2020 Sequestered Data Evaluation for Known Activities in Extended Video: Summary and Results

Afzal Godil[1], Yooyoung Lee[1], Jon Fiscus[1], Andrew Delgado[1], Eliot Godard[3,1], Baptiste Chocot[3,1], Lukas Diduch[2], Jim Golden[1], Jesse Zhang[1]
[1]National Institute of Standards and Technology, USA
[2]Dakota Consulting Inc., USA
[3]Guest researcher

Abstract

This paper presents a summary and results for the ActEV'20 SDL (Activities in Extended Video Sequestered Data Leaderboard) challenge that was held under the CVPR'20 ActivityNet workshop [38]. The primary goal of the challenge was to provide an impetus for advancing research and capabilities in the field of human activity detection in untrimmed multi-camera videos. Advancements in activity detection will help with a wide range of public safety applications. The challenge was administered by the National Institute of Standards and Technology (NIST), where anyone could submit their system which run on sequestered data with the resulting score posted to a public leaderboard. Ten teams submitted their systems for the ActEV'20 SDL competition on the Multiview Extended Video with Activities (MEVA) test set with 37 target activities. The performance metric for the leaderboard ranking is the partial, normalized Area Under the Detection Error Tradeoff (DET) curve ($nAUDC$). The top rank on activity detection was by UCF at 37%, followed by CMU at 39% and OPPO at 41%.

1. Introduction

In recent years, large amount of video data has been collected from multi-camera networks for different purposes. With multi-camera networks, activities in a wide area can be detected, and robustness and reliability of activity detection can be improved by fusing data from multiple camera views. However, there has not been a commensurate increase in the usage of intelligent video analytics for real-time alerting or triaging of video in multi-camera networks. Operators of camera networks are typically overwhelmed with the volume of video they must monitor from multiple cameras, and cannot afford to view or analyze even a small fraction of their video data. Automated methods that identify and localize activities in extended video from multiple camera views are necessary to alleviate the current manual process of monitoring by human operators and provide the capability to alert and triage video that can scale with the growth of multi-camera sensor networks.

Recently, larger visual datasets and deep learning have revolutionized the computer vision field, contributing to significant advances in the performance of activity detection and classification. However, a significant focus of activity detection has been on near field and social media video, while application to wide field-of-view public safety video has not yielded satisfactory results. Particular challenges for public safety video include the presence of long time periods with no activities (requiring detection of temporal regions with activities not just classification), multiple viewpoints, temporal coincidence of multiple simultaneous activities in different spatial regions of the video, and occurrence of many activities in mid and far field views from the video sensor resulting in low resolution.

To understand current state-of-the-art and to promote activity detection technologies, the Activities in Extended Video (ActEV) evaluation series is being conducted. The goal of the ActEV evaluation is to facilitate the development of video analytic technologies that can automatically detect target activities and to reduce the detection error rate. In 2018, the National Institute of Standards and Technology (NIST) developed the ActEV evaluation series [1, 39, 2, 21] to support the metrology needs of the Intelligence Advanced Research Projects Activity (IARPA) Deep Intermodal Video Analytics (DIVA) Program [10]. The Multiview Video with Activities (MEVA) dataset [17, 16] and the Video Retrieval and Analysis Tool (VIRAT) dataset [18] were used in the ActEV Sequestered Data Leaderboard (SDL) competition [27] and the ActEV TRECVID compe-

titions [28] and are far more closely aligned with real-world public safety ground video analytics than alternate activity detection datasets. The ActEV'19 SDL competition started since August 2019, allowing for the instantiation of the competition infrastructure and the development and compilation of related resources. Since the start of ActEV'20 SDL competition, we have provided 27 hours of training data annotations to make it more accessible to the broader computer vision community. The ActEV'20 SDL results reported in this paper are only based on the MEVA Test 3 dataset [17] and ran from March 1st to May 17th, 2020. The MEVA data is much larger and has high-resolution compared to VIRAT dataset, and contains hundreds of hours of video with hundreds of instances of each activity, in indoor and outdoor. The data collected by a multi-camera IP network in a heterogeneous environment, in day and night scenes, from multiple viewpoints and staging scenarios that are close to real-world. The site for the data collection was a group of buildings, the interior of the buildings and grounds and roads surrounding the area. The videos are from both EO (Electro-Optical) and IR (Infrared) sensors.

Major contributions of this paper are threefold. First, we describe the MEVA dataset that was used for the challenge, which we hope will facilitate the development of more advanced solutions for real-time activity detection in public safety video and provide an impetus for more research in the field of computer vision. Second, we describe and implement a set of performance evaluation measures for activity detection. Third, we present the results and research finding for the ActEV'20 SDL challenge. In addition, we have created a OpenStack [29] based private cluster to run the ActEV SDL sequestered evaluation by Command Line Interface (CLI) submission of systems. We also have developed a website to submit the systems, to run, score, and display the results on the leaderboard.

We hope ActEV SDL challenge and the associated MEVA dataset will facilitate the development of advanced solutions for real-time activity detection in public safety video and provide an impetus for more research in the field of activity detection. Advancements in activity detection will impact a wide range of applications such as public safety, transportation and infrastructure monitoring for both real-time alerting and forensic analysis.

The paper is organized as follows: Section 2 describes related work in activity detection and classification. The datasets are described in Section 3. Sections 4 summarizes the evaluation measures. Finally, in Section 5 we present the results and findings.

2. Related work

In recent years, activity detection and classification in videos has been an active area of research and has spanned various target domains and applications. Over the years the size of video datasets for activity detection, classification and recognition have grown. Early datasets on activity recognition in video, KTH [32] and Weizmann [4], employed actors performing a small set of scripted activities under controlled conditions. The Hollywood Human Actions I (HOHA-1) dataset [20] contains videos with 8 activities from 32 movies and the Hollywood Human Actions II (HOHA-1I) dataset [23] has 12 activities and 10 classes of scenes from 69 movies. The HMDB51 dataset [19] provides 3 train-test splits, each of which consists of 6,766 videos. These segments are labeled with 51 classes of human actions, and each video is only labeled with one class. One of the larger video benchmarks is the Sports-1M [14], with 500 sports related classes of activities and 1 million YouTube videos. The YFCC100M data [37] by Yahoo! consists of 0.8 million videos with raw metadata. The UCF101 dataset [33] contains 13,320 videos and 101 classes and is a collection of unconstrained videos downloaded from YouTube with challenges such as poor lighting, cluttered background and camera motion. Both HOHA-1I and UCF101 datasets support evaluation of spatio-temporal localization in untrimmed videos. The THUMOS [11] ran a series of challenges since 2013. The [11] was collected from YouTube: the training set has over 13,000 temporally trimmed videos from 101 action classes and a validation set of over 2100 temporally untrimmed videos with temporal annotations of actions. Around 3000 background videos do not include any instance of the 101 actions; the test set has over 5600 temporally untrimmed videos.

Organized under TRECVID, Multimedia Event Detection (MED) [13] was one of the major undertakings in video analytics and searching. The MED task was to detect the occurrence of an event within a video clip based on an event kit, which contains a text description of the concept and some example videos. MED 2016 used video data from the YFCC100M dataset [37]. Similarly, the Surveillance Event Detection (SED) [24] evaluation was organized under the TRECVID evaluation and focused on event detection in the multi-camera airport video domain. The evaluation was conducted as part of TRECVID from 2008 to 2017. The i-LIDS dataset [5] was used by the SED evaluation. The development data consisted of the full 100-hour dataset used for the 2008 Event Detection [13] evaluation. The video for the evaluation corpus came from the approximate 50-hour i-LIDS MCTTR dataset. Both datasets were collected in the same airport environment.

In 2018, UCF introduced the UCF-Crime dataset [34] of 128 hours of videos. It consists of 1900 long untrimmed real-world public safety videos, with 13 realistic anomalies (such as fighting, road accident, burglary, robbery), as well as normal activities. This dataset can be used for two tasks: 1) for general anomaly detection considering all anomalies as one; and 2) for recognizing each of 13 anomalous activ-

ities separately. The HiEve dataset for human-centric video analysis was introduced in 2020 by UCF [22] for 56K complex event relates to dense crowds, anomalous individual, or collective behavior.

Some of the widely used multi-camera benchmarks for person and vehicle, tracking and re-identification are the Market-1501 [40], the CityFlow [36] and the DukeMTMC [31].

For instructional video analysis, the COIN dataset [35] was introduced, the dataset contains 11,827 videos of 180 different tasks, covering the daily activities related to vehicles, gadgets and many more. The HowTo100M dataset [25] of 136 million video clips sourced from over a million narrated instructional web videos depicting humans performing and describing over 23k different visual tasks was introduced in 2019.

ActivityNet [6] is a large-scale dataset for recognition of human activities. It consists of 203 activity classes with both trimmed and untrimmed videos. The classes are linked through a taxonomy consisting of parent-child relationships. The ActivityNet Challenge workshop [38] has become an umbrella platform that hosted activity challenges held at Computer Vision and Pattern Recognition (CVPR) 2016 through 2020. For CVPR20, there were seven tasks, three based on the ActivityNet dataset, and the other four based on Kinetics-700 (Trimmed Activity Recognition), AVA (Spatio-temporal Action Localization) and Moments in Time (Trimmed Event Recognition) datasets. The Kinetics dataset [15] contains 400 human action classes, with at least 400 video clips for each action. Each clip lasts around 10 seconds and is taken from a different YouTube video and an extended dataset Kinetics-700 contains 700 action classes. The AVA dataset [9] contains densely annotated 80 atomic visual actions in 430 15-minute video clips, where actions are localized both in space and time, resulting in 1.58M action labels with multiple labels per person occurring frequently in the video. The HACS Temporal Action Localization Challenge 2020 [26] goal was to temporally localize actions in untrimmed videos. The forth task ActEV'20 SDL Challenge was run as a sequestered evaluation with the goal to detect and temporally localize instances of 37 different activities in 140 hours of video. This report presents the results of the ActEV'20 SDL Challenge. The previous challenge ActEV'19 SDL with MEVA Test2 dataset was held under the WACV'20 HADCV workshop [12].

In addition, many open competitions have released the test set to participants to submit self-reported results. These competitions have relied on trust that participants will follow the rules and not inspect the test set to obtain an unfair advantage. There was no mechanism in place to guarantee that participants followed the rules, and it was possible for participants to view the test set and obtain an unfair advantage. The ActEV SDL competition has performed the evaluation exclusively on sequestered data with all evaluation runs performed on hardware hosted by the test and evaluation team, instead of an open competition where the test set is released to participants to submit self-reported results.

3. Datasets

The ActEV SDL competition is based on the the Multi-view Video with Activities (MEVA) dataset [17, 16] (mevadata.org) which was collected and annotated specifically for the development and evaluation of public safety video activity detection capabilities at the Muscatatuck Urban Training Center by Kitware for the DIVA program for IARPA and the broader research community. This dataset contains time-synchronized multi-camera, continuous, long-duration video, often taken at significant stand-off ranges from the activities. Metadata and auxiliary data for the site was provided as is typical for public-safely scenarios where detailed knowledge of the site is available to systems. Provided data included a map and 3D site model of the test area, approximate camera locations for the publicly released video data, and camera models for released sensor video. The dataset was collected with both EO (Electro-Optical) and IR (Infrared) sensors, with over 100 actors performing in various scripted and non-scripted activities in various scenarios. The activities included, person and multi-person activities, person object interaction activities, vehicle activities, and person vehicle interaction activities.

The dataset was captured with off-the-shelf cameras with fields of view, which is both overlapping and non-overlapping, and the videos are captured by 38 cameras for both EO and IR. The spatial resolution of the EO cameras is 1920x1080 or 1920x1072 and the thermal IR cameras is 352x240. All the video cameras have frame rate of 30 frames/second, have a fixed orientation except one, and all are synchronized with the GPS time signal. All the IR cameras are paired with EO cameras having the same position and orientation, and are only outdoor.

The dataset has two main parts: the sequestered test data and MEVA public training and development data.

3.1. Sequestered Test datasets

The MEVA test dataset (aka Known Facility (KF) dataset) used for the ActEV'20 SDL challenge was the MEVA Test3 from March 1st 2020 to May 17th 2020. The videos includes indoor and outdoor scenes, night and day, crowds and individuals, and videos from both EO (Electro-Optical) and IR (Infrared) sensors. The number of target activities in EO videos is 37 and the target activity for IR videos is 34 as the IR cameras are only outdoor. Figure 1 shows the montage of randomly selected videos. Figure 2 shows some of the activities that we are currently using for

Figure 1. Montage of randomly selected video clips.

Figure 2. Montage of randomly selected activities.

ActEV SDL tasks. Table 1 lists the names of the activities; the count is not listed because of the on ongoing evaluations.

MEVA Test3 dataset was the sequestered data used for ActEV'20 SDL challenge. This version had 140 hour of multi-camera videos with 37 activities and support was provided for the use of multi-camera videos. The data set consists of both EO and IR cameras, public cameras (examples of which are in the public data set). The leaderboard presents results on the full 140-hour collection reporting separately by EO and IR data. The activities names for the ActEV'20 SDL evaluation are shown in Table 1. The activities can be broadly classified as, person/multi-person activities, person object interaction, vehicle activities, person facility interaction, and person vehicle interaction.

3.2. MEVA Public dataset

The public MEVA dataset that has been released to-date is approximately 500 GB in size. We provided 27 hours of annotations of the publicly released data for the 37 activities. We have also made available a large corpus of annotations performed by external contributors. The 3D model of the facility, UAV video data, public camera information and the GPS tracks for the actors have been provided.

Table 1. The 37 activity names.

Activity Type	
person_abandons_package	person_picks_up_object
person_closes_facility_door	person_purchases
person_closes_trunk	person_reads_document
person_closes_vehicle_door	person_rides_bicycle
person_embraces_person	person_puts_down_object
person_enters_scene_through_structure	person_sits_down
person_enters_vehicle	person_stands_up
person_exits_scene_through_structure	person_talks_on_phone
person_exits_vehicle	person_texts_on_phone
hand_interacts_with_person	person_steals_object
person_carries_heavy_object	person_unloads_vehicle
person_interacts_with_laptop	vehicle_drops_off_person
person_loads_vehicle	vehicle_picks_up_person
person_transfers_object	vehicle_reverses
person_opens_facility_door	vehicle_starts
person_opens_trunk	vehicle_stops
person_opens_vehicle_door	vehicle_turns_left
person_talks_to_person	vehicle_turns_right
	vehicle_makes_u_turn

4. Measures

4.1. Activity Detection Metrics

The commonly used activity detection metrics are based on using the per-frame mean average precision (mAP) [7] or its calibrated version (cAP) [8] and these metrics mainly apply to the forensic activity detection. The Instantaneous Accuracy (IA) metric [3, 30] was introduced for the evaluation of streaming activity detection.

Based on the requirement of our application, the metrics that we have used are measured by Probability of Missed Detection (P_{miss}), and Time-based False Alarm (T_{fa}) criteria. The primary performance metric is the partial Area Under the DET Curve $nAUDC$ from 0 to a fixed, Time-based False Alarm T_{fa} value a, denoted $nAUDCa$. The partial area under DET curve is computed separately for each activity over all videos in the test collection and then is normalized to the range [0, 1] by dividing by the maximum partial area $nAUDC_a = 0$ is a perfect score. The $nAUDC_a$ is defined as:

$$nAUDC_a = \frac{1}{a}\int_{x=0}^{a} P_{miss}(x)dx, x = T_{fa}$$

where x is integrated over the set of T_{fa} values. The instance-based probability of missed detections P_{miss} is defined as:

$$P_{miss}(x) = \frac{N_{md}(x)}{N_{TrueInstance}}$$

where $N_{md}(x)$ is the number of missed detections at the presence confidence threshold that result in $T_{fa} = x$ (see

the below equation for the details). $N_{TrueInstance}$ is the number of true instances in the sequence of reference.

The time-based false alarm T_{fa} is defined as:

$$T_{fa} = \frac{1}{NR} \sum_{i=1}^{N_{frames}} \max(0, S'_i - R'_i)$$

where N_{frames} is the duration of the video and NR is the non-reference duration; the duration of the video without the target activity occurring. S'_i is the total count of system instances for frame i while R'_i is the total count of reference instances for frame i. The detailed calculation of T_{fa} is illustrated in Figure 3. In S, the first number indicates instance id and the second indicates presence confidence score. For example, $S1(.9)$ represents the instance $S1$ with corresponding confidence score .9. Green arrows indicate aligned instances between R and S).

The non-reference duration (NR) of the video where no target activities occurs is computed by constructing a time signal composed of the complement of the union of the reference instances duration. R is the reference instances and S is the system instances. R' is the histogram of the count of reference instances and S' is the histogram of the count of system instances for the target activity. R' and S' both have N_{frames} bins, thus R'_i is the value of the i^{th} bin R' while S'_i is the value of the i^{th} bin S'. S' is the total count of system instances in frame i and R' is the total count of reference instances in frame i. False alarm time is computed by summing over the positive difference of $S' - R'$(shown in red in Figure 4); that is the duration of falsely detected system instances. This value is normalized by the non-reference duration of the video to provide the T_{fa} value in Equation above. Figure 5 shows visual representations of the the DET, we used Time-based false alarms and calculated $nAUDC$ from T_{fa} 0 to 0.2.

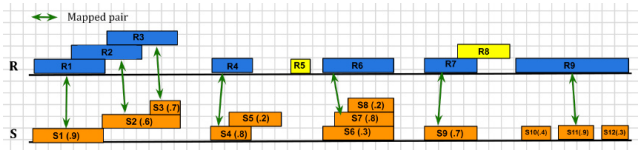

Figure 3. Illustration of activity instance alignment and P_{miss} calculation.

4.2. Runtime Speed Calculations and Time-Limited Scoring

We reported the runtime speed for each submission, since systems are expected process video faster than real time.

If a SDL system takes longer than realtime processing the videos, the results are rescored by NIST. This was done to simulate what the score would be if system execution had stopped in the middle of the sub part when the runtime exceeded realtime. Time-limited metric calculated are, Time limited $nAUDC$ and Time limited $mean-p_{miss}@0.04T_{fa}$ reported on the ActEV SDL leaderboard.

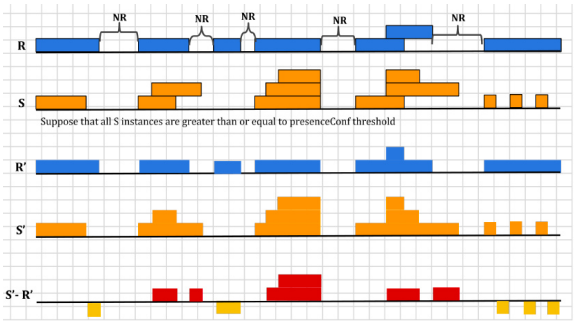

Figure 4. Pictorial depiction of T_{fa} calculation.

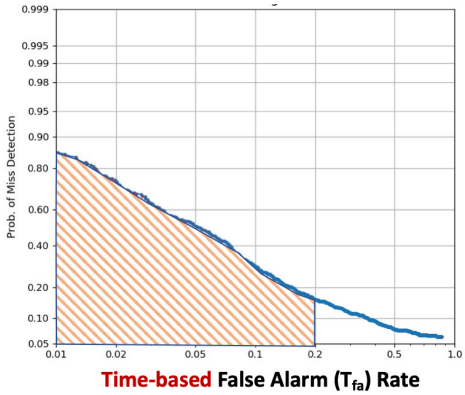

Figure 5. Detection Error Tradeoff (DET) curve, and $nAUDC$ is calculated with a fixed $T_{fa} = 0.2$.

5. Results

In this subsection, we present and compare the results for the ActEV'20 SDL evaluation for the system submitted by May 17th, 2020 on the MEVA Test3 dataset. Ten teams took part in the sequestered leaderboard and submitted 40 submissions and the lowest detection error results for each team are presented.

Table 2 presents a list of the participants ranked by the $nAUDC$ results, the $nAUDC$ results are the best system per team. The top ranking on activity detection is by UCF at 37 %, followed by CMU at 39%, and OPPO at 41%.

Figure 6 the ranks the performance of the ten teams. The results show the ranking of system performance: the x-axis is the ten teams, and the y-axis is the metric $nAUDC$, the lower value is considered better performance. The black points indicates a mean value for each team (marked points in green) and the green error bar indicates the standard deviation. The tan points shows the $nAUDC$ values of the 37

different activities

Figure 7 illustrates the ranking of the activities across systems. The plot shows the ranking of activities across systems the x-axis is the activity type, and the y-axis is the metric $nAUDC$. The points marked in black indicates a mean value across different systems and the green error bar indicates its standard deviation. The tan points shows the values of the ten teams".

The performance ranking by activity by the average system are reported in Table 3. From the table, for example the person_abandon_package and person_steals_object are the most difficult activities to detect given the dataset and target activity list.

Figure 8 shows the ranking of the teams on the relative run time. x-axis shows the ten teams and y-axis is the relative processing time. Based on the Figure 8 vireoJD and IBM are the fastest systems.

Figure 9 shows the activities that are ordered by the level of difficulty for each team. The x-axis shows the team names and average activity ranking (AVG), the y-axis, shows the 37 activities, and the numbers in the matrix, show the the ranking of 37 activities per system.

Finally, the system performance vs submission days for all the teams is shown in Figure 10, x-axis shows days since 2019/07/01 and y axis shows the $nAUDC$. It shows large improvement in system performance over time from the start of the ActEV'19 SDL, which was based on MEVA Test2, compared to ActEV'20 SDL which is based on MEVA Test3.

Table 2. Challenge participants ranking, the results are for the submission deadline of May17th, 2020.

Team	Organization	$nAUDC$	proc_time
UCF	University of Central Florida	0.365	0.684
INF-CMU	Carnegie Mellon University	0.387	0.498
VUS	OPPO Research Institute	0.406	1.344
UMD+UCF	UMD+UCF	0.415	1.253
UMD	University of Maryland	0.466	0.684
IBM-MIT-Purdue	Purdue University, USA	0.505	0.366
Team-vision	International Business Machines	0.530	1.002
vireoJD-MM	City University of Hong Kong	0.539	0.149
BUPT-MCPRL	Beijing University of Posts & Tel	0.615	0.969
CIS_JHU	Johns Hopkins University	0.629	4.520

6. Conclusion

In this report, we presented the results from the ActEV'20 SDL challenge that was held as a task under the CVPR'20 Activity workshop and ran on the MEVA Test3 dataset. The competition was open to the public and run as sequestered evaluation.

Ten teams participated in the ActEV'20 SDL evaluation which was based on MEVA Test3 dataset and a total of the 40 systems were submitted. We observed that out of all the target activities the person_abandons_package and person_steals_object are the hardest activities to detect across

Table 3. Performance Ranking by Activity on the Test3 dataset.

Activity	$nAUDC$	Activity	$nAUDC$
person_rides_bicycle	0.252	person_puts_down_object	0.489
vehicle_makes_u_turn	0.263	person_stands_up	0.506
person_reads_document	0.263	person_embraces_person	0.511
person_purchases	0.273	person_picks_up_object	0.519
vehicle_reverses	0.292	person_closes_trunk	0.536
vehicle_picks_up_person	0.340	hand_interacts_with_person	0.537
vehicle_turns_left	0.368	person_opens_facility_door	0.539
person_texts_on_phone	0.386	person_transfers_object	0.552
person_talks_to_person	0.392	person_exits_vehicle	0.552
vehicle_turns_right	0.398	person_opens_vehicle_door	0.574
vehicle_drops_off_person	0.416	person_closes_vehicle_door	0.593
person_sits_down	0.431	person_enters_scene_through	0.600
vehicle_stops	0.434	person_exits_scene_through	0.602
person_interacts_with_laptop	0.440	person_unloads_vehicle	0.613
vehicle_starts	0.440	person_loads_vehicle	0.655
person_talks_on_phone	0.450	person_closes_facility_door	0.664
person_carries_heavy_object	0.465	person_abandons_package	0.834
person_enters_vehicle	0.476	person_steals_object	0.836
person_opens_trunk	0.479	person_abandons_package	0.838

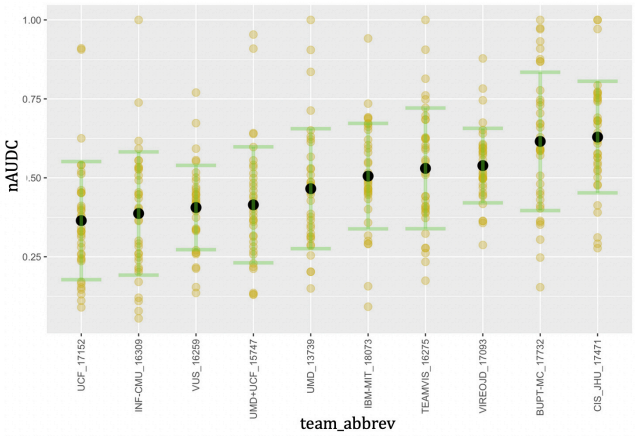

Figure 6. Ranking of System Performance for the ten teams.

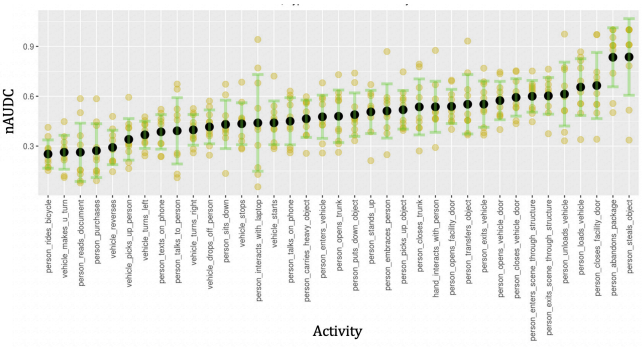

Figure 7. Ranking of Activities across Systems

systems. Based on Figure 10, we observed remarkable improvement of system performance over the SDL challenge

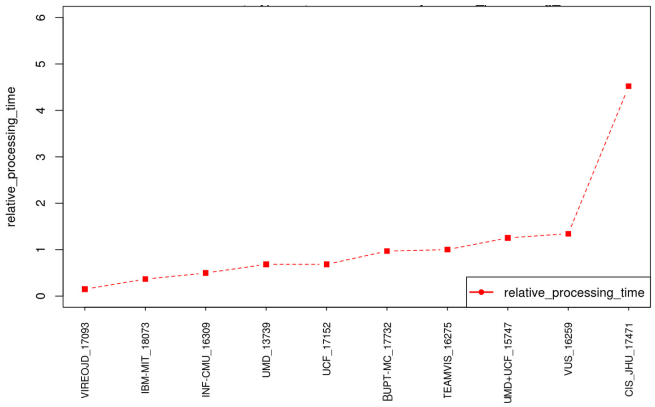

Figure 8. Ranking of the teams on the relative run time

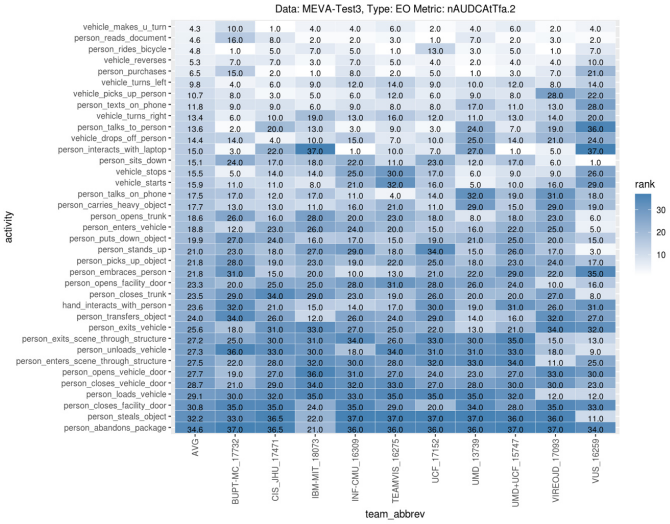

Figure 9. Activity difficulty across systems?

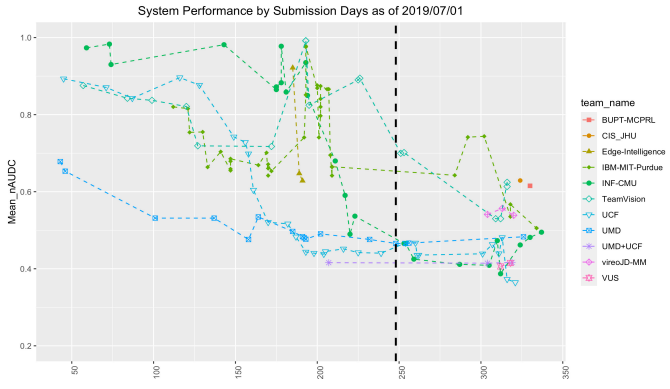

Figure 10. System Performance vs Submission Days

from the ActEV'19 SDL to ActEV'20 SDL. Based on the ActEV SDL submission deadline of May 17th, 2020, we invited the top-three teams for the CVPR'20 Activity workshop presentations: UCF was awarded first place, CMU received second place, and OPPO received third place.

The future ActEV SDL challenge will also include an optional surprise/ad-hoc activity component where a textual description and a limited number of exemplars are provided to the system at test time requiring on-line system training without the developer in the loop.

The ActEV SDL competition provided researchers an opportunity to evaluate their activity detection technologies on sequestered dataset. The competition also resulted in outstanding progress in improving activity detection accuracy. We hope these ActEV SDL challenges, and the associated MEVA datasets will facilitate the development of advanced solutions for real-time activity detection for public safety videos. This will in turn provide provide an impetus for more research in the field of activity detection

Disclaimer: Certain commercial equipment, instruments, software, or materials are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, NIST, or the U.S. Government.

Acknowledgement: The NIST work was supported by the IARPA, agreement IARPA-16002 #D2018-1807230003. The authors would like to thank Kitware, Inc. for collecting and annotating the dataset.

References

[1] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzad Godil, David Joy, Andrew Delgado, Alan Smeaton, Yvette Graham, et al. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. 2018.

[2] George Awad, Asad A Butt, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F Smeaton, and Yvette Graham. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. 2019.

[3] Marcos Baptista-Ríos, Roberto J López-Sastre, Fabian Caba-Heilbron, Jan Van Gemert, F Javier Acevedo-Rodriguez, and Saturnino Maldonado-Bascón. The instantaneous accuracy: a novel metric for the problem of online human behaviour recognition in untrimmed videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[4] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402. IEEE, 2005.

[5] Home Office Scientific Development Branch. Imagery library for intelligent detection systems (i-Lids). In *2006 IET Conference on Crime and Security*, pages 445–448. IET, 2006.

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[7] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *European Conference on Computer Vision*, pages 269–284. Springer, 2016.

[8] Jiyang Gao, Kan Chen, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 68–83, 2018.

[9] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[10] IARPA. Deep Intermodal Video Analytics (DIVA). `https://www.iarpa.gov/index.php/research-programs/diva`, 2018. Accessed: 2020-08-12.

[11] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos in the wild. *Computer Vision and Image Understanding*, 155:1–23, 2017.

[12] IEEE. 2nd International Workshop on Human Activity Detection in multi-camera, Continuous, long-duration Video (HADCV'20). `https://actev.nist.gov/workshop/hadcv20`, 2020. Accessed: 2020-08-30.

[13] David M Joy and Jonathan G Fiscus. 2017 trecvid multimedia event detection evaluation plan. Technical report, 2017.

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[15] 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.

[16] Corona Kellie, Osterdahl Katie, Collins Roderic, and Hoogs Anthony. Meva: A large-scale multiview, multimodal video dataset for activity detection. *IEEE Winter Conf. on Applications of Computer Vision (WACV) Conference, January 5, 2021*, 2021.

[17] Kitware. MEVA Data Website. `http://www.mevadata.org`, 2020. Accessed: 2020-08-12.

[18] Kitware. VIRAT Data Website. `http://www.viratdata.org`, 2020. Accessed: 2020-08-12.

[19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[20] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[21] Yooyoung Lee, Jon Fiscus, Afzal Godil, Andrew Delgado, Jim Golden, Lukas Diduch, and Maxime Hubert. Summary of the 2019 Activity Detection in Extended Videos Prize Challenge. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pages 148–154, 2020.

[22] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Guo-Jun Qi, Rui Qian, Tao Wang, Nicu Sebe, Ning Xu, Hongkai Xiong, et al. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020.

[23] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009.

[24] M Michel, J Fiscus, and D Joy. Trecvid 2017 surveillance event detection evaluation, 2017.

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision*, pages 2630–2640, 2019.

[26] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[27] NIST. ActEV Sequestered Data Leaderboard Website. `https://actev.nist.gov/sdl`, 2019. Accessed: 2020-08-12.

[28] NIST. TREC Video Retrieval Evaluation: TRECVID Website. `https://trecvid.nist.gov/`, 2020. Accessed: 2020-08-12.

[29] OpenStack. The OpenStack Foundation. `https://www.openstack.org/foundation/`, 2020. Accessed: 2020-08-12.

[30] Marcos Baptista Rios, Roberto J López-Sastre, Fabian Caba Heilbron, Jan van Gemert, F Javier Acevedo-Rodríguez, and SATURNINO Maldonado-Bascón. Rethinking online action detection in untrimmed videos: A novel online evaluation protocol. *arXiv preprint arXiv:2003.12041*, 2020.

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[32] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings*

of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., volume 3, pages 32–36. IEEE, 2004.

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[34] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[35] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019.

[36] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019.

[37] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[38] ActivityNet workshop. CVPR'20 International Challenge on Activity Recognition (ActivityNet workshop). `http://activity-net.org/challenges/2020/index.html`, 2020. Accessed: 2020-08-12.

[39] Yooyoung Yooyoung, Jon Fiscus, Afzal Godil, David Joy, Andrew Delgado, and Jim Golden. Actev18: Human activity detection evaluation for extended videos. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 1–8. IEEE, 2019.

[40] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.