WHITE PAPER

Internet of Things

Making a success of computer vision

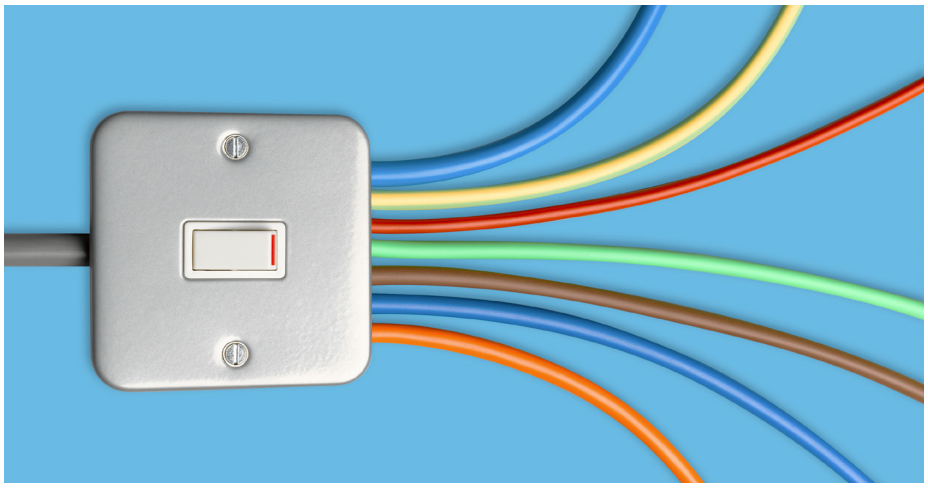
With huge volumes of video being generated, there's more and more demand for computer vision and video analytics solutions at the network edge. Discover the practices that can help system integrators develop them successfully

Executive Summary

In the next few years, there will be a huge growth in the number of systems that use artificial intelligence (AI), computer vision and deep learning to analyze videos and images. The applications are broad: from healthcare to public safety; and from retail to online media.

Enterprises will turn to system integrators (SIs) to implement such solutions. Those SIs will face a number of challenges, including the available compute resources at the edge and the need to rapidly upskill to keep up with demand for video-based applications. SIs can mitigate the risks in video projects by designing with the edge in mind, working with technologies from a trusted solution provider, and running a pilot project at the edge.

Intel supports the ecosystem so that SIs can get optimized solutions that are ready for computer vision for their customers. Intel's technologies include a range of processors, so SIs can balance power consumption and performance to right-size their hardware for the constraints of the edge. There are accelerators, including the purpose-built Intel® Movidius™ Vision Processing Unit (VPU) which was designed specifically to accelerate computer vision and AI inference workloads at the edge. Intel also provides field programmable gate arrays (FPGAs), and a reference design for a vision accelerator based on them.



To develop algorithms, the Intel® Distribution of OpenVINO™ toolkit enables you to more quickly develop solutions that scale across different processors and accelerators, and to take advantage of performance optimizations using them.

As companies such as Outdu, Philips and Advantech have shown, these technologies can form the foundation of powerful computer vision solutions.

Business Challenge

According to Gartner, 65 percent of enterprise-captured videos and images will be analyzed by machines rather than humans by the year 2023[1]. That represents a sharp increase, up from 10 percent in 2018, reflecting the huge growth in AI capabilities and requirements in recent years.

One reason for that success is the growth of deep learning. IHS Markit[2] notes that deep learning algorithms are better than traditional analytics products at simulating human vision, enabling them to better differentiate between objects and behaviors, and increasing accuracy. Deep learning is also enabling greater volumes of video data to be processed more quickly, according to IHS Markit.

There is a wide range of applications for computer vision. Some will see humans supported by computers in existing image processing tasks, such as content moderation, security camera monitoring, and medical diagnoses. Others are entirely new use cases, made possible by computer vision, such as quality control in manufacturing; identity validation; and retail video analytics, where footfall, inventory and queues can be analyzed automatically. New public safety applications will become possible as security and IT technologies come together to enable real-time processing of camera feeds and automatic incident detection.

Whatever the application, the increased use of video and image data creates fresh challenges for those businesses that use computer vision, and their SIs who build the solutions.

While many organizations have become accustomed to more granular and more regular data gathering in recent years, thanks to Internet of Things (IoT) applications, image and video applications represent a significant increase in data volumes. An image at a resolution of 2 megapixels could be as large as about 3 MB in size. When streamed at 30 frames per second (FPS) even in a compressed format such as H.264, that amounts to about 10 Megabits or 1.3 GBytes per second, or 36 Gbits or 4 GBytes per hour.

Data volumes of this magnitude, which may require analysis in real-time, present a number of challenges. Most obviously, all this data must be stored somewhere. In particular, if data must be retained for compliance, record-keeping or evidence reasons, the cumulative data volumes can quickly become large. Bandwidth and network latency may prove to be a bottleneck if visual content has to be sent across a network for analysis in a data center.

Video and real-time image processing are compute-intensive activities, so in some cases the existing compute resources may not be powerful enough. A legacy retail platform used for stock management may struggle to handle the increased processing demand for visual processing. Adding software for visual processing, potentially with accelerators, can also increase the complexity of the solution, complicating operations and support activities.

Security is an important consideration for any computer system today, and capturing image or video content will only increase the need for robust security measures. Sensitive data could be inadvertently recorded, by videoing the location of a person at a particular point in time, for example. Some visual data is inherently sensitive, such as medical imagery, security captures, or facial recognition data.

SIs implementing computer vision systems for their clients need to ensure they have a platform that is capable of storing, processing, communicating, and securing the visual data effectively, so that the solution delivers the business results required.

A Strategy for Implementing Computer Vision

When faced with these challenges in implementing computer vision, there are two facets to the solution: hardware and software. Often, the intelligence for computer vision will be hosted at the edge to avoid the latency, cost and complexity of backhauling video to a data center for processing.

Software can be optimized to extract more performance from the existing hardware. In some cases, it will not be necessary to modify the application code or the AI model, because the underlying frameworks and libraries can be optimized to make better use of processor resources, and any available accelerators. This approach scales well across large organizations, because it does not require any hardware upgrades, but it may be a temporary fix. If the organization wishes to adopt more ambitious AI solutions in the future, the hardware platform might not be powerful enough, even if it can meet today's needs. Where the platform is capable of handling the machine vision workloads, software optimizations enable you to extract optimal performance. Using compute resources effectively helps with overall platform performance if the computer vision software will share resources with other applications.

Upgrading the hardware requires a holistic view to ensure that compute, storage, and networking are all addressed. Tackling just one part of the solution risks creating bottlenecks elsewhere in the system design. Often, it will be possible to use a general-purpose compute platform with appropriate software optimizations, without requiring accelerators to achieve the necessary performance. Using general-purpose processors means that you can share your AI platform with other applications, and it's easier to maintain than having dedicated accelerators in the solution.

For more demanding applications, accelerators can play an important role in enabling real-time image processing.

Properly tiered storage will be part of the solution, enabling current data to be stored in faster storage media, such as Intel® Optane™ DC persistent memory, while archives reside on slower media, such as solid state drives (SSDs). This enables you to strike a balance between cost and performance, and to ensure that live data is available on media that is fast enough to support your application.

In practice, many solutions will require a combination of hardware and software optimizations.

In implementing computer vision solutions for their clients, there are a number of best practices that SIs can follow:

- **Design for the edge:** While some edge locations will be able to support servers, many will have limited space available. Consider the power, thermal dissipation and physical space available for the computer vision equipment. You may need to work with lower-powered edge solutions, and scale your application accordingly. For example, you might choose to analyze frames less frequently in an analytics application, to reduce the load on the processor. You may prefer to use SSDs, even if you don't need their speed, because SSDs offer a far greater storage density than hard drives.
- **Work with a trusted solution provider:** One of the biggest challenges for SIs is integrating video analytics. Although solutions have been around for many years, these have been oversold in the past, with excessive false positives and a high management cost. By working with technologies from a trusted leading IT vendor, such as Intel, you can adopt validated and proven technology, and reduce your risk.
- **Run a pilot project:** It's only when you use a solution that you can truly see what it's capable of. As a first step, you can train and test the AI model using real data, and measure its performance on different architecture options, so you can strike a balance between cost and performance. This stage can be completed in the lab. As far as possible, though, you should also seek to replicate the runtime conditions in a pilot project using your target hardware platform. If you can work at one of the edge locations, within their space, power and thermal constraints, you'll get a fuller picture of any issues arising, and will be able to mitigate them before attempting a roll-out at scale.
- **Use high-quality software:** The effectiveness of your computer vision solution will depend on the software stack you are using. In many cases, you can accelerate your deployment by building a solution on top of proven AI frameworks, such as TensorFlow*. Ensure you're using a software stack that is engineered for enterprise use, and has been proven in edge deployments. Your software needs to have security built in, to ensure visual data is protected.
- **Model the total cost of ownership (TCO):** Cost modelling can be difficult for projects that span a number of edge locations, because there are often externalities that can be hard to incorporate in the cost model, such as the cost of technical support. It may be desirable to use SSDs instead of hard drives, because SSDs are considerably more reliable, even if they cost more up front. Using a more reliable option can cut the number of site visits required, with a dramatic impact on TCO. Where accelerators are required, Intel offers more cost effective options than GPUs, helping you to lower your TCO.
- **Bridge the talent gap:** As video applications take off quickly, SIs may find they need to extend their knowledge and skills rapidly. It may be difficult to hire experienced people, so SIs are investing in training their team to build new visual applications, using training resources provided by Intel, among others.

Solution Overview

Intel provides an end-to-end computer vision portfolio that you can use to build and optimize your solution. The first step is to develop and optimize your software.

You can use the Intel Distribution of OpenVINO toolkit to optimize an existing trained AI model. Over 100 validated models are available from common software frameworks such as TensorFlow*, Caffe* and MXNet*. The OpenVINO toolkit converts the model into an intermediate representation (IR) file, including making optimizations where possible. When the inference engine runs the IR file, it can deploy it using whatever hardware resources are available, so it's easy for SIs to create solutions that run across FPGAs and other accelerators without maintaining separate code bases.

The Intel Distribution of OpenVINO toolkit includes the Intel® Deep Learning Deployment Toolkit (Intel® DLDT) with a model optimizer and inference engine, computer vision functions, and optimized libraries for OpenCV* and OpenVX*, to enhance traditional computer vision applications.

The Intel Distribution of OpenVINO toolkit is free to download. To help train your team and kickstart your projects, Intel provides a number of reference implementations for applications including smart retail analytics, a network video recorder, store aisle monitor, store traffic monitor, safety gear detector, and intruder detector. Code for these implementations is available for download.

Using the OpenVINO toolkit opens up a world of hardware choices, because you can use a single code base and let OpenVINO take care of optimizing your solution for the available execution hardware at runtime, whether CPU, FPGA, Integrated GPU, or video processing unit (VPU).

Looking at CPUs, Intel offers a range of processors so you can balance performance against power consumption, important when hosting video applications at the edge (see Figure 1). You can run the same software at the edge as you do in the data center, and streamline your software development pipeline with consistent tools and languages across all your deployment locations. For low-power applications at the edge, you can use an Intel Atom® or Intel® Core™ processor.

For more demanding applications where power is less of a constraint, choose Intel® Xeon® processors. The Intel® Xeon® Scalable processor includes support for reduced precision (INT8) data, which can help to increase data processing speed, and Intel® Deep Learning Boost (Intel® DL Boost), which adds a new instruction to accelerate deep learning inference. The Intel Xeon Scalable processor includes support for Intel Optane DC persistent memory, which adds an affordable high-performance and high-capacity memory tier in the storage hierarchy.

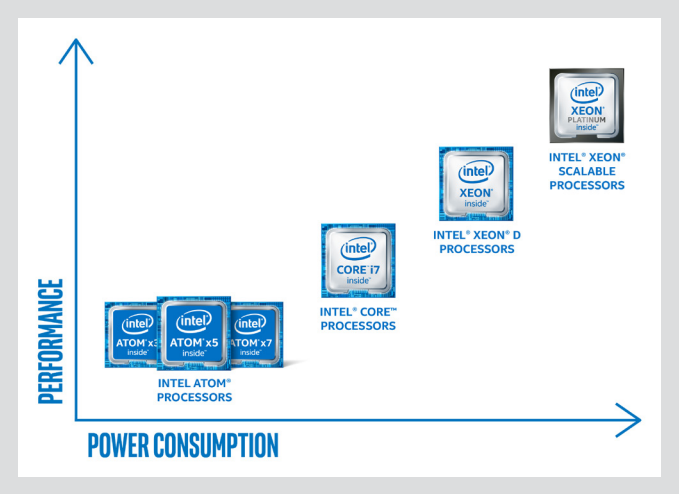


Figure 1. Intel offers a range of processors, so you can choose the right one for the power budget and performance requirements at the edge

Alternatively, FPGAs enable you to reprogram the circuitry of the accelerator to run your application, delivering much faster results than software running on general-purpose hardware. Intel® Programmable Acceleration Cards (Intel® PACs) enable you to connect a pre-validated FPGA board to an Intel Xeon processor-based server, using PCI Express* (PCIe*). Acceleration cards are available based on the Intel® Arria® 10 GX FPGA, or the Intel® Stratix® 10 SX FPGA for more demanding applications. In the past, programming FPGAs required highly specialist skills. Now, Intel has made FPGAs more accessible by creating the Intel® Acceleration Stack for Intel® Xeon® CPU with FPGAs. It provides a standard software stack for FPGAs that enables you to reuse accelerator code, and to use optimized libraries and frameworks to take advantage of the FPGA without any custom programming. One of the strengths of FPGAs is that they can accommodate future innovations in computer vision, because their circuitry can be reprogrammed after shipping.

To enable the vendor ecosystem, Intel has created reference designs for accelerator cards for vision applications. The Intel® Vision Accelerator Designs provide power-efficient deep neural network inference for video analytics and computer vision applications. They can be used in edge servers, network video recorders and edge appliances. The cards are designed to work with the OpenVINO toolkit, so SIs can easily accelerate applications using them.

There are reference designs for two cards:

- The Intel® Vision Accelerator Design with Intel® Movidius™ VPUs incorporates specialized processors designed to deliver high-performance machine vision at ultra-low power. It offers excellent computer vision and deep learning performance per Watt per dollar and helps to drive scalability for well-defined deep learning workloads. It excels in camera and network video recorder applications that have power, size and cost constraints and is good for mainstream deep learning topologies that can be optimized into an ASIC. There is a range of interfaces: PCIe*, mini PCIe, M.2*/Key E*. This reference design typically supports up to 16 video streams per device (depending on frame rate and algorithm complexity), with a batch size of up to 4. Power consumption is typically less than 2W.
- The Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA is an alternative option when you need to run compute-intensive networks such as VGG* or ResNet-101*. It is configured for edge servers, including network video recorders, gateways, and video analytics servers. This reference design offers high performance, low latency, and consistent power consumption (typically around 35W for Intel® Arria® 10 1150 device). Because it is based on an FPGA, the card has dynamic flexibility and future-proofing for custom or new workloads. This accelerator can be connected over PCIe and aggregates between 3 and 32 video streams per device, with batch sizes of up to 144. The reference design reduces the cost and power requirements, compared to deploying most existing FPGA PCIe cards, and has a smaller form factor than a full-size FPGA add-in card. Various precision options (FP16/11/9) are supported to balance speed and accuracy. The card offers especially strong performance in low-batch sizes for mission-critical applications and enables a deep learning network that supports more than 2 million parameters. You can buy a branded version of Intel Vision Accelerator Design with Intel Arria 10 FPGA from IEI Integration Corp.

To develop your software with OpenVINO toolkit, you can use Intel® System Studio. It includes analyzers, compilers and libraries to help you to implement efficient code and improve performance and power efficiency. Using hardware-assisted tracing, it can help you to identify memory leaks and poorly performing code, so you can optimize it. System-wide debuggers and analyzers help you to enhance system stability. Intel System Studio helps you to develop IoT applications faster with access to cloud connectors and over 400 sensors, and easy-to-use code wizards and samples.

Use Cases

Many companies are building computer vision and video analytics solutions based on Intel® technologies. Here are some examples:

- In Outdu's real-time video analytics solution, cameras with built-in AI locally analyze visitors who walk into a site.

The cameras deliver deep insights that lead to greater understanding of visitor behavior. Outdu's video analytics solution incorporates the Intel Distribution of OpenVINO toolkit for optimized speed, performance, and scalability. To enable inference at 60 frames per second (FPS), an Intel® Neural Compute Stick can be added.

- Intel teamed up with Philips to show that servers powered by Intel Xeon Scalable processors could be used to efficiently perform deep learning inference on patients' X-rays and computed tomography (CT) scans, without the need for accelerators. The companies tested two healthcare use cases for deep learning inference models: one on X-rays of bones for bone-age-prediction modeling, and the other on CT scans of lungs for lung segmentation. Using the OpenVINO toolkit and other optimizations, along with efficient multicore processing from Intel Xeon Scalable processors, Philips was able to achieve a speed improvement of 188.1x for the bone-age-prediction model, and a 37.7x speed improvement for the lung-segmentation model over the baseline measurements[3].
- Improving traffic counts in retail environments—ranging from small shops to megastores—is vital in maintaining proper inventory levels and managing staff. With its UShop* EIS AI system, Advantech uses AI video analysis technology, including gateways with Intel Core processors and the new Intel Vision Accelerator Design products, in order to detect human heads and bodies precisely, enabling more precise staff allocation and merchandise adjustment. The Advantech system uses the Intel Vision Accelerator Design with Intel Movidius VPU to generate heatmaps and perform traffic pattern analysis to identify how shoppers walk around the store, information that aids retailers in proper merchandise placement in near-real time. This Intel technology also interprets shopper behavior (such as gestures or reaching for a product), which can be combined with RFID data to identify which items are picked up and purchased versus which items are picked up and put back down, thus gleaning further insight into customers' preferences and behaviors.

Learn more about these use cases and others in the Developer Success Stories library.

Conclusion

As SIs implement vision solutions for their clients, they need a foundation of dependable technologies from a reliable vendor. Intel provides a range of technologies that are optimized for computer vision and deep learning in video environments, and are available from SIs' trusted suppliers. Using these technologies, SIs can work with Intel's broad ecosystem to deliver video-based solutions that are tailored to meet their customers' needs.

Learn More

- Intel® AI: In Production
- Intel Atom® processors
- Intel® Core™ processors
- Intel® Xeon® D processors
- Intel® Xeon® Scalable processors
- Intel® Vision Accelerator Designs
- Intel® Vision Accelerator Design with Intel® Arria® 10 FPGA
- Intel® Distribution of OpenVINO™ toolkit
- Intel® Acceleration Stack for Intel® Xeon® CPU with FPGAs
- Intel® Programmable Acceleration Cards (Intel® PACs)
- Intel® Neural Compute Stick 2 (Intel® NCS2)

Find the solution that is right for your organization. Visit **https://software.intel.com/en-us/ai/ai-in-production.**

Ready to talk?

Solution Provided By:

[1] Market Insight: Tech CEOs Must Leverage Ecosystems and Partnerships for Business Success With AI-Enabled Vision Systems, Gartner, 6 August 2019.

[2] Market Insight: A revolution in video surveillance – Deep learning video analytics, IHS Markit, 24 July 2017

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks

[3] Intel® Xeon® Platinum 8168 processor at 2.70 GHz, Intel® Hyper-Threading Technology (Intel® HT Technology) disabled; BIOS version SE5C6 20.86B.0D.01.0010.072020182008; system memory 192 GB, 2,666 MHz; Intel® Turbo Boost Technology enabled; Intel® SSDSC2CW240A3; Ubuntu* 18.04.1 LTS (GNU/Linux 4.15.0-29-generic x86_64*); Keras* 2.1.1; TensorFlow* 1.2.1; OpenVINO™ Toolkit 2018 R2; Intel Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) v0.14; Bone-Age-Prediction Model 299x299x3 .png images; Lung-Segmentation Model 512x512 .dcm images. The baseline shows zero optimizations. The optimized data used the same system configuration, in conjunction with the noted optimizations. Performance results are based on testing as of August 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Performance results are based on testing as of the date set forth in the Configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

1019/AS/CAT/PDF 341748-001EN