Distributed Processing for Big Data Video Analytics

David Gibbon and Lee Begeja, AT&T Labs-Research, USA
{dcg@research.att.com, lee@research.att.com}

1. Introduction
The growing number of live video streams available today, combined with advances in system integration capabilities and video processing methods, provides a rich source of quantitative information that can act as a source stream for Big Data ingest to enable a wide range of new applications. While the primary application area for video analytics (VA) has been surveillance [1], some of the methods can be applied to entertainment video sources and consumer generated video as well. This e-letter will provide an overview of VA with a focus on distributed architecture considerations for practical deployment.

While video analytics (VA) typically refers to less constrained computer vision problems than manufacturing inspection [2], the extraction of quantitative information from a manufacturing process via video monitoring followed by analysis of that data clearly uses some of the same building blocks and has similar benefit. VA is used in tasks such as retail inventory control and occupancy monitoring in large retail environments [3]. More recently, consumer applications for VA have opened up due to the widespread availability of requisite infrastructure such as broadband internet access, home WiFi and advances in IP camera technology. While USB cameras connected to PCs and cameras integrated into laptops have been prevalent for many years, it is only recently that affordable standalone IP cameras have become available. The precursors to today’s smart IP cameras were standalone web cams that supported video streaming and, in some cases, pan-tilt-zoom control from web clients via HTTP. More basic implementations allowed only for JPEG frames at low frame rates. With this precedent established, today’s IP cameras are targeted for more private applications such as monitoring a baby’s room, a pet, or for home monitoring. Device capability has improved and costs have dropped so H.264 encoding of 1080p HD video is practical. To be competitive, camera vendors have differentiated on features such as inclusion of microphone/speakers and video quality. To improve quality, vendors have focused on correcting problems such as poor lighting conditions and unstable cameras that plague home monitoring and similar surveillance applications. Wide dynamic range (WDR) or High dynamic range (HDR) algorithms [4], digital image stabilization (DIS) [5] and geometric distortion correction can vastly improve image quality in certain situations.

When used to identify segments of interest from seemingly endless video streams, VA can provide a great benefit to professionals as well as home owners. VA also holds the promise of easing the burden of monitoring an aging parent and raising an alert in the case of a potentially life-threatening fall. Aggregated VA stream storage can provide valuable information on historical patterns and can enable anomaly detection. Systems can incorporate estimates of basic demographic information based on attribute detection (such as age range) in a manner that does not rely on identification of individuals.

2. Architectures
Cost effective implementation of these capabilities requires a high degree of integration and system-on-chip (SoC) architectures [6] are used which typically employ fully customized blocks for tasks such as H.264 encoding along with general purpose computing blocks such as ARM cores. The SoC can implement...
In enterprise applications, several cameras are monitored centrally by a dedicated PC with recording capability. Low-cost IP Network Video Recorders (NVR) for the consumer market are now available. Here, traditional CPU-based VA implementations or those that leverage GPU can play a role, in cases where cameras fall short in processing capabilities. However, in this scenario, these CPU/GPU based implementations will have to handle a number of streams simultaneously. Cloud approaches offer some promise of reduced cost, but considering that numerous streams of HD video would require high upload bandwidth, this may not be a good fit. Further, if this approach were taken, a significant component of the computational load would be fixed and would not take full advantage of the elasticity of the cloud. Camera SoC implementations also have a clear advantage in that they have access to the pixel data prior to compression (see Fig. 1). A hybrid approach where data reduction occurs in several phases seems to be the best choice. The figure suggests the components required for pixel-based processing at different access nodes along the video acquisition to cloud storage path. Algorithms that operate in the compressed domain do not require decoding, and systems that perform feature extraction on the camera advantageously do not require the pixel data (either compressed or decoded) at the later stages.

In order to optimize the processing capability for doing real time VA it is critical that the individual components of the system are designed in an integrated manner. One solution would be for a single entity to design the entire structure from end to end. An alternative is to design the end to end system using standard methods, data structures and APIs through the entire process. The difficulty in that solution is getting a standard established in a timely fashion.

Despite these difficulties, the hybrid approach is the best bet to succeed because it will combine ubiquitous distributed computing using local devices (laptops, tablets, phones) with cloud distributed computing. The need for the cloud is the need to incorporate the benefits of big data into the mass of local computing using ever improving networking.

3. Event Detection

Event detection is the goal of many VA applications, but detailed metadata streams intrinsic to this process can be tapped for additional value. For surveillance tasks, accurately determining when a person enters a zone or leaves an object such as a backpack are typical design goals and have clear utility if they can be achieved with low false alarm rates. However, platforms designed for this can also be used to generate streams of metadata describing the visual content of the scene quantitatively. Analyzed in aggregate, this data can describe trends such as pedestrian or vehicular traffic patterns. Video camera data can be augmented with other sensor data and for some application areas, low-cost depth sensors can provide very robust object segmentation without relying on background removal algorithms.

One branch of current research is focused on reliably detecting events under more challenging conditions such as in public spaces with camera views that encompass a large number of people. The TRECVID Surveillance Event Detection (SED) [9] evaluates algorithms’ performance for a range of events such as a person running, placing an object, or putting a cellphone to their ear. Top performing algorithms extract several types of motion features that are often computationally expensive and apply these in a learning framework for event detection [8]. The primary focus is detection accuracy; further work will be required to improve efficiency so that these novel approaches can be brought to widespread application.

Big data approaches to event detection will not only have to be top performing, they will have to be efficient, and of course respect privacy concerns. The definition of efficiency can be contextual. If processing is distributed locally then there can be various layers of event detection, from local in the camera, to in an aggregating device (like an NVR) and finally to the cloud. Each layer would process and pass up analytics results and meta-data. The higher layer could focus on the most interesting events and use computationally expensive but highly accurate algorithms.
4. Metadata Representation

Open standards for exchange of media and metadata from IP camera systems to Big Data analytics services and storage lakes are critical for the success of this ecosystem. For media, adoption of H.264 is well-established both for camera encoding and mobile device decoding but transcoding may be required to match display device capabilities. Newer cameras support multiple simultaneous encoding streams to address this situation. H.265 HEVC and approaches targeted at surveillance video [10] will offer improved coding efficiencies and suggest that analytics preprocessing (background segmentation) can be performed jointly with coding. At the transport layer, RTSP is used for low latency applications while adaptive HTTP streaming such as HLS is typically used for off-line retrieval of recorded clips. MPEG DASH provides an open standard alternative, but has not yet seen wide adoption in mobile devices.

MPEG-7 provides a flexible framework for content description [11] and can be used to represent extracted metadata. While it can also encode intermediate-level image and video features, true interoperability at the feature level has not been embraced by vendors and systems implementers. Even at the higher level, application-specific profiles or alternative schema may provide developers with a lightweight option to metadata exchange. The industry forum, ONVIF™, has drafted a specification for VA services [12] that includes its own syntax for scene and object description, with spatial and temporal relations. While adoption of open standards in this area is at the early stages and specifications are still under development, the progress to date shows promise toward truly interoperable VA systems.

5. Conclusions

Given the intrinsic bitrates involved with representing high resolution sensor data, video promises to be the biggest of the big data in terms of its computing and networking demands. In order to take advantage of video big data, several elements must come together including advanced IP-connected cameras, distributed video processing architectures, and standards for representation of extracted meta-data of video events. Further work is needed in developing systems and algorithms that can be easily distributed and take advantage of a camera-to-cloud processing path. Combining aggregated video analytics with traditional structured and unstructured big data sources such as location and text based media, facilitated by exchange in agreed-upon metadata representations, with appropriate privacy protections will yield a platform capable of supporting a wide range of novel information services.

References

David Gibbon is Lead Member of Technical Staff in the Video and Multimedia Technologies and Services Research Department at AT&T Labs - Research. His current research focus includes multimedia processing for automated metadata extraction with applications in media and entertainment services including video retrieval and content adaptation. In 2007, David received the AT&T Science and Technology Medal for outstanding technical leadership and innovation in the field of Video and Multimedia Processing and Digital Content Management and in 2001, the AT&T Sparks Award for Video Indexing Technology Commercialization. David has contributed to standards efforts through the Metadata Committee of the ATIS IPTV Interoperability Forum. He serves on the Editorial Board for the Journal of Multimedia Tools and Applications and is a senior member of the ACM and the IEEE. He joined AT&T Bell Labs in 1985 and has and holds over 55 U.S. patents in areas such as multimedia indexing, streaming, and video analysis. He has written a book on video search, several book chapters and encyclopedia articles as well as numerous technical papers. David is an adjunct professor at Columbia University where he teaches a graduate level digital image processing course.

Lee Begeja is Principle Member of Technical Staff in the Video and Multimedia Technologies and Services Research Department at AT&T Labs - Research. His current research focus includes distributed processing for visual analysis with applications in home security and surveillance. His most recent awards are from 2012 when Lee received the AT&T IP Bronze Award for generating shareholder value from the license of intellectual property and 2011 when he received the AT&T Social Innovation Award. Lee holds 69 patents in various areas including network optimization, audio enhancement and video analysis. He has written multiple book chapters and numerous technical papers.