Data, Algorithms, and Framework for Automated Analytics of Surveillance Camera Networks

Roddy Collins†, Katie Osterdahl†, Ameya Shringi†, Kellie Corona†,
Eran Swears†, Reuven Meth•, Anthony Hoogs†
† Kitware, Inc., 28 Corporate Dr, Clifton Park, NY 12065 firstname.lastname@kitware.com
• Engility, firstname.lastname@iarpa.gov

Abstract

Recent advances in areas such as video classification, captioning, and activity detection have significantly expanded the scope of automated analytics that can be performed on videos. These advances are supported by large datasets designed to facilitate training algorithms that can scale up to leverage data at the scale of gigabytes and terabytes. However, almost all these datasets have limited number of hours and annotations or use web-based services like flickr or YouTube to obtain more data, which results in data biased towards consumer expectations in terms of distance to the object of interest, object resolution, ratio of "interesting" to "uninteresting" video, and so forth. These videos differ from those of visual surveillance and public safety data, in which activities of interest are rare, may not be centered in the field of view, and may occur across multiple video streams. These differences limit the scope of transfer learning of models trained on consumer datasets when applied to public safety data.

In this paper we address these challenges by presenting new datasets developed for the IARPA Deep Intermodal Video Analytics (DIVA) program. The first dataset, DIVA-V1, extends the annotations for existing videos collected for the VIRAT Video Data project. The second dataset, DIVA-M1, was designed expressly for the DIVA program and collected approximately 9300 hours of video, using a 38 camera network to image over 100 actors executing scripted and unscripted activities across approximately two weeks. Annotation of the DIVA-M1 data is ongoing.

We additionally discuss results of baseline activity and object detection algorithms on the DIVA-V1 data.

Portions of this data are available for public research via NIST's ActEV (Activities in Extended Video) challenge.

1. Introduction

The volume of video data collected from ground-based video cameras has grown dramatically in recent years. However, there has not been a commensurate increase in the usage of intelligent analytics for real-time alerting or triaging of video archives. In many cases, security personnel or operators of camera networks are overwhelmed with the volume of video they must monitor, and cannot afford to view or analyze even a small fraction of their video footage in real-time. In non-real-time forensic scenarios, the analyst's efficiency when processing large volumes of video is hindered by the lack of widely available tools for automated analysis.

The IARPA Deep Intermodal Video Analytics (DIVA) program [1] is addressing these issues by developing robust automated activity detection (e.g. loading vehicle, carrying object) in a multi-camera streaming environment for faster real-time and forensic analysis. As a Test & Evaluation partner on the program, Kitware is responsible for developing and distributing of a large corpus of annotated data, developing an easily-accessible software framework that allows the community to experiment on that data, and determining baseline performance for object and activity detection.

The data developed for the DIVA program consists of four unique datasets, two completed and two planned, with each dataset providing distinct features including location, sensor configuration, arrangement of fields of view (overlapping and non-overlapping), activity types, and time of collection (time of year, time of day), resulting in a robust corpus of data. The goal for these datasets is to have a large number of realistic behaviors for a wide variety of activities to support algorithm development and analysis. The first dataset, DIVA-V1, is an extension of the annotations on the VIRAT public ground dataset that was released in 2012 [16], which represented 11 distinct scenes. We have increased the number of annotated activity

types from 12 to 47 and the number of annotated object types from 5 to 17. This corresponds to a 16x increase in the average number of activity instances per video clip for the training portion of the VIRAT dataset, with similar enhancements for the sequestered test data.

The second dataset, DIVA-M1, is a novel collection executed at a controlled facility with urban characteristics where 100+ actors with diverse demographics performing realistic activities over fifteen days. This second collection featured a combination of scripted and unscripted scenarios, captured by 38 interior / exterior ground-based cameras, which were a combination of fixed / PTZ and RGB / thermal IR sensors; two RGB cameras on unmanned aerial vehicles; and fourteen handheld or body-worn cameras. Additionally, all actors were provided with a unique GPS logger for tracking movement over the course of the data collect to facilitate re-identification during the annotation process. This dataset contains more than 9,300 total hours of raw ground camera data; portions will be annotated for 41 unique activity types, including embedded threat activities. The remaining two future dataset collections are planned to increase scene, actor, and activity diversity in both scripted and unscripted collections. Taken as a whole, the effort aims to create a benchmark for surveillance datasets that is comparable in terms of data samples to datasets like ImageNet [22] and MS-COCO [15]. These datasets played a pivotal role in obtaining near human accuracy in tasks such as image classification. DIVA aims to bring these advances and improve up them in a more focused surveillance setting. Based on the diversity and the size of the DIVA datasets, we believe they will be useful for developing online algorithms that can be deployed across large surveillance networks.

To this end, we benchmark a variety of baseline approaches representing state-of-the-art in object detection, activity recognition and activity detection. These algorithms demonstrate the open problems in multi-camera surveillance environments that have been largely unaddressed by state of the art object and activity detectors. Additionally, these algorithms are being used as test cases in the development of an open source framework that scales up to the camera networks and abstracts the scale of these networks from the algorithm developers by providing a well-defined API that provides data, evaluation, and execution support.

In this paper, we provide:

- An overview of the DIVA-V1 and DIVA-M1 datasets,
- A discussion of our annotation methodology,
- Baseline results on activity and object detection.
- A brief discussion of the framework, which is under active development.

1.1. The NIST ActEV Challenge

Portions of this data are being used to support the NIST Activities in Extended Video (ActEV) public challenge [4]; resources include the evaluation plan, scoring code, and a leaderboard tracking the performance of submitted systems on sequestered data.

2. Related Work

Activity detection is one of the core problems in computer vision with application numerous domains like visual surveillance [6], augmented reality [7, 23] and human computer interaction [17]. Thus development of activity detectors have been supported by small scale datasets like Weizmann [8], PETS04[11] and KTH [24]. These datasets had limited data instances, centered the actors in the scene, and were captured in a controlled environment. They were followed by datasets such as UCFSports [21], UCF50 [18] and HMDB51 [14], which used videos from the web to increase the data volume and content diversity. These datasets primarily focused on single actor activities. The VIRAT video dataset [16] introduced multi-actor activities in a real world environment; rather than relying on videos from the internet, this dataset is comprised of long scenes from fixed cameras obtained across multiple locations.

With the recent surge of interest in video understanding, several new datasets such as UCF101 [25], Kinetics [13] and Youtube-8M [5] have been introduced. These datasets are significantly larger and more diverse than their predecessors. UCF101 [25] is comprised of videos that have been spatially and temporally segmented to have a single activity. Kinetics [13] and Youtube-8M [5] are more focused on general video understanding rather than activity detection.

The most similar dataset to ours is WILDTRACK [9]. This dataset is consists of seven static surveillance style cameras viewing a single public open area with unscripted pedestrians. The cameras have overlapping fields of view and annotations are provided for people on 400 frames of each camera at 2 frames per second (i.e. about 3.5 minutes per camera). This is a

step in the right direction. In comparison, our dataset offers a wider variety of scenes/actors, significantly more raw data and annotations, multi-modal data, as well as both moving and stationary aerial and ground imagery.

Thus the task of complex activity detection in multi-view environment largely remains an open problem with little effort to create algorithms that operate on the same scale as image classifiers or object detectors. This lack of progress can be attributed to the lack of a large dataset to support the research. The DIVA datasets represent the first effort to make a major contribution to filling this gap and to facilitating development of algorithm for this setting.

3. DIVA Datasets

3.1. DIVA-V1 Dataset

3.1.1 Video Collection

As part of the DARPA Video and Image Retrieval and Analysis Tool (VIRAT) program [12], Kitware coordinated the collection of the ground camera component of the VIRAT Video Dataset [16]. The dataset consists of approximately 28 hours of stationary ground videos across 16 different scenes. The data contains both scripted activities and unscripted incidental background activity.

3.1.2 Annotation Process

Kitware's in-house annotation team was used to improve and expand the existing VIRAT annotations. These annotations were track-centric: all people and vehicles were tracked, along with certain object types. Activities were then defined to be temporal segments of these object tracks. Quality control included automated checks for common errors and an audit process. Using this process, 14.6 hours of video data were annotated with 17 object types and 45 activity types. Figure 1 visualizes the differences between the original VIRAT and new DIVA annotations.

As of this writing, 118 V1 clips have been released for training and validation; these total roughly 4.3 hours and contain 7,768 annotated activities.

3.2. DIVA-M1 Dataset

3.2.1 Video Collection

We designed and executed a large-scale video and audio data collection with more than 100 actors and 22 vehicles at an access-controlled training venue. The recording infrastructure was designed to provide a mix of overlapping and non-overlapping fields of view at different resolutions of various activities across indoor and outdoor venues at varying times of day. The scripted portions of data were designed to reflect completely realistic, natural scenarios including background behaviors and ranged in time from 20 minutes to 8 hours. Actors were managed with a hierarchical structure which designated a core set of actors as "squad leads" who in turned were responsible for a group of actors to whom they delegated roles, activities, and props. The resulting data set contains 9,283 hours of scripted and unscripted video data over two weeks across 38 ground-level cameras (PTZ, thermal IR, fixed IP). This data is being annotated to provide spatio-temporal annotations of selected activities. Figure 2 shows four views collected from the M1 Pilot collection exercise.

3.2.2 Annotation Process

We utilize a multi-step, modular annotation pipeline to provide activity-centric annotations. This contrasts to the track-centric approach taken with V1. To accommodate the range of complexity in annotation subtasks, we use a combination of in-house annotators and Mechanical Turk tasks to most effectively process the large volume of video data. Multiple annotators perform each task to minimize missed instances and increase overall accuracy. Using this process, we are annotating 40 activities and 7 unique objects.

3.2.3 Data Statistics

This data is currently undergoing the annotation process to provide activity-centric, spatio-temporal annotations of selected activities.

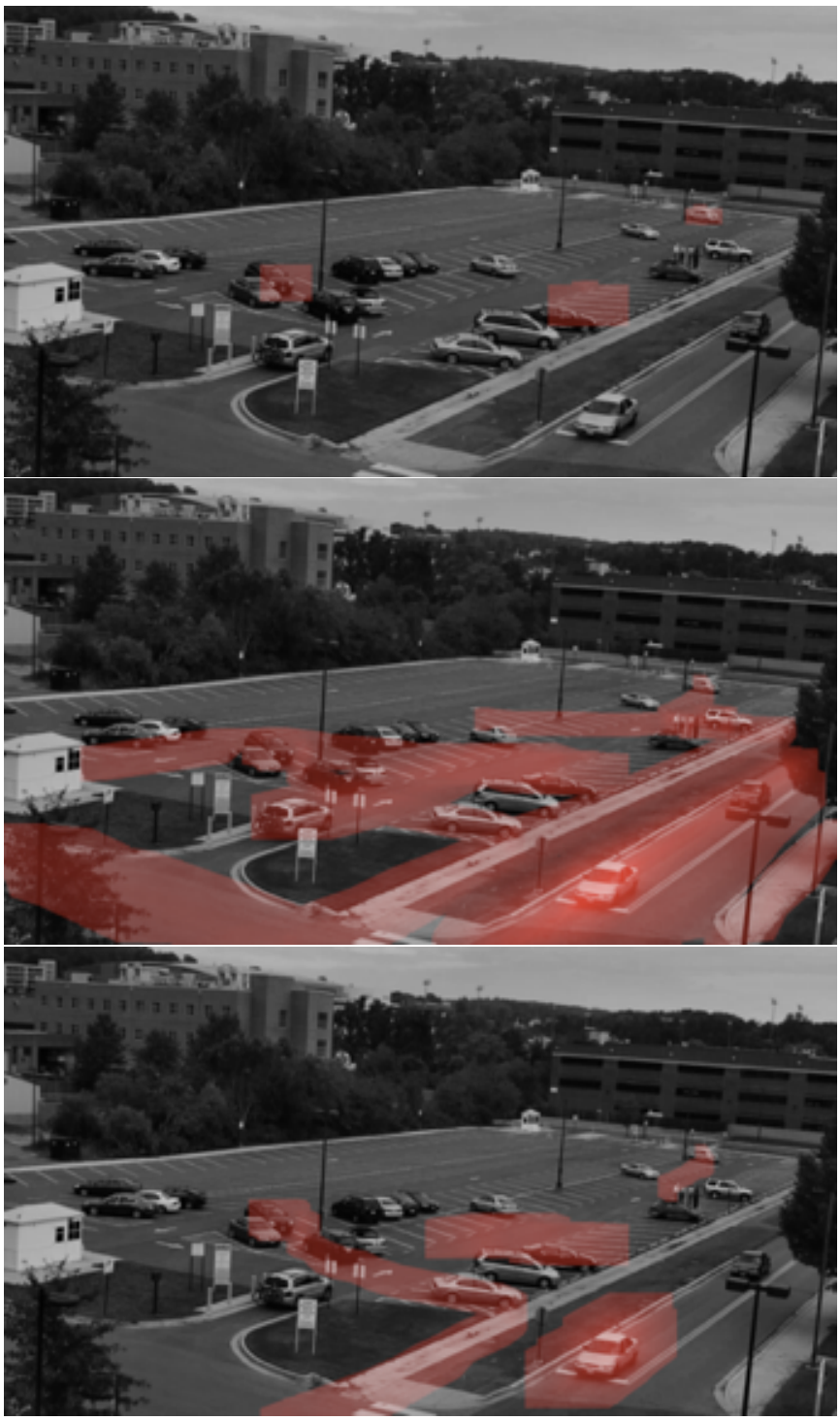

Figure 1. Comparison of VIRAT and DIVA annotations on a 2m13s clip. Top, VIRAT annotated six activities. Middle, DIVA annotated 51 activities. Bottom, after excluding simple person and vehicle motion, DIVA still has annotated 14 activities.

4. Baseline Tasks and Models

The DIVA evaluation plan, available at [4], describes three primary tasks, and one secondary tasks:

- **Activity Detection (AD):** Detect and temporally (but not spatially) localize activities.

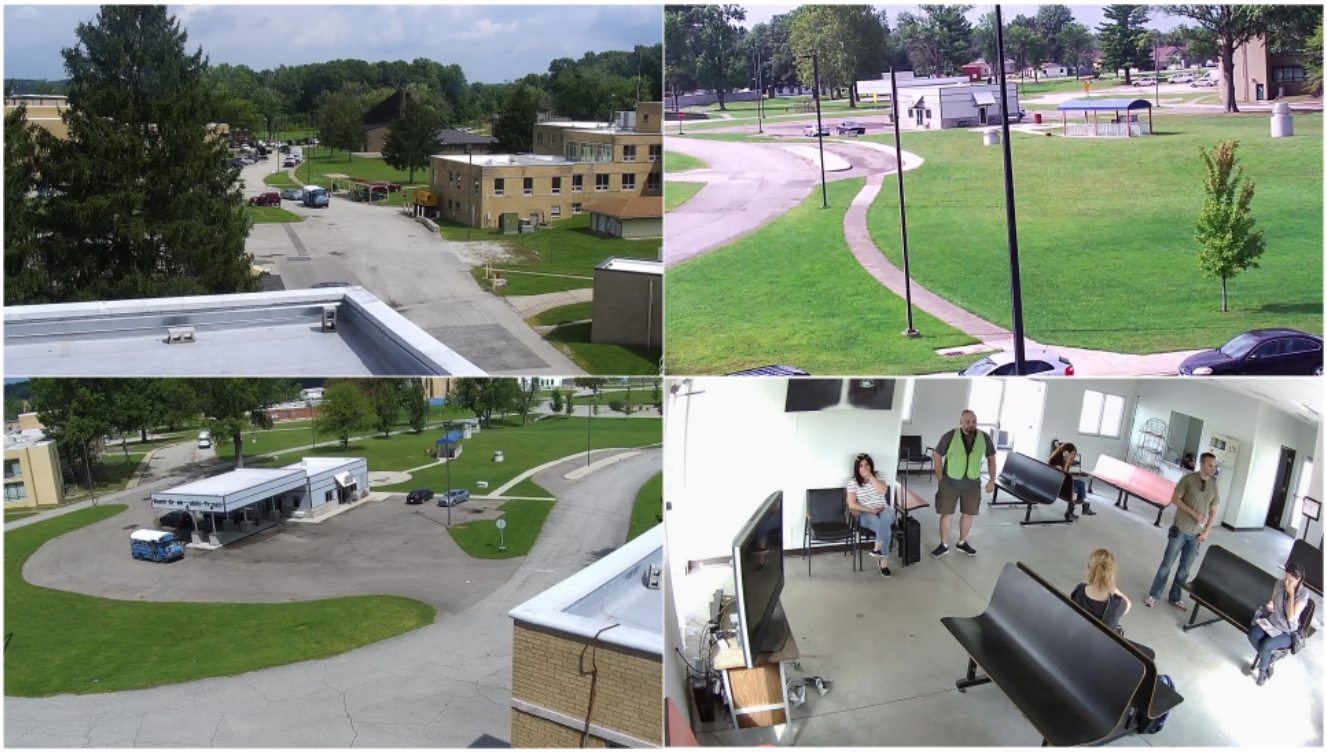

Figure 2. Four images from the M1 Pilot collection

- **Activity and Object Detection (AOD):** Detect and temporally localize activities, as well as detect and spatially localize objects and people associated with the activity.
- **Activity and Object Detection and Tracking (AODT):** Similar to AOD, except the participating objects must be tracked.
- **Reference Temporal Segmentation** *(Secondary)* Given a temporally pre-segmented clip, classify the activity it contains.

To support program evaluation, we are providing baseline implementations for these tasks. We have no mandate to conduct research for these tasks; rather, our role is to identify state-of-the-art algorithms in the academic literature and adapt them as lightly as possible in order to assess their performance on the DIVA tasks. Additionally, these implementations are used to inform the development of the DIVA software framework (Sec 6).

To date, we have implemented algorithms for AD (Sec 4.3), reference temporal segmentation (Sec 4.2), and object detection in partial support of AOD and AODT (Sec 4.1).

4.1. Object Detection

To identify the participants of an activity and obtain near real-time performance, we used YOLOv2 [19] as benchmark object detector. The primary challenge in training a deep network on a surveillance dataset is the number of pixels occupied by the object of interest. Downsampling the image to meet YOLOv2's [19] input resolution further exacerbates this problem. Additionally, the objects tend to have drastically different resolution depending on the scene and distance from camera.

Since YOLOv2 [19] uses regression in a grid cell to localize objects, it tends to perform poorly against small objects that are clustered in a cell of the grid. The model predicts 2 bounding boxes per cell and a cell is responsible for localization if the center of the object falls in the cell. This results in a performance decay that has been discussed in more detail in section 5.2 when compared with PASCAL VOC [10].

4.2. Activity Recognition

We used Convolutional 3D network (C3D) [26] to classify temporally segmented activities. C3D [26] consists of 3D convolution and pooling layers which extract spatial and temporal features for an input video. The model uses 8 images to

determine the activity label. These images are downsampled to 112x112 and treated as a single entity by 3D convolutional network. 3D convolution operation is an extension of conventional 2D convolution, where the third dimension considers the neighboring frames of the input feature map or video. This enables the 3D convolution layer to consider both motion between frames and appearance at the same time. Similarly, the 3D pooling operation performs nonlinear max operation in a spatial and temporal neighborhood.

The two major challenges in training C3D [26] are image resolution of the network and class imbalance in DIVA-V1. C3D uses 112x112 image for training and evaluation, these dimensions are approximately 10 times smaller than the original dimensions. Additionally, similar to object annotations, the activity annotation do not occupy significant portion of the image. Thus C3D experiences similar performance drop as YOLOv2 [19]. To overcome, class imbalance we oversampled the less frequently occurring classes at the cost of generalization of network over these classes. The overall and class-wise scores are available in section 5.3.

4.3. Activity Detection

To temporally localize activities, we used an extension of C3D. Region Convolutional 3D network (RC3D) [27] combines region proposal phase from Faster RCNN [20] with 3D convolution to determine the temporal bounds along with the label for an activity. RC3D defines a set of temporal anchors as a starting point to localize the activity. These anchors slide along the temporal extent of the video to determine if an activity occurs in the region. If activation for an activity is beyond certain threshold, the temporal extents are improved using regression and 3D convolution is used to classify the temporal tubes. R-C3D [27] uses a video buffer of 768 frames and 1 frame is sampled out of 8 frames. Thus the temporal extent of the buffer is 6144 frames. If the video is larger than 6144 frames, temporal window creating the buffer shifts with a stride of 1536. RC3D uses 27 anchors that comprises of all even numbers from 2 to 56 frames. Since these operate on the video buffer, the temporal extent of these anchors vary between 16 to 448 frames.

As with C3D [26], the generalization for RC3D [27] suffers from the class imbalance. However, while in C3D [27] this has to be explicitly addressed using oversampling, RC3D handles class imbalance out of the box. Due to its C3D [26] roots, the image resolution of RC3D is 112x112 which results in significant information loss from downsampling.

5. Experiments

All the models detailed below were trained using the training and validation of DIVA-V1 dataset.

5.1. Evaluation Metric

To evaluate object detectors we used mean average precision (mAP) for the two classes and F1-score. This is the standard metric for object detection and allows for direct comparison of algorithms on multiple dataset to determine the performance change across datasets.

For activity recognition and activity detection, we rely on two metric. The primary metric is probability of missed detection (P_{miss}) at fixed rates of false alarms ($Rate_{FA}$). This is a binary detection measure to analyze the performance trends as the rate of false alarms increases and is suited for surveillance streams. The secondary metric is Normalized Multiple Instance Detection Error (NMIDE) which measures the quality of temporal localization of the activity.

To compute $P_{miss}@Rate_{FA}$, we need to determine the mapping between temporal detection of an algorithm with the ground truth. The metric treats this as a Bipartite matching problem and uses Hungarian solution to find the mapping. The kernel function used to assign score to matches is defined by equation 1.

$$\begin{aligned} K(I_{R_i}, \phi) &= 0 \\ K(\phi, I_{S_j}) &= -1 \\ K(I_{R_i}, I_{S_j}) &= \phi \; if \; Activity(I_{S_j}) \,! = Activity(I_{R_i}) \\ &= \phi \; if IOU(I_{R_i}, I_{S_j}) <= \Delta_{IOU} \\ &= 1 + E_{IOU} * IOU(I_{R_i}, I_{S_j}) + \\ &\quad E_{AP} * AP_c(I_{S_j}) \; , otherwise \end{aligned} \tag{1}$$

where, I_{R_i}, I_{S_j} are ground-truth and activity detector output respectively, IOU is temporal intersection over union that is computed using equation 2, K is the kernel score, $Delta_{IOU} = 0.2$, $E_{IOU} = 10^{-8}$, $E_{AP} = 10^{-6}$ and $AP_c(I_{S_j})$ is computed using equation 3

$$IOU(I_{R_i}, I_{S_j}) = \frac{Intersection(I_{R_i}, I_{S_j})}{Union(I_{R_i}, I_{S_j})} \tag{2}$$

$$AP_c(I_{S_j}) = \frac{AP(I_{S_k}) - AP_{min}(S_{AP})}{AP_{max}(S_D) - AP_{min}(S_{AP})} \tag{3}$$

where, $AP(I_{S_j})$ is the presence confidence of activity I_{S_j}, S_{AP} is an algorithm's confidence score and $AP_{min}(S_{AP})$, $AP_{max}(S_{AP})$ are minimum and maximum values of confidence score in S_{AP}.

Based on the matching, $P_{miss}(\tau)$ and $Rate_{FA}(\tau)$ is given by equation 4 and 5 respectively.

$$P_{miss}(\tau) = \frac{N_{MD}(\tau)}{N_{TrueInstances}} \tag{4}$$

$$Rate_{FA}(\tau) = \frac{N_{FA}(\tau)}{VideoDurationInMinutes} \tag{5}$$

The NMIDE score is computed using equation

$$NMIDE = \frac{\sum_{I=1}^{N_{mapped}} (C_{MD} * \frac{MD_I}{MD_I + CD_I}) + C_{FA} * \frac{FA_I}{V - (MD_I + CD_I + NS_I)}}{N_{mapped}} \tag{6}$$

where, MD_I, FA_I and NS_I are missed detection, false alarms and no score for activity instance I. C_{MD} and C_{FA} are constants with values 1, V is video duration and N_{mapped} number of instances mapped between the algorithm's output and ground truth.

5.2. Object Detection Results

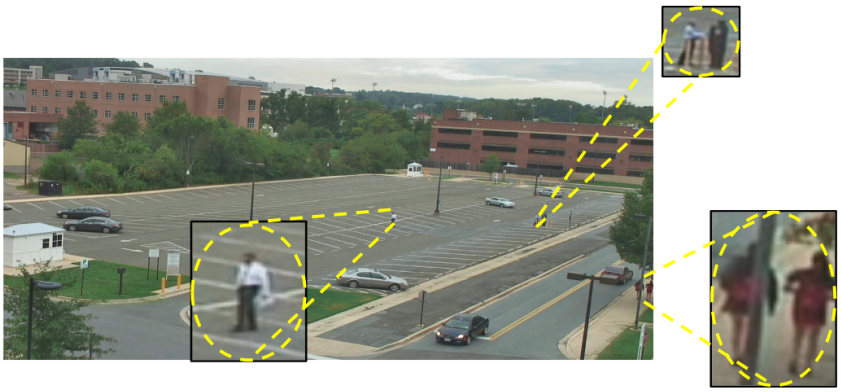

Figure 3. Example of participants in a scene in DIVA-V1

Table 1 shows the significant decrease in performance for both people and car. As discussed earlier, this can be primarily attributed to the size of an object of interest. Figure 3 show an example with 3 individuals in the scene along with the difference in image resolution and object resolution. Although this can be resolved to a certain extent by lowering the confidence threshold. Lowering the confidence threshold tends to increase the false positive detection in an image.

Dataset	Precision		mAP	F1-Score
	Person	Car		
Pascal VOC	0.87	0.87	0.76*	-
DIVA-V1	0.59	0.53	0.56	0.25

Table 1. Performance comparison of YOLO-V2 on Pascal VOC [10] and DIVA-V1 dataset. $*$ represents mAP score for the entire Pascal VOC dataset.

$Rate_{FA}$	0.15	0.1	0.2	1.0	n-mide
P_{miss}	0.997	0.997	0.993	0.958	0.692

Table 2. Performance of C3D on DIVA-V1 dataset

5.3. Activity Recognition Results

Table 2 shows the overall probability of missing detection at different false positive rates along with the n-mide score for C3D [26]. Compared to R-C3D [27], C3D tends to miss more detection across all the frame rates but has a very high n-mide score which suggests that confidence value C3D associates with the detection is much higher R-C3D. The overall low scores would be due to high miss-classification rate at high confidence. We believe this to be due to the short temporal extent that c3d uses to make predictions. Figure 4 show the activity wise breakout of missed detection at different false positive rates.

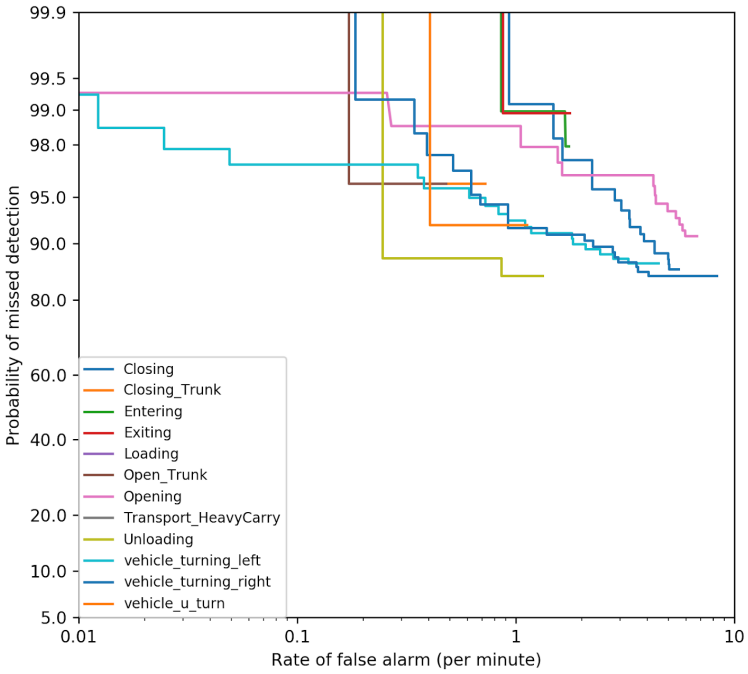

Figure 4. Performance curves for C3D on DIVA-V1 dataset

5.4. Activity Detection Results

Table 3 shows the overall probability of missing detection at different false positive rates along with the n-mide score for RC3D [27]. Low n-mide score suggests that the detector has very low confidence associated with the temporal regions of the activity. This can be attributed to the loss of information that occurs from image downsampling. Although surprisingly it performs better than C3D in the overall evaluation and for every activity that was scored.

$Rate_{FA}$	0.15	0.1	0.2	1.0	n-mide
P_{miss}	0.855	0.866	0.846	0.724	0.214

Table 3. Performance of RC3D on DIVA-V1 dataset

6. Framework

To enable performers to focus on research fundamental to the DIVA tasks, we are developing an open-source framework designed to allow researchers to deploy their algorithms at scale. Our framework is based on KWIVER [3], a cross-platform

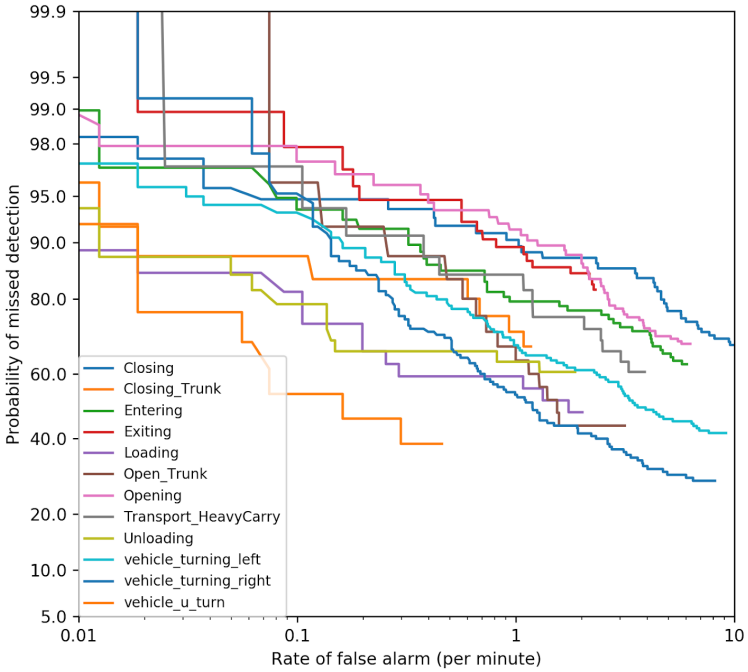

Figure 5. Performance curves for RC3D on DIVA-V1 dataset

computer vision systems toolkit which supplies support for fundamental computer vision data types, processing pipelines, and algorithm implementation dynamically selectable at run-time via shared library loading. Recent work on KWIVER has extended the capabilities to include multiprocessing, using ZeroMQ to transfer data between system nodes. The intent is that by developing against the DIVA framework API, researchers can seamlessly transfer their algorithms from development systems to large-scale deployments. This work is related to the Baseline effort (Sec 4), which are used as prototype algorithms to inform framework development. The DIVA framework may be found on github [2].

7. Conclusion

In this paper, we have presented our work supporting the Test and Evaluation effort on the IARPA DIVA program. In particular, we have described two new datasets focused on activity detection and recognition in surveillance and public safety data, which we hope will become a benchmark reference for the research community. This type of data has video and activity distribution characteristics which differ from typical ”consumer-oriented” video of the sort normally addressed in the literature, motivating the need for a new dataset. We have additionally described our work implementing baseline algorithms to establish reference performance levels against this data. Portions of this data are being used in the ongoing NIST ActEV challenge, in whose leaderboard our baseline algorithms participate. Finally, to support transition of research products to large-scale evaluation and deployment environment, we discussed our development of an open-source software framework for multi-camera video processing and analytics.

References

[1] Deep Intermodal Video Analytics (DIVA). `https://www.iarpa.gov/index.php/research-programs/diva`.

[2] DIVA framework. `https://github.com/Kitware/DIVA`.

[3] KWIVER: The Kitware Image and Video Exploitation and Retrieval toolkit, howpublished = `http://www.kwiver.org`.

[4] TRECVID 2018: Activities in Extended Video (ActEV). `https://actev.nist.gov/`.

[5] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[6] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet. Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[7] A. Bannat, J. Gast, G. Rigoll, and F. Wallhoff. Event analysis and interpretation of human activity for augmented reality-based assistant systems. In *Intelligent Computer Communication and Processing, 2008. ICCP 2008. 4th International Conference on*, pages 1–8. IEEE, 2008.

[8] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[9] T. Chavdarova, P. Baqu, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[11] R. B. Fisher. The pets04 surveillance ground-truth data sets. In *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–5, 2004.

[12] A. Hoogs, A. G. A. Perera, R. Collins, and et al. An end-to-end system for content-based video retrieval using behavior, actions, and appearance with interactive query refinement. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Aug 2015.

[13] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] S. Oh, A. Hoogs, A. Perera, and et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 3153–3160. IEEE, 2011.

[17] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):677–695, 1997.

[18] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[19] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[21] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[23] M. Schröder and H. Ritter. Deep learning for action recognition in augmented reality assistance systems. In *ACM SIGGRAPH 2017 Posters*, page 75. ACM, 2017.

[24] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[25] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[27] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.