extension of Simple Online Realtime Tracking (SORT) [36] that shows greater performance in terms of ID switches and occlusions. In order to track objects successfully, DeepSORT [30] uses appearance descriptors apart from extracting only velocity and motion cues from the objects. To fill the potential temporal gaps within an object's trajectory, the interpolation algorithm was used to give the object's coordinates for the missing frames. The output of this step is a set of tracked objects $O = \{o_i\}$, and every one of them is characterized by the bounding boxes for all the video frames that it was tracked $o_i = \{(x_{left}, y_{top}, width, height)_{t_1}, ...\}$.

3.1.3 Activity Recognition

The final step of our pipeline is activity classification. In this context, the 3D-ResNet [31] is employed to label the tracked objects. 3D-ResNet [31] involves a deep learning architecture and effectively performs on spatiotemporal data due to its 3D convolutional layers. Its architecture consists of four sequential bottleneck blocks, where each block includes three 3D-convolution layers (with variant kernel sizes), batch normalization, and ReLU activation layers. More specifically, we initialized the model of 3D-ResNet [31] using the Kinetics [37] dataset's pre-trained weights and then fine-tuned it in a multi-label manner using the MEVA dataset [32]. Regarding the input, the 3D-ResNet model performs in a 16-length frame batch.

In order to ensure greater learning ability, two separate activity classifiers are used, trained in two different sets of activity classes as depicted in Table 2. One set of classes is dedicated to PR activities only, while the other involves VR classes as well as PVR classes. The PVR classes demonstrate the interaction between people and vehicles, and provide spatial data for both the person and vehicle involved in an activity. In most cases, the bounding boxes of persons overlap with the ones of vehicles, and thereby we make the assumption that vehicles' bounding boxes can capture, to a great extent, the interaction with persons. Therefore, to exclude redundant information, we keep only the vehicles' spatial information for the PVR classes and concatenate them with the VR ones.

3.1.4 Activities refinement

The activity classification step assigns scored labels to every batch (16) of frames of a detected object's trajectory. In most cases, objects perform more than one activity simultaneously or for different time intervals during their trajectories. For this reason, some post-processing steps are needed in order to generate more accurate activity proposals and reduce the number of false alarms. The first step of the refinement algorithm is to threshold the scored activity labels. Thus, we set a high threshold T_{hight} and a low threshold T_{low}. The T_{low} sets a strict limit for the activity proposals, as those with scores lower than this value are less possible to happen, which implies their exclusion. The T_{hight} also sets a lower limit for an activity type but gives a chance for lower-scored frame batches (at most 3 sequential batches, namely 48 frames) to be included in a created activity proposal that mostly consists of high-scored frame batches for a specific activity type. This serves the case that the activity classifier can incorrectly give lower scores to activity instances due to misleading factors such as occlusions, background noise, etc. Another refinement step is to add the NMS in order to deduplicate the activity proposals.

Our last refinement rule is related to the semantic relation between the activity labels. As an attempt to reduce the number of false positives, this rule sets four groups of activity labels that can not characterize the same activity proposal. Alternatively, if an activity classifier assigns high scores to labels belonging to the same group, only the highest-scored label will be kept. For example, in case that there were assigned the labels "vehicle_starts" and "vehicle_stops" to the same activity proposal, only the one with the highest score value will be kept as semantically it is impossible for a vehicle to perform both activities simultaneously. In Table 3, there are mutually exclusive groups of activity labels. Groups 1 and 2 refer to activities for VR and PVR classes, while groups 3 and 4 to PR classes.

Table 3: Mutually exclusive groups of activity labels.

Activity Labels Groups	
Group 1	vehicle_starts, vehicle_stops, person_closes_vehicle_door, person_opens_vehicle_door, person_enters_vehicle, person_exits_vehicle
Group 2	vehicle_turns_left, vehicle_turns_right, person_closes_vehicle_door, person_opens_vehicle_door, person_enters_vehicle, person_exits_vehicle
Group 3	person_stands_up, person_sits_down, person_enters_scene_through_structure, person_exits_scene_through_structure
Group 4	person_picks_up_object, person_puts_down_object, person_reads_document

3.2 Submissions

In this section, we present the two submitted systems, as depicted in Figure 2 and in Table 4:

- baseline: This system was deployed using YOLOv4 [29] for object detection and the DeepSORT [30] algorithm for tracking. The MEVA dataset [32] was used to train and validate the two activity classifiers, splitting the classes according to Table 2. The refinement was achieved by adjusting the T_{hight} for the final scores to 40% (T_{low}=0%).
- M4DSYS_1: This system is an extension of the baseline system, due to the differences in refinement steps. According to experiments, the values of T_{hight} and T_{low} are set to 65% and 10% respectively. The NMS and the semantic rules of mutually exclusive groups of activity classes were performed subsequently.

Table 4: The evaluation results for MEVA [32] validation set and ActEV challenge test set.

	Validation Set		**Test Set**	
	Baseline	M4DSYS_1	Baseline	M4DSYS_1
p_miss@0.1rfa	0.9787	**0.9513**	0.9823	**0.9603**
nAUDC@0.2rfa	0.9802	**0.9528**	0.9819	**0.9639**
Correct Detections	**3142**	1233	-	-
False Detections	198059	**23269**	-	-
Missed Detections	**2670**	4579	-	-
Number of Activities	201201	24502	144071	23572

3.3 Experimental Results

In this subsection, further discussion about the performance of the submitted systems is reported. In Table 4, there are results for two different setups as described in the Submissions section - baseline and M4DSYS_1 - for both validation and test sets. The results for validation sets came out from evaluations that run locally and the ones for test sets gained from the public leaderboard[2] of the ActEV challenge. Both sets were evaluated using the "SRL_AD_V1" scoring protocol. For the test set, we could not fill the numbers for correct, false and missed detections, as this information is not available to us.

According to Table 4, the results for the M4DSYS_1 setup are slightly better than those of the baseline, which is due to the changes in the refinement steps that lead to the reduction in the number of false detections. Consequently, there is also a reduction in the number of correct detections and a rise in missed detections, which resulted from the exclusion of a high number of activities due to stricter control. In spite of the reduction in correct detections, the values of p_miss@0.1rfa and nAUDC@0.2rfa metrics suggest better results were influenced by fewer false detections in the M4DSYS_1 system. One

[2]https://actev.nist.gov/SRL#tab_leaderboard

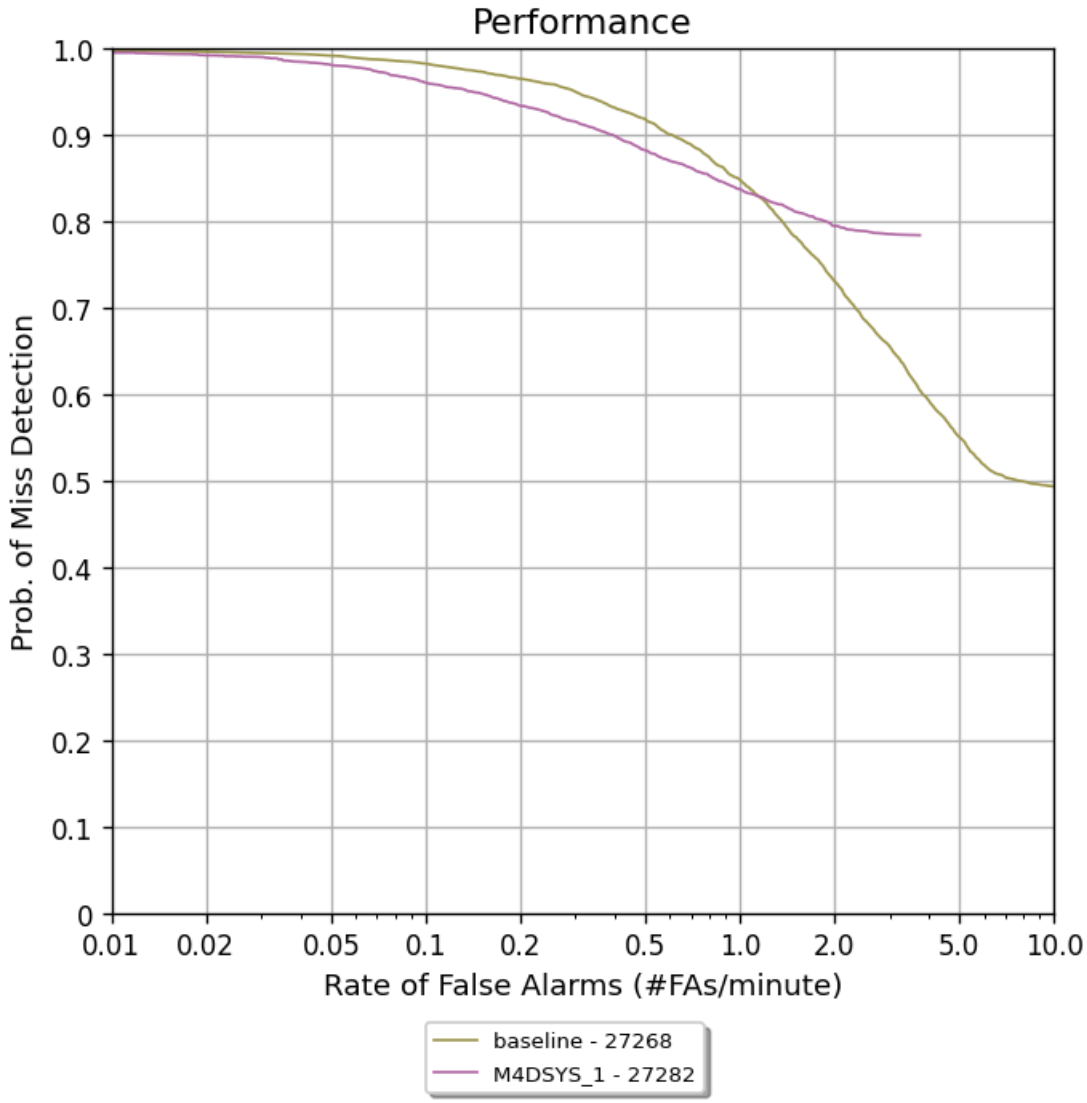

Figure 2: CERTH-ITI systems' reported performance in the context of the ActEV 2022 challenge.

reason for the high values in evaluation metrics for both system setups is that our method assigns activity labels to the whole trajectory of a tracked object, while in the annotation set only parts of an object trajectory are annotated. Taking as input the whole trajectory of an object, activity classifiers are misled, yielding non-negligible scores even for part of the objects' trajectory that should not be labelled.

4 Conclusions

In this paper, the evaluation of ITI-CERTH during the TRECVID 2022 challenge [15] is reported. ITI-CERTH this year participated by developing new techniques and algorithms in the context of AVS and ActEV tasks. Regarding the AVS task, we used a cross-modal network for text-to-video retrieval to combine textual and visual features and learn multiple joint feature spaces. Moreover, we utilize background queries and a dual-softmax operation to revise query-video similarities. We showed that using individual queries from the examined ones leads to compatible results while serving real-case scenarios. Regarding the ActEV task, a three-step pipeline was deployed in order to effectively detect objects, track them and recognize their activities in a multi-label manner. The classification of the

detected activities is performed spatio-temporally using two separate classifiers; one for the person-related activities and one for the vehicle-related and person-vehicle interaction activities. Though the results are not expected, some aspects of the process seem promising.

5 Acknowledgements

This work was partially supported by the projects INFINITY (H2020-883293), CRiTERIA (H2020-101021866), CREST (H2020-833464) and ODYSSEUS (H2020-101021857), funded by the European Commission.

References

[1] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] A. Moumtzidou, A. Dimou, and P. King et al. ITI-CERTH participation to TRECVID 2009 HLFE and Search. In *Proc. TRECVID 2009 Workshop*, pages 665–668. 7th TRECVID Workshop, Gaithersburg, USA, November 2009.

[3] A. Moumtzidou, A. Dimou, and N. Gkalelis et al. ITI-CERTH participation to TRECVID 2010. In *Proc. TRECVID 2010 Workshop*. 8th TRECVID Workshop, Gaithersburg, MD, USA, November 2010.

[4] A. Moumtzidou, P. Sidiropoulos, and S. Vrochidis et al. ITI-CERTH participation to TRECVID 2011. In *Proc. TRECVID 2011 Workshop*. 9th TRECVID Workshop, Gaithersburg, MD, USA, December 2011.

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

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

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

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

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

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

[11] 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.

[12] Konstantinos Gkountakos, Konstantinos Ioannidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Iti-certh participation in trecvid 2019. In *TRECVID 2019 Workshop*, 2019.

[13] Konstantinos Gkountakos, Damianos Galanopoulos, Marios Mpakratsas, Despoina Touska, Anastasia Moumtzidou, Konstantinos Ioannidis, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. Iti-certh participation in trecvid 2020. In *TRECVID 2020 Workshop*, Gaithersburg, MD, USA, 2020.

[14] Konstantinos Gkountakos, Damianos Galanopoulos, Despoina Touska, Konstantinos Ioannidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. Iti-certh participation in actev and avs tracks of trecvid 2021. In *TRECVID 2021 Workshop*, Gaithersburg, MD, USA, 2021.

[15] George Awad, Keith Curtis, Asad A. Butt, et al. An overview on the evaluated video retrieval tasks at trecvid 2022. In *Proceedings of TRECVID 2022*. NIST, USA, 2022.

[16] Damianos Galanopoulos and Vasileios Mezaris. Are all combinations equal? combining textual and visual features with multiple space learning for text-based video retrieval. In *ECCVW*. Springer, October 2022.

[17] J. Xu, T. Mei, et al. MSR-VTT: A large video description dataset for bridging video and language. In *Proc. of IEEE CVPR 2016*, pages 5288–5296, 2016.

[18] Y. Li, Y. Song, L. Cao, J. Tetreault, et al. TGIF: A new dataset and benchmark on animated gif description. In *Proc. of IEEE CVPR 2016*, 2016.

[19] F. Caba Heilbron et al. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proc. of IEEE CVPR 2015*, pages 961–970, 2015.

[20] X. Wang et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proc. of IEEE/CVF ICCV 2019*, pages 4581–4591, 2019.

[21] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt. V3C–a research video collection. In *Proc. of MMM 2019*, pages 349–360. Springer, 2019.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, Workshop Track Proceedings*, ICLR '13, 2013.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, , et al. Learning transferable visual models from natural language supervision. In *Proc. of the 38th Int. Conf. on Machine Learning (ICML)*, 2021.

[25] D. Galanopoulos and V. Mezaris. Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In *Proc. of ACM ICMR 2020*, 2020.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[28] D. Galanopoulos and V. Mezaris. Hard-negatives or Non-negatives? A hard-negative selection strategy for cross-modal retrieval using the improved marginal ranking loss. In *Proc. of IEEE/CVF ICCVW 2021*, 2021.

[29] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[30] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[31] 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.

[32] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1060–1068, 2021.

[33] Konstantinos Gkountakos, Despoina Touska, Konstantinos Ioannidis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. Spatio-temporal activity detection and recognition in untrimmed surveillance videos. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 451–455, 2021.

[34] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[36] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[37] 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.