From Leaderboard To Operations: DIVA Transition Experiences

Bharadwaj Ravichandran, Roderic Collins, Keith Fieldhouse, Kellie Corona, Anthony Hoogs
{firstname.lastname}@kitware.com
Kitware Inc.
1712 Route 9, Suite 300, Clifton Park, NY 12065 USA

Abstract

The IARPA Deep Intermodal Video Analytics (DIVA) program has sponsored the development of systems that detect and recognize activities in security video. During the period from September 2017 to March 2021, the development and evaluation of these systems was focused on optimizing accuracy, embodied in quantified metrics, against a large but relatively static corpus of video collected and annotated by the program. This focus was aided by various software engineering decisions collaboratively reached by the program performers and Test & Evaluation (T&E) team, which established a common software framework enabling ongoing quantitative evaluation via software submissions to a leaderboard.

While continuing to support the leaderboard, in March 2021 the program began efforts, still in progress, to transition capabilities developed on DIVA from the research environment to operational evaluation and deployment. As an operational system is a different use case than a research environment, it is not surprising that design decisions favoring the former will not always align with the latter. This paper discusses our work to transition DIVA systems into an operational setting, particularly identifying and resolving conflicts between the evaluation framework and operational requirements. We describe transition efforts to date, propose future work, and conclude with lessons learned from the overall transition effort.

1. Introduction

It has been estimated that in 2019, 180 million security cameras were shipped worldwide [9], while the attention span of a human camera operator has been estimated at only 20 minutes of continuous manual monitoring [4, 3]. The gap between the massive volume of data available and the scarce capacity of human analysts has been closed, but not eliminated, by the rapid advancement of computer vision techniques, particularly deep-learning based methods.

The IARPA Deep Intermodal Video Analytics (DIVA) program [5] seeks to address this imbalance by sponsoring research into systems that detect and recognize activities in security and public safety video. Since the program's start in September 2017, teams from academia and industry have been developing systems capable of processing multiple videos at or near real-time. These systems have been evaluated by NIST via the Activities in Extended Video (ActEV) leaderboard [11], which computes the system's probability of missing an activity against its time-based false alarm rate [13] for a suite of 37 activities defined by the DIVA program and annotated across a very large video collection [6].

This focus on algorithmic performance has spurred development of systems whose capabilities warrant consideration by transition partners in operational settings. This same focus, however, has meant that other aspects of system development required for transition have not received the same attention. In particular, the partner assessment paradigm differs from the ActEV evaluation in one crucial respect: **partners may only have bandwidth to assess a small number of systems in an open-ended, interactive setting, whereas ActEV conducts numerous evaluation loops, computing a static suite of metrics on a large video data corpus processed in batch mode.** Partners may have their own use-cases and datasets which differ from those used to generally advance the state of the art.

Training systems for new activities, a crucial requirement for transition partners, has been a particularly tricky issue. Moving from a closed-universe design of 37 activities to an open-ended, trainable concept of operations exposed assumptions and dependencies at all levels of the DIVA ecosystem: data, annotations, and system interfaces. Section 5 describes how we assessed trainability; ideas for further support are discussed in Section 7.

In this paper, we report on our experiences adapting DIVA performer systems for assessment by transition partners, ranging from low-level items such as hard-coded constants and file formats, to more complex questions of enabling local system training by users with minimal machine learning experience. The paper is organized as follows: Section 2 briefly describes the DIVA task and evaluation

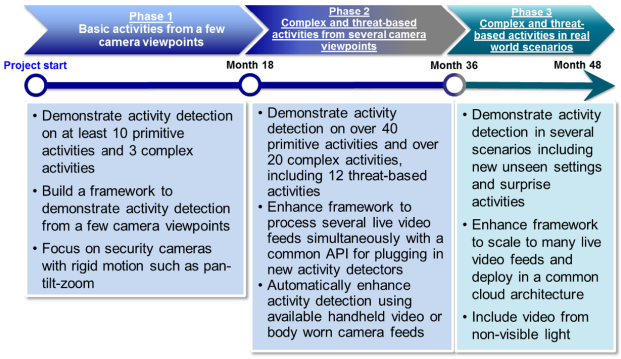

Figure 1. DIVA program goals; from the Broad Area Announcement (BAA) available via [5].

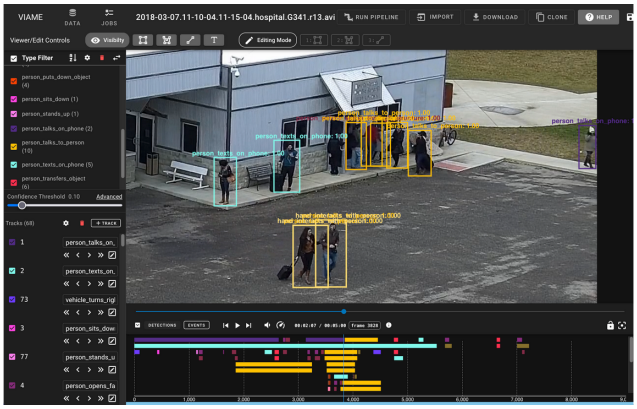

Figure 2. MEVA footage with ground-truth, in the DIVE [7] viewer. Spatial output is visualized in the main panel; the timeline in the bottom provides a temporal-only view of the data.

data. Section 3 describes the evaluation software environment developed by NIST in which performer systems are delivered. Section 4 details challenges and solutions to adapting systems for operational use; training systems for new activities is discussed in Section 5. Our transition efforts to date are described in Section 6, followed by possible future work in Section 7 and lessons learned in Section 8.

2. DIVA Program Overview

The fundamental task for DIVA performers is to identify activities in video streams. In support of this, the program Test and Evaluation (T&E) team has collected and annotated video data, established evaluation metrics, and implemented a leaderboard [11]. The program's goals, Figure 1, emphasize research in activity detection throughout the program, and also imply significant system engineering in the later phases. Since June 2019, the evaluations have been based on the "Known Facility" (KF) and "Unknown Facility" (UF) datasets. In the KF scenario, performers have access to sample footage from the dataset, annotation samples, and metadata such as camera models. In the UF scenario, no such prior knowledge is provided. In both cases, the video is electro-optical (EO) 1080p, 30Hz video from commercial off-the-shelf security cameras. The KF data also includes video from thermal IR cameras (352x240, 30Hz).

The DIVA program metrics [13] defined three tasks: **AD** (activity detection), **AOD**, (activity and object detection), and **AODT** (activity and object detection and tracking). The ActEV leaderboard has focused on the AD task, which is temporal-only: performers are required to only report *when* the activity is happening, not *where*. (This is analogous to the timeline view in the bottom panel of Figure 2.) The system bandwidth for communicating temporal-only intervals is significantly lower than for spatio-temporal intervals, a design decision addressed in Section 4.

Portions of the KF dataset have been released as the Multiview Extended Video with Activities (MEVA) dataset [6], whose collection and annotation are described in [2]. (Procedures for collecting and annotating the UF dataset were nearly identical). Annotation starts with the activity definitions [8], which specify when an activity starts and stops, and what actors and props must be annotated. All DIVA program annotations include bounding boxes, as shown in the central pane of the viewer in Figure 2.

The UF evaluation is intended to facilitate transition by providing feedback for how systems perform on wholly unknown data. As of this writing, for a time-based false alarm point of 0.02, teams tend to have a probability-of-missed-detection about 0.1 to 0.2 higher for UF than KF. The need to keep the data sequestered has limited the ability of the teams to respond to this feedback, although general guidance regarding differences between KF and UF has been provided by the program office.

3. The ActEV CLI

DIVA's evaluation plan relies on running performer systems on sequestered videos and comparing the output to sequestered annotations. NIST defined the ActEV Command Line Interface (CLI) [12] to facilitate performers submitting their systems for execution and evaluation in a sequestered environment. The CLI makes few assumptions about how the systems are designed, and implements a state machine for downloading, building, and validating the system. It also specifies a set of executables the system must provide for initializing, running, and closing out experiments. Each experiment processes the evaluation video set (which may include hundreds or thousands of five-minute video clips) as a set of *chunks*, whose size the system is responsible for selecting. Systems must support chunk-level setup, processing, and cleanup; this allows the CLI to run a large number of clips across many chunks while minimizing the risk of lost work should the system crash. It also permits a degree of parallelization.

Several desiderata for the ActEV CLI's goal of leader-

board support are, unsurprisingly, at odds with those sought by transition partners:

1. **Installation.** The CLI allows for varied and changing system implementations. In particular, the use of container technology such as Docker [10] was not required. As a result, performer system implementation technologies varied widely.
2. **Bias for long runs.** The systems must support the chunking paradigm, which offers little value for interactive runs of a few clips.
3. **Sequestered evaluation.** To support sequestered evaluation, the systems are designed to require no external network connectivity after the validation stage.
4. **Scoring support.** System output includes chunking status and is optimized for the scoring system, rather than interactive display.

These design decisions, while entirely appropriate for evaluating multiple research systems, required adjustment when supplying individual systems to transition partners for assessment.

4. Transition Overview

The goal of our transition efforts is to stand up DIVA systems in the partner's environment of choice, typically a lab setting with access to hardware comparable to the BAA baseline described in Section 6.1.

4.1. Challenges

As mentioned in Section 1, partners assess systems differently than the DIVA program at large evaluates them. Some of these differences include:

Installation cycle. The CLI re-installs the system each time; partners typically prefer to install a system once.

Datasets. The DIVA program data was collected in accordance with human subject research protocols to be acceptable for academic and industry research use. Operational data may be restricted.

Activity suite. The DIVA program selected 37 publicly-known activities for performers to focus on; around December 2020, an additional evaluation protocol allowed for so-called "surprise" activities to be presented to the system during sequestered execution for few-shot learning. But partners may be interested in entirely different activities than those selected by the DIVA program.

Activity concept. A related issue is that partners may target activities whose scope falls outside those of the program evaluation, for example, a long-duration complex activity with multiple actors.

Training. A related issue is how a transition partner would train the system for a new activity. This is discussed in detail in Section 5, in particular the issue of **one-vs-many**: DIVA systems have always operated in multi-class paradigm, but partners may be interested in only a single activity, which has implications for how the system is trained.

Evaluation. Transition partners, at least during initial assessment, are more interested in rapid qualitative assessment on a few videos than on comprehensive statistics on a large corpus of partially sequestered data. Partners will require **interactivity** in the form of a GUI used to visualize results on their data.

Related to evaluation, the issue of **file formats** arose once the need to keep and transmit spatiotemporal results arose. The formats written by the system and read by the scoring code are designed to organize results from batch runs of hundreds or thousands of videos, and are not always efficient for interactive use.

4.2. Solutions

These thematic differences have systemic implications both large and small. Actions and issues addressed to date include:

A "few-shot" installation script supplied by NIST to simplify the download and installation of an ActEV CLI system. As described in Section 6.1, this addresses some, but not all, of the installation issues; further work remains before systems are easily installable.

Data modality. The DIVA data is all h264 in an AVI container at 30Hz, but the same may not be true of partner data. We have developed a suite of videos with various codecs, containers, bitrates, etc. to map out what systems can accept.

Low-level software issues. Writing software is an iterative process not unlike fitting data to a model, and is similarly prone to over-fitting its test environment. We have identified and addressed low-level assumptions such as hard-coded pathnames for videos and classification vectors with a fixed length of 37 (to match the long-standing set of 37 public activities.) Such issues are identified more from simply varying the installation environment, rather than from any particular partner requirement.

Visualization. We are adapting our open-source video analytics and visualization tool - DIVE [7] - to work with DIVA, as seen in Figure 3. This enables us to create an overlay of the predicted activity detections over the corresponding ground truth detections.

Annotating partner data. When possible, we have accepted partner data and annotated it for events wholly novel to the DIVA program, enabling assessment of the trainable-system capabilities described in Section 5 on operationally relevant data.

5. Transitioning Trainability

All of the DIVA performer systems are based on contemporary deep-learning algorithms, which require extensive training on labeled data. For the primary DIVA metric

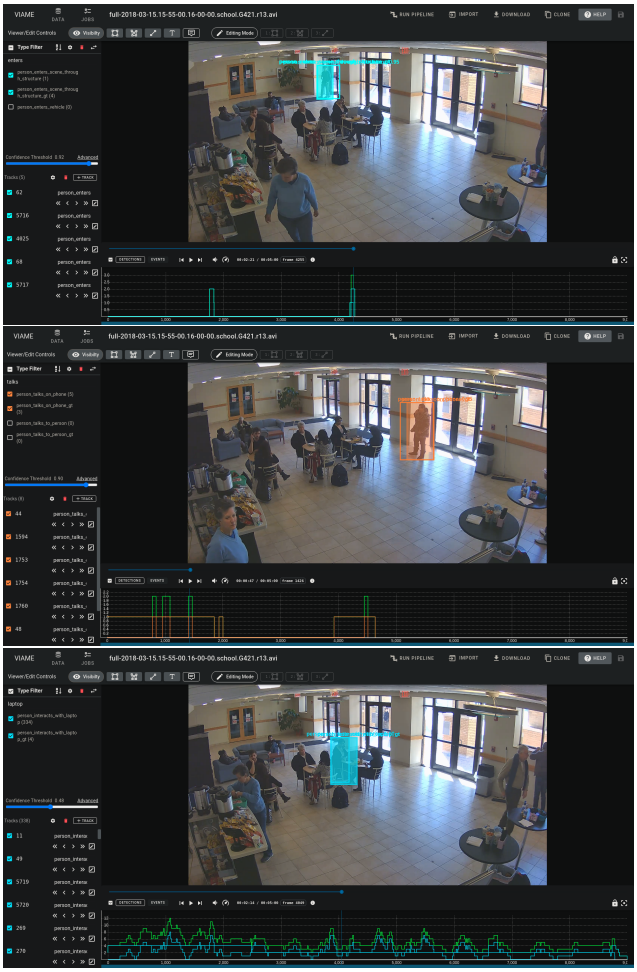

Figure 3. DIVE [7] visualization examples from a DIVA performer system's activity predictions (non-filled bounding box) vs ground truth data (bounding box with color mask); Top: Visualization of activity "person_enters_scene_through_structure"; Middle: Visualization of activity "person_talks_on_phone"; Bottom: Visualization of activity "person_interacts_with_laptop".

of performance on the publicly known set of 37 activities, performers are free to iterate and fine-tune the algorithm models before submission to the ActEV leaderboard. In contrast, an operational deployment will require clear procedures for training the system to recognize new activities.

With a similar goal to that of the UF evaluation to simulate the unforseen nature of data in an operational environment, the ActEV challenge defines a "surprise activity" protocol where, during evaluation, systems implement few-shot learning to recognize a new activity given a textual description and a small set of exemplars. Performance on this task is measured on the leaderboard under "Surprise Activities."

From the transition perspective, we are interested in two additional aspects of trainability: **reproducibility**, to verify that the T&E team can reproduce the trained model the

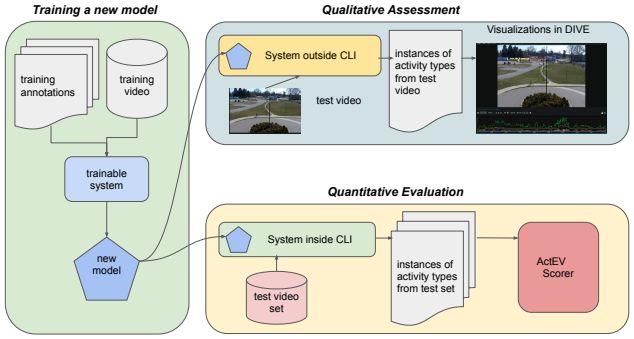

Figure 4. Overview of the trainable system validation workflow. On the left, a new model is trained; on top right, the results are qualitatively visualized; on bottom right, the new model is quantitatively scored.

performers submit to the leaderboard, and **extensibility**, to measure how new activities can be added to the system in situations more flexible than that of the few-shot protocol but more constrained than the open-ended research environment.

To measure reproducibility, the T&E team obtained a description of the training data and protocol from the performer and independently trained the system. As shown in Figure 4, the locally-trained system was then scored using the same procedure as that for the leaderboard for both the KF (Known Facility) and UF (Unknown Facility) datasets. If the scores agreed to within roughly 0.1%, the test was considered to have passed.

Measuring extensibility, that is, the ability of transition partners to train a system on totally new activity types with no assistance from T&E or the performer, required performers to specify a training protocol which could be executed by somebody with minimal machine learning experience. We evaluated extensibility by applying the performer's protocol to the same data used for testing reproducibility, which would contain a variable (and, to the performer, unknown) number of surprise activity instances. The T&E team has access to the surprise annotations in the training data, and thus can train on them. We evaluate against the UF microset, containing both known and unknown activities, and check two results:

- The results for the 37 known activities were compared to the leaderboard results, with the expectation that results should be similar.
- The results for the 37 known activities were compared to those for the 10 surprise activities, with the expectation that the average results should be roughly similar.

If both comparisons passed, the extensibility protocol was judged to be suitable for use by transition partners for training on non-DIVA activities.

One aspect which was highlighted by a transition partner's experiences was DIVA's bias towards multiclass classification, rather than binary classification. Training a system requires both positive and negative examples of the activity in question. When training for the multiclass 37 public activities, each activity implicitly has negative examples available from the other 36 activities. However, when a transition partner wished to train DIVA for a **single** activity, the annotations they generated provided no means for obtaining negative examples. Resolving this issue required iterating with the teams to properly document and test the binary classification use case.

6. Transition Efforts

6.1. AWS usage

At the start of the DIVA program, a $10,000USD reference system was proposed as the hardware baseline. Based on this, at the end of Phase 3, the common specification used by all performers and the T&E team was a 4-GPU system with at least 12GB of VRAM in each GPU. To ensure reproducibility, we required systems which could be easily installed from scratch for each testing run. Hence, we chose to use p3.8xlarge [1] AWS EC2 instances for testing DIVA systems. A "Deep Learning Base AMI" image with Ubuntu 18.04 is used as the base image; this allows each instance to boot up pre-installed with the required deep learning dependencies such as NVIDIA-CUDA drivers and libraries (including Docker). The p3.8xlarge specification provides 4×16GB V100 GPUs, 244 GB RAM and 32 virtual CPUs. The storage is specified separately during instance creation. By default, each instance is started with 1TB of storage space and additional storage is added based on demand.

After setting up an instance, an ActEV DIVA pretrained performer system is downloaded, installed and executed as part of a dry test run. The entire process from downloading to execution is controlled by the ActEV CLI's "few-shot" script which consists of the following steps:

1. **ActEV CLI download**. The pretrained system is downloaded and the latest git version of the CLI is checked out.
2. **ActEV CLI install**. The installation sets up the ActEV CLI commands and the necessary python virtual environments for evaluation and scoring.
3. **MEVA test set download**. An ActEV set consisting of 60 MEVA videos is downloaded as a test set.
4. **ActEV CLI execution**. The final step is the data chunk processing, evaluation and scoring using the ActEV test set as input to the pretrained system.

Through this exercise, we can obtain a working DIVA pretrained system that has cleared all the necessary dependency checks and is able to produce activity detection output and DET curve visualizations on a set of MEVA videos. If there are no dependency issues, these steps will run end-to-end within 60 minutes.

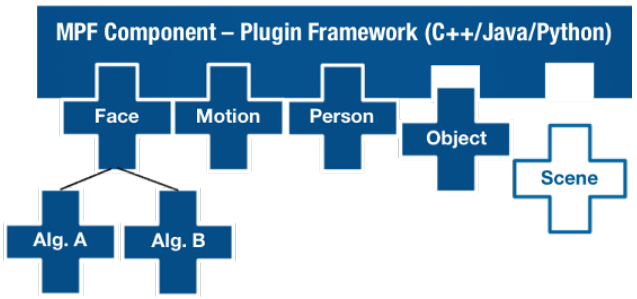

Figure 5. OpenMPF plugin architecture. Figure from [14].

6.2. OpenMPF

The Open Media Processing Framework, or OpenMPF [14], is a popular open-source framework for media analytics. OpenMPF consists of a set of APIs to help integrate standalone systems into an open-source, customizable and intuitive end-to-end pipeline. Figure 5 illustrates OpenMPF's plugin architecture.

To demonstrate DIVA's compatibility with OpenMPF, we implemented an OpenMPF pipeline to generate activity detections, which were visualized using OpenMPF's video overlay mechanism. We tested the pipeline on a specific performer's pretrained system. As part of the implementation, the system was unwrapped from the ActEV CLI and integrated into the OpenMPF framework. The input to the pipeline was either a single video or a file index consisting of a list of videos. The system evaluates each video and generates activity detections with corresponding spatial localizations, writing output to a JSON file. Finally, the JSON file is processed by the OpenCV libraries within OpenMPF which generates a video output with the detected activity bounding box overlays. This is a rapid way of building a prototype framework to test DIVA systems on partner data.

6.3. Transition Customer Experiences

Here we discuss various issues that have arisen in the process of assisting transition partners evaluate DIVA systems at their local sites.

Dependency checks. The task of passing the dependency checks for a DIVA system is one of hardest challenges transition partners face. As discussed in Section 6.1, the default spec specifies the use of at least 4 GPUs. Some partners had access to AWS, but others were using a single-GPU system, which required changes to the CLI's configuration. We also encountered issues configuring python virtual environments. Live debugging support was necessary to address all such issues.

Video format. The MEVA dataset consists of h.264 videos in avi containers recorded at 30Hz. The DIVA systems were built under the assumption that all input videos are at 30Hz. One of our partner's data was in the correct format, but recorded at 60Hz, which produced low-quality detections when run through a DIVA system.

Customer data vs DIVA data. Partners typically collect and annotate much lower quantities of video than what the DIVA program has made available. As discussed in Section 5, DIVA systems are designed and optimized to predict the public 37 activities, whereas some partners seek to train the same system for a single activity.

Machine Learning expertise. Operational environments may or may not be staffed by ML engineers, and one of our goals is to identify when ML expertise is or is not required. Transition partners would expect the system documentation to include appropriate guidance about when and how parameters should be changed. Based on documentation from the performer teams and our experience training and testing the systems, it is our responsibility as ML experts to provide transition documentation that explains research parameters for fine-tuning system performance at the customer's end.

6.4. DIVA-like partner data

Typically, the robustness of a trained deep learning system is verified by evaluating the system across unseen test sets [15]. In addition, a system which converges faster during training on new data and activities by learning from transferred pretrained weights [16] provides preliminary evidence that the system can be fine-tuned for different types of data. To test this theory, we obtained DIVA-like video from a potential transition partner. We annotated additional activities which are not part of the public and surprise DIVA activities set. The partner data is in a similar style to DIVA data, and tests are ongoing.

7. Future work

As of this writing, both the DIVA program and our transition efforts are ongoing. As we gain experience mediating between the systems developed by the program teams and the requirements of transition partners, future work includes:

Containerization. To maximize performer flexibility at the start of the program, the CLI did not require containerization. Systems such as Docker [10] simplify installation and library dependencies, and can smooth certification and authorization processes in customer facilities. Redeploying performer systems in a Docker-like environment will likely require a moderate amount of systems engineering, greatly assisted by the familiarity we have gained with the systems as we worked through the set of tasks described in Sections 6 and 5.

Continued visualization improvements. In particular, DIVE's data model is confined to detections and tracks; there is no concept of a multi-track activity, nor is there any capability to have hierarchies of activities. Other improvements include convenient visualizations of multiple datasets in a single session, for example, system output and ground-truth.

Enhancing trainability. As discussed in Section 6.4, we are investigating how a research system's trained model can be fine-tuned on operational data, rather than trained from scratch; this would reduce the equipment and resource burden for deploying a DIVA system into a new environment.

Development of partner datasets. As partners work with the DIVA systems, we will continue, when possible, to advise and assist with the development of annotations on their data.

Close to the end of the program, we envision a final DIVA transition system to have an architecture similar to Figure 6.

8. Conclusions and Lessons Learned

In this report, we have described our efforts to take research systems designed to solve the IARPA DIVA problem of activity recognition in a multi-camera environment, and assist in their transition to government partners. In collaboration with the T&E team, the DIVA performers have produced systems which can be delivered, installed, run, and evaluated in a sequestered environment; this is a significant achievement and is a testimony to the efforts of all involved. Adjusting the systems to deal with the inevitable differences in system implementation, use case, and data between evaluation and operation has acted as a forcing function to improve documentation, clarify corner cases, and as such has already yielded increased robustness.

Some of the lessons, we feel, that have emerged from this effort, might be applied to future end-to-end system research efforts that include:

Containerization. When DIVA started, technologies such as Docker were still nascent; their use in contemporary and future programs should be less controversial. We did receive feedback during the program that Docker was forbidden at some performer team sites as its requirements for administrative access conflicted with local security policies; development of a best-practices document to identify and resolve such issues would be a worthwhile endeavor.

Common Program I/O Implementation. Related to containerization, to avoid performers having to re-invent the wheels of reading videos and ensuring their output routines conformed to program schemata, future programs could supply a Docker base image with video I/O and program output pre-implemented. This would facilitate testing and allow a measure of transition work to continue in parallel as a preliminary or "dummy" system could be supplied to

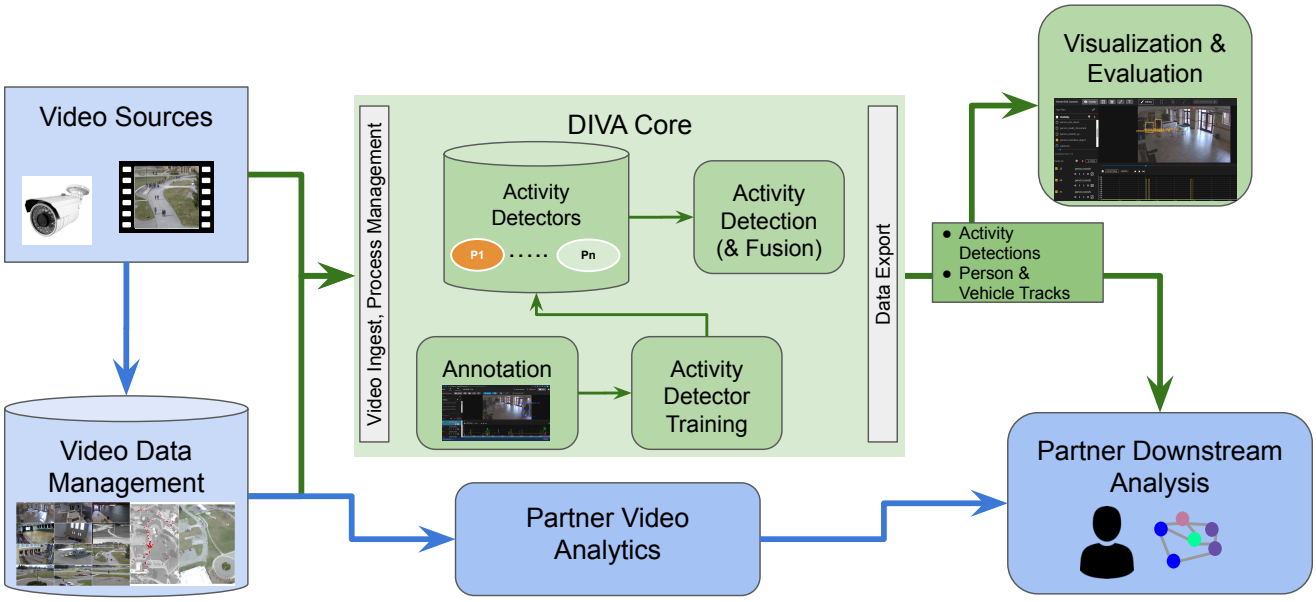

Figure 6. Proposed architecture for a deployed DIVA transition system.

partners while the program is in progress, allowing early identification of any data or output issues.

Incorporating the "data cliff" into the evaluation. As a system transitions from research to deployment, it notionally falls off a "data cliff" where the developers can no longer adjust how the system responds to new data. Operational data may vary from program data at many levels: different frame and compression rates, different codecs, different camera angles, greater variety in weather conditions, etc. Making a system robust to these varying factors is a combination of research and engineering; the DIVA program attempted to stimulate both by the inclusion of the UF evaluation and the surprise activity protocol.

Early incorporation of partner use cases. An example from DIVA would be developing an evaluation component that only identified a single activity, unknown to the performer at submission time, to expose issues such as how hard negative examples are obtained.

9. Acknowledgements

The authors would like to thank both the DIVA performers and T&E team for their unstinting and generous support over the past four years.

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via contract 2017-16110300001. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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 or the U.S. Government.

References

[1] Amazon Web Services. Amazon EC2 P3. https://aws.amazon.com/ec2/instance-types/p3/.

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

[3] Mary W Green. *The appropriate and effective use of security technologies in US schools: A guide for schools and law enforcement agencies*. US Department of Justice, Office of Justice Programs, National Institute of . . . , 1999.

[4] Niels Haering, Péter L Venetianer, and Alan Lipton. The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19(5-6):279–290, 2008.

[5] IARPA. DIVA: Deep Intermodal Video Analytics. https://www.iarpa.gov/research-programs/diva.

[6] Kitware Inc. Multiview Extended Video with Activity dataset. http://mevadata.org.

[7] Kitware Inc. DIVE. https://kitware.github.io/dive/.

[8] Kitware Inc. MEVA Annotation Definitions. https://gitlab.kitware.com/meva/meva-data-repo/blob/master/documents/MEVA-Annotation-Definitions.pdf.

[9] IHS Markit. Security technologies top trends for 2019. https://cdn.ihs.com/www/pdf/1218/IHSMarkit-Security-Technologies-Trends-2019.pdf.

[10] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

[11] National Institute of Standards and Technology. ActEV: Activities in Extended Video. https://www.nist.gov/programs-projects/activities-extended-videos-evaluation.

[12] National Institute of Standards and Technology. ActEV SDL CLI. https://actev.nist.gov/pub/ActEV_SDL_CLI_20190611.pdf.

[13] National Institute of Standards and Technology. ActEV Sequestered Data Leaderboard (SDL) Evaluation Plan. https://actev.nist.gov/pub/ActEV_SDL_EvaluationPlan_Feb_28_2020.pdf.

[14] OpenMPF. Open media processing framework. https://openmpf.github.io/.

[15] Benjamin Taskar, Ming Fai Wong, and Daphne Koller. Learning on the test data: Leveraging unseen features. In *ICML*, pages 744–751, 2003.

[16] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.