# APPLICATION FOR MACHINE LEARNING TO REDUCE DOWNSTREAM PROCESSING

Malcolm Weir

Kevin Hudson

Tyler Whitlock

Ampex Data Systems Corporation, Hayward, CA, USA

**Abstract**: This paper explores the potential for the use of airborne content analytics in both the Test and Evaluation (T&E) and production/operational environments. Using machine learning and neural networks to examine large datasets for potentially interesting events can dramatically reduce the time taken to identify anomalies, and represents a significant part of the development of the capabilities required by innovations such as "self-driving cars." Development of parallel processing solutions using components originally designed for graphics display ("General Purpose Graphics Processor Units", or GPGPU) has provided processing resources that promise to make machine learning systems a viable component of a space- (and power-)constrained on-board content analytic system. Examples of applications for such systems are discussed, focused on those for which a GPGPU with a few hundred cores might be practical.

**Keywords**: Machine Learning, Content Analytics, GPU accelerated computing

## 1. Introduction

Instrumentation platforms are subject to their own variant of "Moore's law", the observation by Intel's Gordon Moore that transistor density will double every two years. Over the last three decades data storage bandwidth has increased from tens of Megabytes/sec to thousands of MB/sec while storage density has increased from the tens of Gigabytes to tens of Terabytes – the latter drive particularly by the direct application of Moore's Law to flash memory devices. The disparity between the increase in data rates and the increase in storage density effectively means that it can take an order of magnitude longer to process the resultant data sets. The sensors and data systems are continuously evolving to produce and process higher data rates for improved resolution and fidelity. As the data sets get larger, there is an increased need to provide useful context and structure to the data. Several mechanisms, such as multicasting, have been employed to attempt to improve data flow, but most suffer from a lack of "understanding" of the relative importance of any given piece of data.

Frequently cited as a probable cause for human induced error is "too much information." Momentary distractions, longer work hours (and flight durations) and multiple data feeds all serve to reduce each human operator's available attention and ability to identify a pattern in a data stream. Often, the data sets required to detect are either large by nature or potentially so but unknown ahead of time. In a test, instrumentation or collection environment, it is generally impossible to predict when an event of interest will occur, as the value of the data is unknown before it is analysed. However, as the data sets get larger, and as sensors and data acquisition systems get more capable, the time required to move the data and analysed the data increases.

At the same time, the demand for instant access to pertinent information (however defined) is omnipresent. For example, the evolution of flight test data processing at Boeing has been continually toward real-time analysis. The original 1969 747-100 flight test regimen was almost entirely dependent on post-processing; more often than not, that post-processing consisted of engineers – possibly on-board the aircraft – looking at strip charts. A contemporary flight test regimen (as used for the 747-8 Intercontinental, 42 years later) requires flight test information be instantaneously available on the ground while the test article is still in flight

Numerous mechanisms have been implemented over the years to work through these problems. Some solutions target the reduction of the volume of data produced in the acquisition phase (e.g. by processing at point of capture), while others focused instead on providing metadata alongside the acquired data in order to facilitate downstream processing. As no single solution will ever work for all cases, a data acquisition system will naturally integrate multiple techniques to manage the data.

Fortunately, advances in chip development have resulted in extraordinary processing power that can now be ruggedized (reasonably easily) and then integrated into a data acquisition system for harsh environments. The improved processing power can be directed to manage the data in real-time while the data is being acquired.

This data management potentially includes: reducing the size of the acquired data, improving the description of the acquired data (adding metadata), and/or validating the acquired data to produce a summary data set. In addition to traditional rule-based data validators, recent developments in machine learning technology are highly applicable to assisting with these problems. Using processors originally designed to be Graphics Processing Units (GPUs), highly-parallelized accelerated computing is now being utilized to realize practical neural networks capable of performing useful data analysis in a timely fashion.

For example, the Intel® Core™ i7-7567U Processor has, in addition to its two standard "x86" cores, forty-eight integrated GPU cores ("execution units") and dedicated graphics chips, which can contain thousands of GPU cores (the nVidia "Pascal" family has devices with approximately 3,600).

Ideally, in an environment where the value of the acquired data is unknown ahead of time but the cost is fixed, it is desirable to analyze the data at acquisition time to determine the value. Valuable data can generate meta-data tags or system events, anomalous data can generate an error, and other data decimated or discarded. A traditional instrumentation recorder with integrated content analysis capabilities can:

- Reduce the amount of unnecessary/ erroneous data.
- Reduce the post-processing requirements.
- Reduce the time required to access data of interest.
- Improve diagnostic capabilities for fault tolerance.

## 2. Content Analytics

Content Analytics refers to the ability, in real-time, to make decisions based on the content of the data being acquired. One of the differences between a dedicated instrumentation recorder and a generic storage device is that the instrumentation recorder provides the functionality to add metadata to the data streams it receives and stores the combination in such a way as to enhance data reproduction and extraction in real-time. Examples of auto-generated metadata include time stamps and timing information, as well as fault information for out-of-band data. Traditional instrumentation recorders have frequently lacked the bandwidth to store *and* process high-volume, high-speed data streams. The traditional instrumentation recorders typically rely on externally generated and supplied metadata for identifying data of interest, and in some cases requiring operator intervention in the acquisition or post-processing phase; at its simplest, perhaps pushing an "Event Marker" button. The goal of Content Analytics, then, is to provide that meaningful metadata as well as an intelligent control interface together with useful diagnostics while retaining the core functionality, so that overall the system provides enhanced data extraction, reproduction and archiving.

New tools for machine learning are enabling a wide range of capabilities that are enabling significant improvement towards meaningful content analytics on a wide variety of new platforms. As the cost of a useful parallel processor has fallen (driven, as it is, by the demand for ever more realistic graphics which demand ever more powerful engines), there has been a significant acceleration in the development of machine learning technology. The tremendous investment in self-driving cars (and the similar, but less dramatic driver assistance tools) has reaped a number of hardware and software tools for implementing and exploiting neural network architectures. The GPUs provide capabilities to parallelize both large numbers of floating point and matrix operations for which traditional CPUs (Central Processing Units) are not well optimized. Separated from the "video card" role, these devices become sophisticated co-processors, or GPGPUs (General Purpose Graphics Processing Units ). The increased performance provided by advancements in GPU technology makes neural networks practical for a wide range of new applications. In other words, the "compute density" of modern GPU's, functioning as GPGPUs have made long-established machine learning algorithms practical. And using machine learning tools to analyze data in real-time allows for sophisticated data-based decision capabilities.

Neural networks are designed for pattern recognition, which is a technique applicable to a large number of problems, including predicting future events and correlating historical ones. In an environment where the structure and content of acquired data types are well understood, neural networks can be a powerful tool for data analysis. But in an environment where the acquired data structure is not well understood, neural networks can provide useful diagnostic and validation capabilities.

Neural networks are model-based. The operating principle is to compare input data to a hidden model that is created in the training process to see if any previously learned patterns exist in arbitrarily complex data sets. By automating the modeling process, a GPU accelerated data acquisition system can utilize multiple machine learning algorithms.

Unsupervised algorithms (such as Hebbian Learning, Self-Organizing Maps or the Apriori Algorithm) may be appropriate for diagnostics, or control and data validation based on historical system behavior, while supervised machine learning algorithms such as Decision Trees, Random Forests, or other "Classifier learning" technique that may be more appropriate for tagging or otherwise identifying data of interest

## 3. Reduction / Validation

The first step in managing large data sets is managing the physical size of the data set. That is, eliminating invalid, duplicated, counterfeit or otherwise unwanted data. Typical data acquisition systems provide mechanisms for controlling the flow of data as well as mechanisms for validating the data path.

A prerequisite to data validation is "source validation." Data must be received from a valid, known source to have any value. For example, a network-based acquisition system would include network security, e.g. a firewall, to validate the data path as well as the control path. The data itself may then be validated for structure and content. In a GPGPU-accelerated data acquisition system, the validation of both the data path and the control path can be improved – the characteristics used to define the overall data source and data sets can be analyzed in great detail. For example, the traffic patterns can be monitored and anomalies detected, so that data received through an anomalous control route or at an anomalous rate might signal the existence of a significant event or a more serious error.

For data in which the structure and content can be easily modeled, there are additional mechanisms for data validation. For example, consider an IRIG 106 Chapter 11 Analog packet data format, where the range and precision of an A/D converter data source is known at (and before) acquisition time. In this case, additional characteristics such as variance, rate of change, or absolute limits can be modeled and identified and thus used to validate the input data stream.

Another classic example is the surveillance camera application; in this case GPGPU-accelerated data analysis is performed – in real-time – for change detection or object recognition to further validate the video data and so assign a value to the data.

## 4. Prioritization

The primary goal of content analytics is to determine the value of acquisition data in real-time. The value of the data depends on multiple factors, including the (calculated) validity of the data and application specific characteristics. The goal is to provide a description of the data so that a weighted "value" metric can be assigned to the data. That metric can then be used to perform some subsequent action.

For example, the data could be stored or not, pre-processed, forwarded to an external system, or used to trigger an alert to an operator or another part of the data system. More importantly, assigning a weighted value to data at acquisition time will greatly improve the data extraction and down-stream processing capabilities by providing a mechanism for identifying some, if not all, areas of interest in the data set. Autonomous, GPGPU-accelerated data acquisition systems are able to identify areas of interest by comparing incoming data to known data patterns of interest.

The secondary objective is to provide a model for the data set. This model is developed continuously, based initially on historical data and then with feedback from the extraction and analysis process. Maintaining a data model throughout the life cycle of the data set provides a means for continuous improvement of the description of the data set. The weighted value of a data set can be refined based on application specific areas of interest.

## 5. Diagnostics

Diagnostics is a core component of any data acquisition system. As more data is acquired by autonomous systems, diagnostics becomes more critical. The ability to perform diagnostics based on data content can significantly improve the fault detection and fault recovery processes. By enhancing the descriptive characteristics of both erroneous and valid data, common failure modes can be recognized.

For example, a disconnected strain gauge may produce an "out of range" signal, while an incorrectly installed gauge might produce fixed data in the valid range. Or perhaps a video sensor which could recognize a lens cap could generate a "lens cap fault" or even perform an appropriate corrective action. It is not necessary to know the structure and content of acquisition data to perform behavior-based fault analysis.

Consider, for example, the binary sequence 0xFFFF, 0x000, 0x001. If the values are actually 16-bit signed integers (-1, 0, 1), they represent a small variance. But if the sequence was actually unsigned values (65535, 0, 1), then the variance is large. A neural network, however, can be "trained" to recognize that pattern as a low variance signal. GPGPU-accelerated systems can also offer insight into traffic flow patterns that cannot be easily predicted ahead of time, but which can then be used to develop simulated data environments.

More importantly, providing improved diagnostics can significantly improve the resiliency of the overall system, by providing real-time diagnostic information. Content Analytics provides a greater visibility into the overall health of the data acquisition system. The increased visibility provides improved fault detection and diagnostics.

Improved diagnostics can equally be used to automate fault recovery processes. For example, a data acquisition system with redundant inputs can implement an automatic fail-over based on the quality of the data.

Platform-based diagnostic information, as metadata, can improve the downstream processing by providing additional quality of data information. For example, identifying the location and source of anomalous data can differentiate processing errors from acquisition errors, or determine the appropriate data recovery scheme. Improving the visibility of the overall system improves the opportunity to provide meaningful diagnostics which in turn provides the opportunity for improved resiliency of the data management process.

## 6. Implementation

Implementing a system capable of supporting advanced content analytics is a primarily software-centric solution. Software components are added to existing data acquisition systems to take advantage of GPGPU-accelerated processing and the available processing power available in modern computing platforms.

The overall implementation can be separated into three key components:

(1) The data validation and diagnostic component,
(2) The data identification and tagging component,
(3) The data model creation and management component.

The data validation and diagnostic component provides all of the platform specific functionality. This includes the control path and data path validation, signal validation and traffic pattern monitoring. The data model used for this component is based on the data flow through the system; it is responsible for identifying and allocating available system resources, identifying anomalous input or system behavior and providing diagnostic information for problem resolution and recovery. This software is primarily event-based, the component will respond to external triggers, generate fault and data flow events and detect anomalous control or data path behavior.

The data identification and tagging component provides all of the application-specific functionality. That is the functionality required to assign value to validated data. The data model used by this component is based on application requirements for identifying areas of interest. The data identification and tagging component also includes the software for data transformation between known data or encoding formats. The primary function of this component is to provide a rich description of the data set in the form of meta-data to facilitate data management.

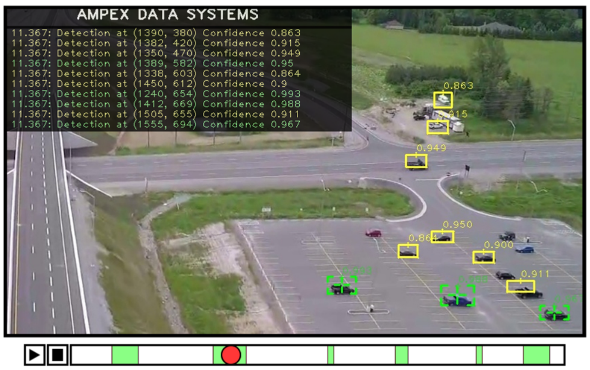


**Figure 1 Ampex SMART Tagging**

For example, Figure 1 shows several “areas of interest” in a video frame, together with the confidence level that the area matches the target description; in this case, green designators are overlaid on targets that have a confidence level in excess of 95%. Simultaneously, the location of this particular frame is indexed, so that an analyst can jump to the sections of the video that have the greatest probability of containing objects of interest.

A secondary function of this component is to improve data extraction and post-processing by pushing some of the functionality for tagging and classifying significant data in the acquisition process. As such, this component includes functionality that can be deferred to the post-processing stage.

The data model creation and management component is the most critical component. This component provides all of the functionality required to generate and maintain the data models required for the first two software components. It includes software integrated across system configuration, data acquisition, and post-processing platforms. This component manages the incorporation of new data into the data models used during the data acquisition and post-process phases. This software includes the software used to provide application specific training for data of interest and it provides the feedback path for human operators.

The primary function of the data model/management component is to maintain data flow and data content models for incorporation into the data management process. The models are typically maintained in a form of distributed database, where new information can be communicated between similar systems when they are in contact – this approach facilitates “swarm” systems, but is also applicable to cases where multiple test articles “share their experiences.” The ultimate goal of this component is to provide continuous improvement of the data models used to validate and classify acquisition data by allowing acquisition and processing systems to learn from each other.

## 7. Conclusion

Recent development in computing performance and machine learning algorithms for autonomous systems are changing the nature of data acquisition – particularly in ruggedized environments. Integrating GPGPU-accelerated computing and the infrastructure to support it provides a practical and affordable mechanism for significantly improving the validity and description of large data sets. In turn, improving the description of the acquired data can significantly reduce the time to process relevant data and thereby reduce post processing over-head.

How much time (and money) gets saved is, of course, application specific. It is axiomatic that the vast majority of test data gets minimal, if any, attention … but how much processing time could be saved if one can first check to see if there is anything of interest to process? Finding a needle in a haystack is hard so anything that shrinks the amount of hay is to be welcomed!

## 8. Acknowledgements

The authors acknowledge with thanks the openness of the extensive and vibrant machine-learning community, and the work performed by researchers everywhere, but particularly in the San Francisco Bay Area, who have developed the toolsets to exploit this emerging technology.