Summary of the 2019 Activity Detection in Extended Videos Prize Challenge

Yooyoung Lee, Jon Fiscus, Afzal Godil, Andrew Delgado, Jim Golden, Lukas Diduch[+], Maxime Hubert[*]

National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899
{yooyoung.lee, jonathan.fiscus, afzal.godil, andrew.delgado, jim.golden}@nist.gov

[+]Dakota Consulting Inc. 1110 Bonifant Street, Suite 310, Silver Spring, MD 20910
lukas.diduch@dakota-consulting.com

[*]Thalès Alenia Space, 26 Avenue Jean François Champollion, 31100 Toulouse, France
maxime.hubert@outlook.com

Abstract

Despite previous data collection efforts and benchmark studies, progress in activity detection technologies has been slow, especially with applications that meet practical needs for the video analytics domain. In this paper, we discuss the results from the Activity detection in Extended Video Prize Challenge (ActEV-PC) that was sponsored by IARPA. The goal of the ActEV-PC was to promote robust automatic activity detection system development and to reduce the detection error rate. To examine the ability of activity detection systems in different aspects, we opened a competition to the public and ran evaluations (as a task under the ActivityNet workshop at CVPR 2019) with two different phases: an open leaderboard evaluation and a sequestered data evaluation. The Video and Image Retrieval and Analysis Tool (VIRAT) dataset was used for the open leaderboard evaluation while the Multiview Extended Video with Activities (MEVA) dataset was used for the sequestered data evaluation. Eighteen target activities were defined for detection. In this paper, we present results and findings from the two-phase ActEV-PC competition. Eighteen teams from academia and industry participated in the competitions and three top performers received a cash award (funded by IARPA). The winners were presented at the ActivityNet Workshop at CVPR 2019.

1. Introduction

Despite previous data collection efforts and benchmark studies, progress in automatically detecting and understanding human activities in video has been slow, especially with applications that meet practical needs for the video analytics domain. Impeding challenges [1] include the large variability in human activity instantiation styles, complexity of the visual stimuli in terms of camera frame motions, background clutter and viewpoint changes, as well as the level of detail of the activities.

In 2017, the National Institute of Standards and Technology (NIST) developed the Activities in Extended Video (ActEV) evaluation series [2][3][4] to support the metrology needs of the Intelligence Advanced Research Projects Activity (IARPA) Deep Intermodal Video Analytics (DIVA) Program [5].

To understand current state-of-the-art and to promote activity detection technologies, the ActEV prize challenge (ActEV-PC) [6] was open to the public (sponsored by IARPA) and competitions were conducted as a task under the ActivityNet challenge at CVPR2019 [7]. The goal of the ActEV-PC was to facilitate the development of video analytic technologies that can automatically detect target activities and to reduce the detection error rate.

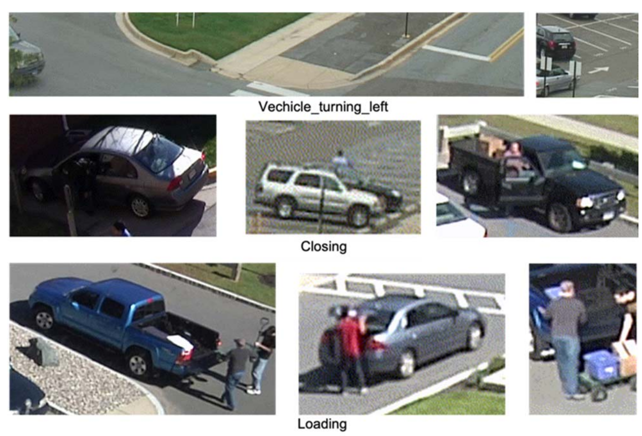

Figure 1 Examples of activity types
IRB (Institutional Review Board) number: 00000755

The ActEV-PC was comprised of a two-phase competition: an open leaderboard and a sequestered data evaluation. Two datasets were used: VIRAT [8] for the open leaderboard evaluation and the MEVA M1 dataset [9] for the sequestered data evaluation. Figure 1 illustrates examples of the activity types for both competitions.

In this paper, we discuss the evaluation task, performance measures, and datasets, and present results and observations for the ActEV-PC competitions. The paper is organized as follows: Section 2 describes related work in activity detection and classification. Section 3 describes the ActEV-PC evaluation task and performance measure. Sections 4 and 5 summarize the evaluation framework and datasets respectively. Finally, in Section 6 we present the results and findings.

2. Related Works

The recent evolution of system development with machine learning techniques and large visual datasets have revolutionized the computer vision and video analytics communities.

In this section, we provide a comparison of existing datasets associated with activity detection and classification. Table 1 illustrates a detailed comparison of the datasets used for activity classification and localization (temporal and spatio-temporal localization).

Table 1 Comparison of datasets for activity classification and detection/localization

Datasets	Source	Activity Classes	Temporal Localization	Spatio-Temporal Localization
ActEV VIRAT [8]	Multi-Camera Security Video	50	Yes	Yes
ActEV MEVA [9]	Multi-Camera Security Video	>37	Yes	Yes
SED/i-LIDS [10]	Multi-Camera Security Video	10	No	No
USF101 [11]	YouTube	101	No	No
HMDB51 [12]	Movies, YouTube	51	No	No
THUMOS 15 [13]	YouTube	101	Yes	No
ActivityNet [1]	YouTube	200	Yes	No
AVA [14]	Movies	80	Yes	Yes
HACS [15]	YouTube	200	Yes	No
The Sports-1M [16]	YouTube	487	No	No
Charades [17]	266 Homes	157	Yes	No
Kinetics-700 [18]	YouTube	600	No	No
YouTube-8M [19]	YouTube	3862	No	No
Moments in Time Recognition [20]	YouTube	339	No	No

The existing datasets listed in this table, are mainly derived from social media (YouTube videos) or from movies, except the VIRAT and MEVA datasets used by the ActEV evaluation series. These datasets contain multi-camera, continuous, long-duration video, often taken at significant stand-off ranges from the activities of interest. Although the ActEV series addresses both activity detection and temporal (and spatio-temporal) localization of the activity, the ActEV-PC competition has primarily focused only on temporal activity detection. In the MEVA and VIRAT datasets, multiple activities can happen at any time, anywhere in the frame and across cameras. The VIRAT dataset is a large-scale video dataset designed to assess the performance of activity detection algorithms in realistic scenes. The MEVA dataset is much larger and has a higher resolution than the VIRAT dataset; it contains hundreds of video hours with a number of instances of each activity from multiple viewpoints that are collected by a multi-camera IP network in a heterogeneous environment. The stage for the data collection contains the interior and exterior of a group of buildings, grounds of the buildings and the surrounding roads. The VIRAT and MEVA datasets facilitate both detections of activities and localizations of the corresponding spatio-temporal location of objects associated with activities.

Both the VIRAT and MEVA datasets are unique relative to other datasets in that they are far more closely aligned with real-world public safety video analytics. The primary purpose of the data is to stimulate the computer vision community to develop advanced human activity detection algorithms with improved performance and robustness for multi-camera systems that cover a large area.

3. Evaluation Tasks and Measures

The purpose of the ActEV evaluation series is to promote the development of systems that automatically:

- identify a target activity along with the time span of the activity (activity detection/localization)
- detect objects associated with the activity instance (activity and object detection), and
- track multiple objects associated with the activity instance (activity and object detection and tracking).

The ActEV-PC evaluation primarily focused on the development of robust automatic activity detection systems in the context of extended videos. The extended videos in this paper are defined as video with long duration for days/weeks/months. ActEV-PC systems ran on a set of activities previously known to the system.

In the activity detection (AD) task for the ActEV-PC competitions, given a target activity, a system automatically detected and temporally localized all instances of the target activity in a single-camera video. The system was required to provide the start and end frames indicating the temporal location of the target activity and a presence confidence score with higher values indicating the activity instance was more likely to have occurred.

To evaluate system performance, we modified the metrics from TREC Video Retrieval Evaluation (TRECVID) surveillance event detection (SED) [10] and Classification of Events Activities and Relationships (CLEAR) [21] evaluations.

The primary metric evaluated how accurately the system detected the occurrences of the activity. The scoring method comparing the reference and system output had four distinct steps: 1) instance alignment, 2) confusion matrix computation, 3) summary performance metrics, and 4) graphical analysis of the Type I/II error tradeoff space.

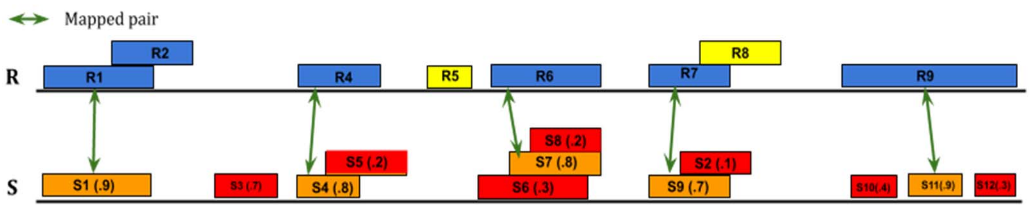

Figure 2 A pictorial depict of activity instance alignment, MD and FA calculation

The goal of the alignment step was to find a one-to-one correspondence between the reference and system output instances. This step was required because a single system instance cannot be counted as correct for multiple reference instances. For example, if there are two "closing _trunk" instances that occur at the same time but in separate regions of the video and there was a single detection by the system, one of the reference instances was missed. Thus, we utilized the Hungarian algorithm [22] to find an optimal mapping while reducing the computational complexity.

The next step was to calculate the detection confusion matrix for activity instance occurrence. Correct Detection (CD) indicates that the reference instance (R) and system output instance (S), were correctly mapped. Missed Detection (MD) indicates that an instance in the reference had no correspondence in the system output while False Alarm (FA) indicates that an instance in the system output had no correspondence in the reference. In Figure 2, the first number shown following the S is the instance ID and the second shown in parentheses is the presence confidence score that indicates how likely the instance is associated with the target activity. For example, S1 (.9) represents the instance S1 with corresponding presence confidence score of 0.9. Green arrows indicate alignment between R and S. It also identifies system instance S4 as a better match (than S5) to reference instance R4 when considering the presence confidence values. Yellow instances {R5, R8} are missed detections and red instances {S2, S3, S5, S6, S8, S10, S12} are false alarms.

After calculating the confusion matrix, we summarized system performance. The confidence score was used as a decision threshold, enabling a probability of missed detections (P_{miss}) and a rate of false alarms (R_{FA}) to be computed at a given threshold:

$$P_{miss}(\tau) = \frac{8 + N_{MD}(\tau)}{10 + N_{TrueInstance}}$$

$$Rate_{FA}(\tau) = \frac{N_{FA}(\tau)}{VideoDurInMinutes}$$

where $N_{MD}(\tau)$ is the number of missed detections at the threshold τ, $N_{FA}(\tau)$ is the number of false alarms, $VideoDurInMinutes$ is the number of minutes of video, and $N_{TrueInstance}$ is the number of reference instances annotated in the sequence.

The P_{miss} was calculated with a weighted value that is most relevant for activities that have few instances, reflecting a prior belief on P_{miss} being around 0.8. Activities for which there are many instances to detect would overcome this prior, and activities for which there are fewer instances would be more weighted by the prior. This value was then averaged over all activities in the video. The total instance count of each activity on the leaderboard and in the sequestered data will not be published, but activities' relative instance counts do differ from public datasets.

For the ActEV-PC evaluation, therefore, we evaluated system performance on the probability of missed detection with a weighted value (wP_{miss}) at a specific operating point (wP_{miss} at $R_{FA} = 0.15$) and then averaged over activity types.

Lastly, as illustrated in Figure 3, the Detection Error Tradeoff (DET) curve [23] was used to visualize system performance.

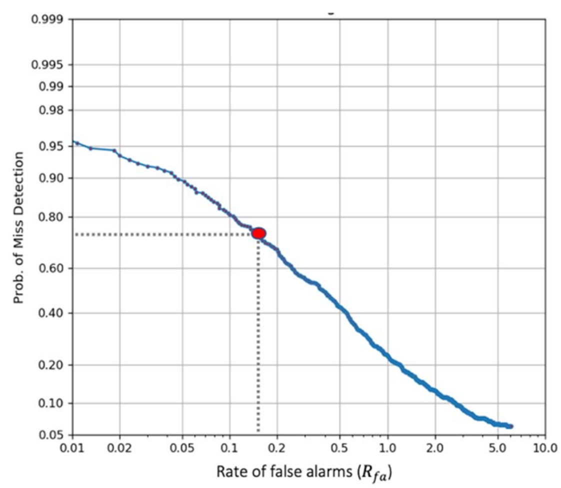

Figure 3 An example of Detection Error Tradeoff (DET) curve and the operating point of interest

4. Evaluation Framework

For ActEV-PC, there were the two evaluation phases: 1) open leaderboard and 2) sequestered data. In the open leaderboard evaluation, the participants ran their software systems on their own in-house computing hardware and submitted the system output in a defined format to the NIST public scoring server. The leaderboard evaluation provided an overall performance score after aggregating system performance across all target activities. Developers could process the test collection multiple times and receive performance scores immediately.

For the sequestered data evaluation, the participants submitted their runnable system to the NIST public scoring server, which was independently evaluated on the sequestered data using the NIST evaluation hardware.

5. Datasets

In the ActEV-PC competitions, we used the VIRAT dataset for the open leaderboard evaluation and the MEVA M1 dataset for the sequestered data evaluation. Both datasets were annotated by Kitware, Inc. [9][24].

The same 18 target activities were used in both the open leaderboard and sequestered data evaluations. However, the number of instances for each activity between the VIRAT and MEVA test sets differ. The detailed definition of each activity is described in the evaluation plan [6]. Table 2 lists the number of instances for each activity for the training and validation sets only. Due to ongoing evaluations, test set statistics are not included in the table. The number of instances for the test sets are not balanced across activities (similar to the training and validation sets shown table below), which may affect the system performance results.

Table 2 A list of 18 activities and their associated number of instances for the train and validation sets

Activity Type	Train	Validation
Closing	126	132
Closing_trunk	31	21
Entering	70	71
Exiting	72	65
Loading	38	37
Open_trunk	35	22
Opening	125	127
Transport_HeavyCarry	45	31
Unloading	44	32
Vehicle_turning_left	152	133
Vehicle_turning_right	165	137
Vehicle_u_turn	13	8
Pull	21	22
Riding	21	22
Talking	67	41
Activity_carrying	364	237
Specialized_talking_phone	16	17
Specialized_texting_phone	20	5

6. Results and Analyses

In this section, we present a summary of the evaluation results and speed measurements from the ActEV-PC open leaderboard and sequestered data evaluations.

6.1. Phase 1: Open Leaderboard Evaluation

A total of 18 teams from academia and industry participated in this competition. Each team was allowed to upload multiple submissions, and each team's submission with the lowest detection error based on the mean weighted P_{miss} at $R_{FA} = 0.15$ was selected for the following results.

For the given 18 activities on the VIRAT dataset, Table 3 summarizes the best performance per team for the AD task (submission deadline as of 03/21/19). We had a total of 19 systems from 18 challenge participants plus one baseline system. wP_{miss} at $R_{FA} = 0.15$ was used to rank activity detection performance. For simplicity, we list the values of the metrics using the average values across all 18 activities for each system. The systems are alphabetically ordered and the primary measure is the mean wP_{miss} at $R_{FA} = 0.15$ (μwP_{miss} at $R_{FA} = 0.15$: WPR.15, marked in gray)—a smaller value denotes better performance.

Table 3 ActEV-PC open leaderboard results (submission deadline as of 03/21/19) on the VIRAT dataset

WPR.15: μwP_{miss} at R_{FA} = .15, PR.15:μP_{miss} at R_{FA} = .15

↓: lower value is considered as better system performance

Team	WPR.15↓	PR.15↓	Eligibility
Baseline_ACT	0.907	0.917	N
Baseline_RC3D	0.913	0.922	N
BUPT-MCPRL	0.699	0.678	Y
IBM-MIT-Purdue	0.757	0.743	N
INF (CMU)	0.736	0.718	N
IVP	0.937	0.944	Y
JHUDIVATeam	0.793	0.790	N
NtechLab	0.806	0.803	Y
MUDSML	0.702	0.683	N
Shandong Normal Univ.	0.858	0.858	Y
SRI	0.805	0.801	N
STARK (IBM)	0.758	0.744	N
STR-DIVA Team	0.762	0.749	N
UCF	0.750	0.735	N
UMD	0.750	0.735	N
UNSW_InsData_PC	0.742	0.730	Y
USF Bulls	0.888	0.896	Y
vireoJD-MM	0.768	0.759	Y
XXR	0.972	0.971	Y

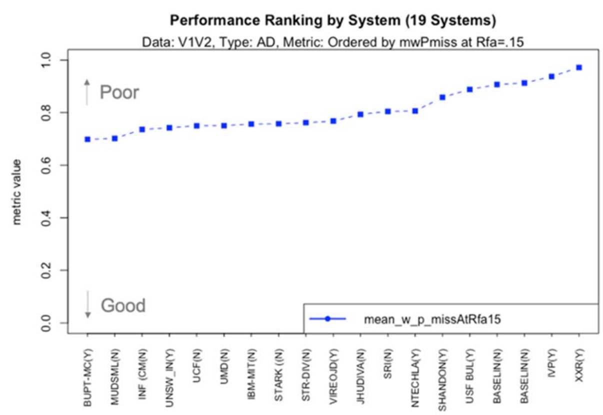

Figure 4 The ranked list of system performance (AD)

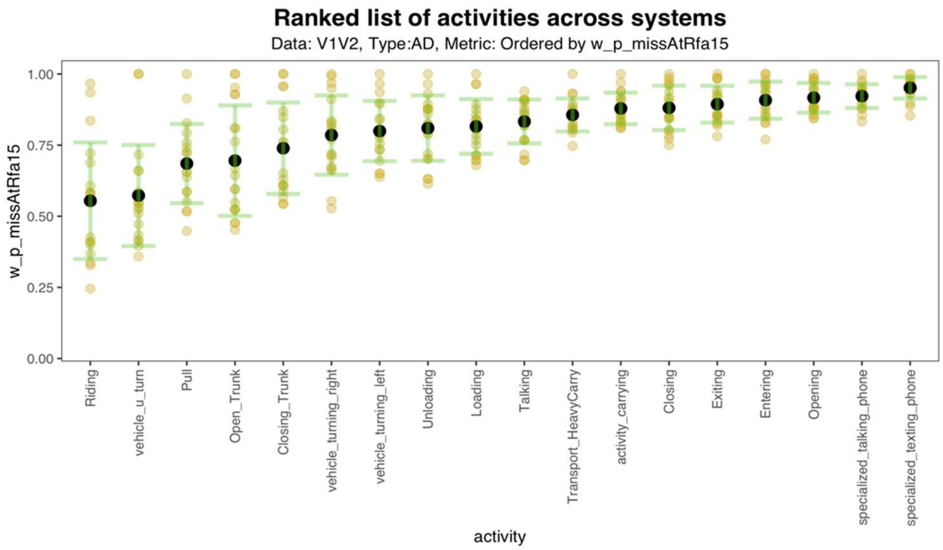

Figure 5 Summary of the different levels in detection difficulty among the 18 activities from the phase-1 evaluation

The metric μP_{miss} at $R_{FA} = .15$ that was used as a performance measure during the previous year's ActEV 2018 evaluation [2] is listed as PR.15: μP_{miss} at $R_{FA} = 0.15$ in this table for comparison purposes. Some of the participants were not eligible for a prize since they received funding from the IARPA DIVA program; hence, we included the prize eligibility for each team.

Figure 4 shows the ranking of the 19 systems (ordered by WPR.15: μwP_{miss} at $R_{FA} = 0.15$). The x-axis lists the systems and the y-axis shows the metric value of μwP_{miss} at $R_{FA} = 0.15$, where lower values are considered better performance.

The results show that, for activity detection, BUPT-MCPRL achieved the lowest error rate (WPR.15: 66.9%) followed by MUDSML (WPR.15: 70.2%).

Figure 5 addresses the question of the different levels in detection difficulty among the 18 activities for a given test dataset. To determine the activity detection difficulty, the activity types are characterized by the average performance across all 19 system outputs from the open leaderboard sub-missions. In Figure 5, the x-axis contains activities and the y-axis is wP_{miss} at $R_{FA} = 0.15$. For the VIRAT dataset, "riding" and "vehicle_u_turn" activities are generally easier to detect compared to the rest of the other activities.

6.2. Phase 2: Sequestered Data Evaluation

The top six performers (from the open leaderboard participants) who were eligible for the prize were invited to submit their systems to the sequestered data evaluation. Three teams out of the six submitted systems to NIST.

Table 4 summarizes the invited teams, their submission status and system performance across all 18 activities. The metrics were first calculated on each activity and averaged across all activities on the entire dataset. The results show that the BUPT-MCPRL [25] team has the lowest error rate on μwP_{miss} at $R_{FA} = 0.15$ (WPR.15) followed by the vireoJD-MM [26] team.

Table 4 Sequestered data evaluation results on the MEVA M1 (ordered by WPR.15: μwP_{miss} at $R_{FA} = 0.15$ marked in gray)

Teams Invited	Submitted System	WPR.15↓
BUPT-MCPRL	Y	0.889
vireoJD-MM	Y	0.906
NtechLab	Y	0.925
UNSW_InsData_PC	N	NA
Shandong Normal University	N	NA
USF Bulls	N	NA

In addition, ActEV-PC had a speed requirement that states that systems should not be more than 20 times slower than real-time; real-time processing runtime in this evaluation refers to the processing at the same rate as the input video on a defined hardware specification.

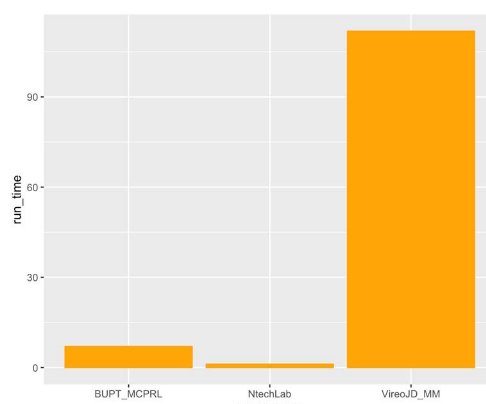

Figure 6 System processing run-time relative to real-time on MEVA M1

Figure 6 summarizes video processing run-time on the MEVA M1 dataset for each system. The x-axis is the system developer and the y-axis denotes the processing runtime relative to real-time.

The ActEV-PC competitions took both detection accuracy and system video processing run-time into account when ranking systems for the prize awards.

BUPT_MCPRL and NtechLab submissions successfully processed all videos while VireoJD_MM failed to process a subset of videos. In addition, as illustrated in Figure 6, vireoJD-MM did not meet the speed requirement (and operated more than 20 times slower than real-time).

7. Summary

In this paper, we presented the results from the Activities in Extended Video Prize Challenge (ActEV-PC). The competition was open to the public and evaluations were run (as a task under the ActivityNet workshop at CVPR 2019) with two different phases: an open leaderboard evaluation and a sequestered data evaluation. We used 18 target activities from the VIRAT dataset for the open leaderboard evaluation and from the MEVA M1 dataset for the sequestered data evaluation.

Eighteen teams participated in the phase 1 open leaderboard and three teams submitted their systems to the phase 2 sequestered data evaluation; BUPT-MCPRL, NtechLab, and vireoJD-MM. Figure 7 illustrates a summary result of the two evaluation phases for the three teams. The first set of histograms (left) represents the results from the open leaderboard on the VIRAT dataset while the last set of histograms (right) indicates the results from sequestered data evaluation on the MEVA M1 dataset. The center set of histograms show results from a common subset of the MEVA data where all submissions successfully processed the data, which is shown since vireoJD-MM did not complete processing for some of the videos.

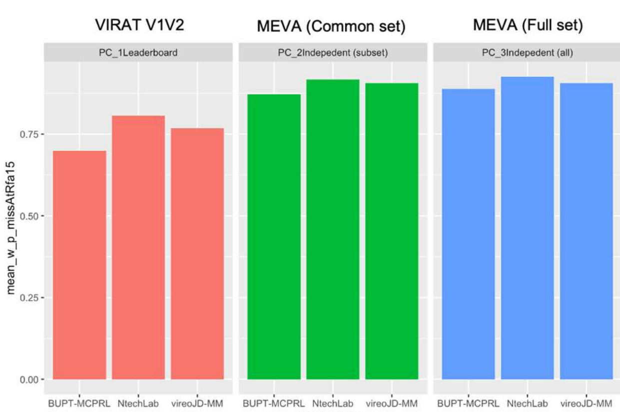

Figure 7 Summary results for open leaderboard and sequestered data evaluations for the three top performers

For system performance, the results show that BUPT-MCPRL had the lowest detection error rate (based on $\mu w P_{miss}$ at $R_{FA} = 0.15$) followed by vireoJD-MM, and NtechLab. However, vireoJD-MM did not meet the required speed time bound.

BUPT-MCPRL was awarded first place (1st most accurate detection within the run-time bound), Ntechlab received second place (3rd most accurate detection within the run-time bound), and vireoJD-MM received third place (even though the system provided the 2nd most accurate detection, it did not meet the of run-time requirement).

We observed that given the target activities in the test set, riding and "vehicle_u_turn" activities were easiest to detect across systems.

The ActEV-PC competitions provided researchers an opportunity to evaluate their activity detection technologies on both public and sequestered datasets. The competition also resulted in outstanding progress in improving activity detection accuracy.

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

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.

References

[1] 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), 2015, pp. 961–970, doi: 10.1109/CVPR.2015.7298698.

[2] Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, and J. Golden, "ActEV18: Human Activity Detection Evaluation for Extended Videos," in 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), doi: 10.1109/WACVW.2019.00008.

[3] G. Awad et al., "TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search," p. 38.

[4] "ActEV: Activity Detection in Extended Videos." https://actev.nist.gov/.

[5] "IARPA Deep Intermodal Video Analytics (DIVA) Program." https://www.iarpa.gov/index.php/research-programs/diva/diva-baa.
[6] "ActEV-PC: ActEV Prize Challenge and Evaluation Plan," 2019. https://actev.nist.gov/prizechallenge.
[7] International Challenge on Activity Recognition (ActivityNet) workshop in 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2019, http://activity-net.org/challenges/2019/.
[8] S. Oh et al., "A large-scale benchmark dataset for event recognition in surveillance video," in CVPR 2011, pp. 3153–3160, doi: 10.1109/CVPR.2011.5995586.
[9] Kitware Inc, "The Multiview Extended Video with Activities (MEVA) dataset." https://mevadata.org/.
[10] M. Michel, J. Fiscus, and D. Joy, "TRECVID 2017 Surveillance Event Detection Evaluation.", https://www.nist.gov/itl/iad/mig/trecvid-surveillance-event-detection-evaluation-track.
[11] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," ArXiv12120402 Cs, Dec. 2012.
[12] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition,"
[13] "THUMOS Challenge 2015." http://www.thumos.info
[14] C. Gu et al., "AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions."
[15] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization," ArXiv171209374 Cs, Sep. 2019.
[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725–1732, doi: 10.1109/CVPR.2014.223.
[17] "Charades Challenge." http://vuchallenge.org/charades.html.
[18] "DeepMind Research – Kinetics | DeepMind." https://deepmind.com/research/open-source/open-source-datasets/kinetics/.
[19] "YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research." https://research.google.com/youtube8m/.
[20] "Moments in Time Challenge." http://moments.csail.mit.edu/challenge.html.
[21] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," EURASIP J. Image Video Process., vol. 2008, doi: 10.1155/2008/246309.
[22] J. Munkres, "Algorithms for the Assignment and Transportation Problems," J. Soc. Ind. Appl. Math., vol. 5(1), pp. 32–38, 1957.
[23] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance", National Inst of Standards and Technology Gaithersburg MD, 1997.
[24] Kitware Inc., "The Video and Image Retrieval and Analysis Tool (VIRAT) dataset." https://viratdata.org.
[25] Y. Li, S. Xu, X. Cheng, Y. Zhao, Z. Zhao, F. Su, and B. Zhuang, "An Effective Detection Framework for Activities in Surveillance Videos," BUPT-MCPRL team report: https://www.mcprl.com/essay/BUPT-MCPRL_report_for_ActEV-PC.pdf, 2019.
[26] F. Long, Q. Cai, Z. Qiu, Z. Hou, Y. Pan, T. Yao, and CW. Ngo, "vireoJD-MM at Activity Detection in Extended Videos," arXiv preprint arXiv:1906.08547, 2019.