Demo: Video Analytics - Killer App for Edge Computing

Ganesh Ananthanarayanan, Victor Bahl, Landon Cox, Alex Crown, Shadi Nogbahi, Yuanchao Shu
Microsoft Research

CCS CONCEPTS

• **Computer systems organization** → **Distributed architectures**; *Cloud computing*; • **Computing methodologies** → *Computer vision tasks*; • **Information systems** → Data analytics; • **Networks** → Network algorithms.

KEYWORDS

edge computing; video analytics; camera; DNN; cloud

ACM Reference Format:
Ganesh Ananthanarayanan, Victor Bahl, Landon Cox, Alex Crown, Shadi Nogbahi, Yuanchao Shu. 2019. Demo: Video Analytics - Killer App for Edge Computing. In *The 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '19), June 17–21, 2019, Seoul, Republic of Korea.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3307334.3328589

1 INTRODUCTION

The world is witnessing an unprecedented increase in camera deployment. The USA and UK, for instance, have one camera for every 8 people. Video analytics from these cameras are becoming more and more pervasive, exerting important functions on a wide range of verticals including manufacturing, transportation, and retails. While vision techniques have seen considerable advancement, they have come at the expense of compute and network cost.

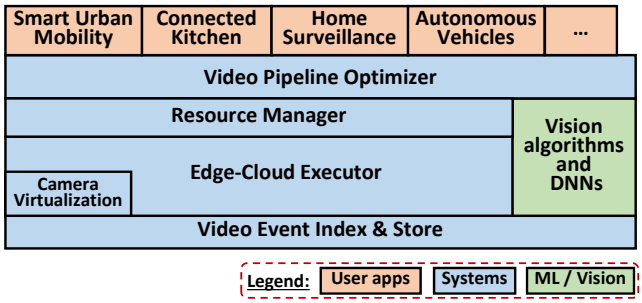


Figure 1: Rocket Video Analytics Stack.

As an alternative to the centralized, in-the-cloud compute paradigm, edge computing offers the promise of near real-time insights, faster localized actions, and cost reduction because of efficient data management and operations. We believe video analytics may represent the "killer application" for edge computing due to its demanding requirements on compute, bandwidth and latency. In this demo, we showcase a live video analytics system (Rocket; see Figure 1) that spans across the cloud and edge, with low cost, and produces results with high accuracy. We highlight three main aspects.

1) Our hybrid architecture intelligently splits the video analytics between the edge and the cloud, along with a cascaded mode of operation that uses CPU-based operators to invoke expensive models selectively.

2) We use resource-accuracy tradeoff in video analytics with multi-dimensional configurations for scheduling.

3) We piggyback on the live video analytics to intelligently generate an index of the video frames that enable interactive querying for after-the-fact analysis.

2 ROCKET VIDEO ANALYTICS STACK

Figure 1 shows our video analytics stack, Rocket [1], that supports multiple applications including traffic camera analytics for smart cities, retail store intelligence scenarios, and home assistants. The "queries" of these applications are converted into a pipeline of vision modules by the *video pipeline optimizer* to process live video streams. The video pipeline consists of multiple modules including the decoder, background subtractor, and deep neural network (DNN) models. Figure 2 shows an example video analytics pipeline that we use for our pilot deployments in smart cities (with the City of Bellevue and others), retail monitoring, and connected kitchens with a large fast-food store (details in §3).

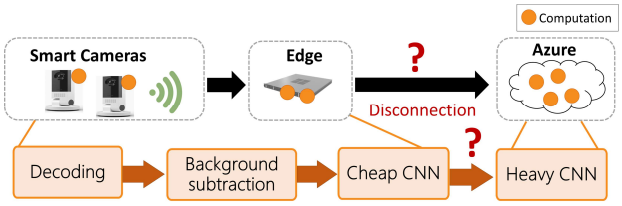


Figure 2: Video analytics pipeline with cascaded operators executing across the edge and cloud.

Cascaded Operators: Intrinsic to the pipeline in Figure 2 is a *cascade of operators with increasing cost.* The background subtraction module detects changes in each frame and can be run even on CPUs at full frame rate of HD videos. If this module notices a change in the region of interest, it invokes a lightweight DNN model (e.g., tiny Yolo [8]) that checks if there is indeed an object of the queried type (e.g., we may be looking only for cars). Only if the lightweight DNN model does not have sufficient confidence do we invoke the heavy DNN model (e.g., full YoloV3 [8]). Such cascading leads to judicious usage of the expensive resources like GPUs.

Edge-cloud distributor: Rocket *partitions* the video pipeline *across the edge and the cloud.* For instance, it is preferable to run

MobiSys '19, June 17–21, 2019, Seoul, Republic of Korea

ACM ISBN 978-1-4503-6661-8/19/06.
https://doi.org/10.1145/3307334.3328589

the heavy DNN on the cloud where the resources are plentiful; see Figure 2. Rocket's edge-cloud partitioning ensures that: (*i*) the compute (CPU and GPU) on the edge device is not overloaded and only used for cheap filtering, (*ii*) the data sent between the edge and the cloud does not overload the network link; see [3].

Network Unavailability: We periodically check the connectivity to the cloud and fall back to an "edge-only" mode when disconnected. This avoids any disruption to the video analytics but may produce outputs of lower accuracy due to relying only on lightweight models.

Resource-accuracy trade-off: The resource manager in Rocket *trades off between resource usage and accuracy* of the outputs by smartly choosing the "configurations" of the video analytics. Configurations are multi-dimensional including the choice of frame resolution, frame rate, and which DNN model(s) to use (both the lightweight and heavy models) [4]. The configuration choice has considerable impact on the resource usage of the video pipeline as well as the accuracy of the output produced. Processing videos at low frame rate by sampling off frames and using DNNs with many convolutional layers stripped out drastically reduces the compute needed but at the expense of lower accuracy in the detected objects. The resource manager allocates resources to competing pipelines such as to maximize the average accuracy of the outputs [5].

Efficient cross-camera analytics: The video analytics stack also features a spatio-temporal profiler that learns spatio-temporal correlations in camera networks. In scenarios where large networks of cameras are being deployed, applications like object tracking can largely benefit from these correlations by narrowing down the inference-time search toward cameras and frames most likely to contain the query identity, thus substantially cutting down inference workload [9]. In Rocket stack, the profiler learns the model on historical data, and updates the model on-the-fly when it detects model variations.

Querying stored videos: Finally, we piggyback on the live video analytics to use its results as an *index* for after-the-fact *interactive querying on stored videos*. Specifically, we support asks of the form, *find frames with red car in the last week*. We answer such asks without processing a week's volume of videos because the live video analytics allows us to generate an index of frames in which objects (e.g., red car) occurs. We also perform additional processing when user queries to improve the precision of the final result [6].

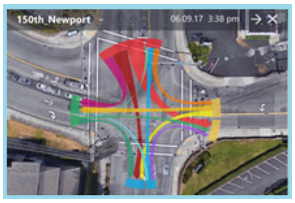
3 DEMONSTRATION SCENARIOS

We will be demonstrating the components of our system Rocket (§2) for different application verticals.

Smart crosswalk: We will setup a camera and analyze its live video to automatically trigger the "walk" signal for a pedestrian in a wheelchair when they come to a crosswalk; Figure 3(a). Further, the walk signal will be extended to provide additional time when it is needed to safely finish crossing, as in this video [2].

Connected Kitchen: Fast food outlets use video analytics for alerts on cars pulling into their driveway and on customers walking in. These alerts enable them to dispatch ushers as well as pre-make certain popular food items so that customer wait-times are reduced. We have an ongoing pilot effort with a major fast-food store, and we will be demonstrating the work.

(a) Smart crosswalk



(b) Traffic dashboard

Figure 3: Applications of video analytics.

Traffic dashboard for Urban Mobility: Partnering with the City of Bellevue, WA, Rocket analyzed their live traffic camera feeds for a 24X7 dashboard of traffic counts (see Figure 3(b)). The dashboard flags abnormal traffic volumes and helps trigger appropriate measures. We will be displaying the dashboard using traffic videos.

Retail Intelligence: Supermarkets and grocery stores use video analytics to understand customer movements for both product placement as well as intervention by their staff. We will demonstrate continuous tracking of people using videos from a retail store.

4 IMPLEMENTATION

We implemented the video analytics system as a .NET Core app and a set of backend cloud services on Azure. The entire system consists of ≈ 33.8k lines of code (loc). We have containerized the local analytics algorithms including the line-based counting and cascaded DNN-based detection and deployed them to various edge devices running either Windows or Linux. Our pipeline can be executed on heterogeneous computing platforms with CPU, GPU or FPGA (e.g., Azure Data Box Edge) [7]. We have also extended Rocket to invoke state-of-the-art machine learning services in the cloud like Azure Machine Learning and Microsoft Cognitive Services.

REFERENCES

[1] 2018. Project Rocket. http://aka.ms/rocket.
[2] 2019. Crosswalk Demo. http://aka.ms/crosswalkdemo.
[3] Chien-Chun Hung, Ganesh Ananthanarayanan, Peter Bodík, Leana Golubchik, Minlan Yu, Victor Bahl, Matthai Philipose. 2018. VideoEdge: Processing Camera Streams using Hierarchical Clusters. In *ACM SEC*.
[4] Haoyu Zhang, Ganesh Ananthanarayanan, Peter Bodík, Matthai Philipose, Victor Bahl, Michael Freedman. 2017. Live Video Analytics at Scale with Approximation and Delay-Tolerance. In *USENIX NSDI*.
[5] Junchen Jiang, Ganesh Ananthanarayanan, Peter Bodík, Siddhartha Sen, Ion Stoica. 2018. Chameleon: Scalable Adaptation of Video Analytics. In *ACM SIGCOMM*.
[6] Kevin Hsieh, Ganesh Ananthanarayanan, Peter Bodík, Shivaram Venkataraman, Victor Bahl, Matthai Philipose, Phillip B. Gibbons , Onur Mutlu. 2018. Focus: Querying Large Video Datasets with Low Latency and Low Cost. In *USENIX OSDI*.
[7] Microsoft. 2018. Azure Brainwave. https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-accelerate-with-fpgas.
[8] Joseph Redmon. 2018. YOLO: Real-Time Object Detection. https://pjreddie.com/darknet/yolo/.
[9] Samvit Jain, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, Joseph E. Gonzalez. 2019. Scaling Video Analytics Systems to Large Camera Deployments. In *ACM HotMobile*.