Solution Brief

Public Safety, Security Enforcement, Venue Experience

On-Site Crowd Intelligence Using Vision Transformer AI

An advanced AI solution for public safety and smart venue experience, deployed on Intel® Xeon® 6 Edge Server and Intel® Arc™ Pro GPU

Introduction

With the advancement of deep learning, we witnessed the widespread adoption of Convolutional Neural Network (CNN) models across various edge applications for image classification and object detection. Recently the emergence of Transformer architecture has given rise to a new category of computer vision models known as Vision Transformers (ViT). Despite their larger size and higher computational complexity, these models offer superior capabilities and value-adds for cutting-edge solutions in situation awareness, traffic monitoring, security surveillance, and defect detection.

Modern edge hardware equipped with built-in AI acceleration now readily supports the execution of ViT models locally, eliminating network connectivity and latency constraints. This enables powerful Vision AI solutions that operate efficiently at the edge, delivering real-time intelligence across many vertical sectors like healthcare, manufacturing, energy, retail, transportation, sports and entertainment.

Deloitte Consulting LLP and Intel have collaborated to develop a minimum viable product solution that demonstrates the application of Vision Transformer AI at the edge using readily available off-the-shelf Intel-based edge hardware available from Supermicro. It is an advanced multi-camera video analytics platform that is designed to deliver real-time intelligent crowd monitoring, people safety, and venue experience solution for large-scale events. This paper details the solution architecture and its implementation, which is built on Intel CPU and GPU hardware along with Intel® software components.

Use Case: On-Site Crowd Intelligence

Public safety and security represent fundamental priorities in municipal governance and civic administration. Advanced computer vision technology offers significant potential to enhance crowd safety and experience management across large public venues, including stadiums, arenas, and recreational parks.

As illustrated in Figure 1, the On-Site Crowd Intelligence solution is made up of a high-performance edge server and a network of cameras strategically positioned throughout a venue facility to provide comprehensive surveillance capabilities. The edge server runs a computer vision application which makes use of sophisticated software algorithms to continuously monitor crowd dynamics, analyze pedestrian movement patterns, and detect potential safety incidents in their early stages.

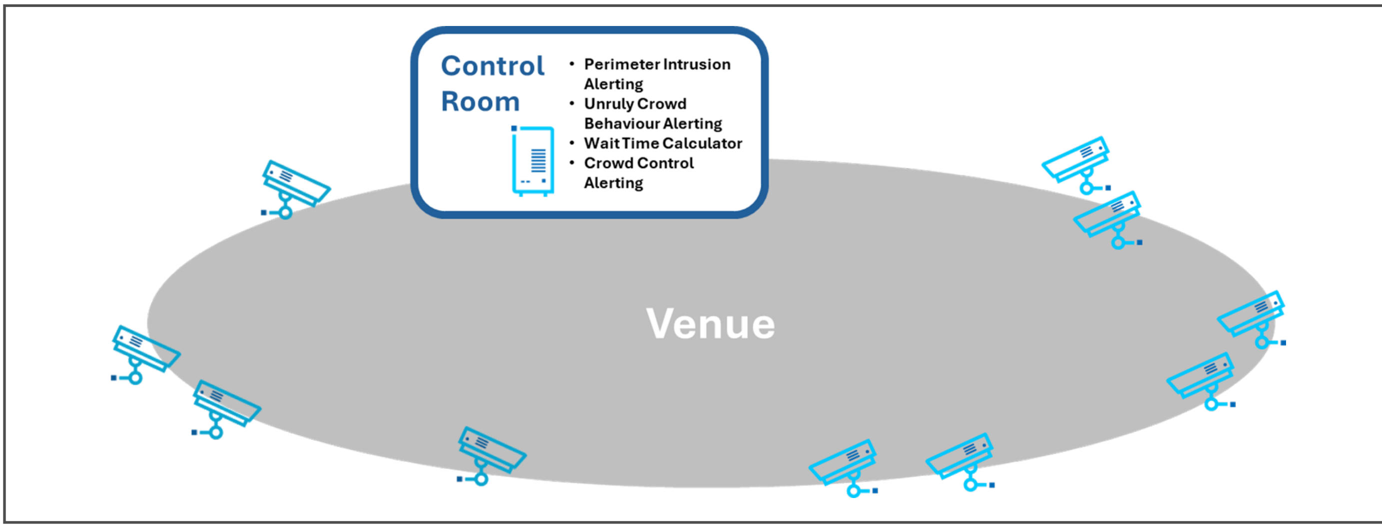

Figure 1. Using AI at edge to enforce security and safety in public venues.

Through automated detection and alert mechanisms, it enables crowd control, proactive intervention, corrective response protocols.

The primary objective of this computer vision solution is to augment the operational capabilities of venue management teams responsible for overseeing public events. By providing real-time situational awareness and predictive analytics, the system enhances decision-making processes for event operations and supports rapid response coordination during critical incidents.

Key operational benefits include improved crowd flow management, early warning systems for potential safety hazards, and enhanced overall security posture for public gatherings. Additionally, the solution can be leveraged to enhance visitor experience in public venues by providing wait time estimations at food and beverage stands, improved wayfinding assistance, and optimized traffic flow management.

This technology-driven approach represents a significant advancement in public safety infrastructure and venue crowd support, enabling more effective protection of citizens and improving visitor experience, while maintaining the operational efficiency required for successful large-scale event management.

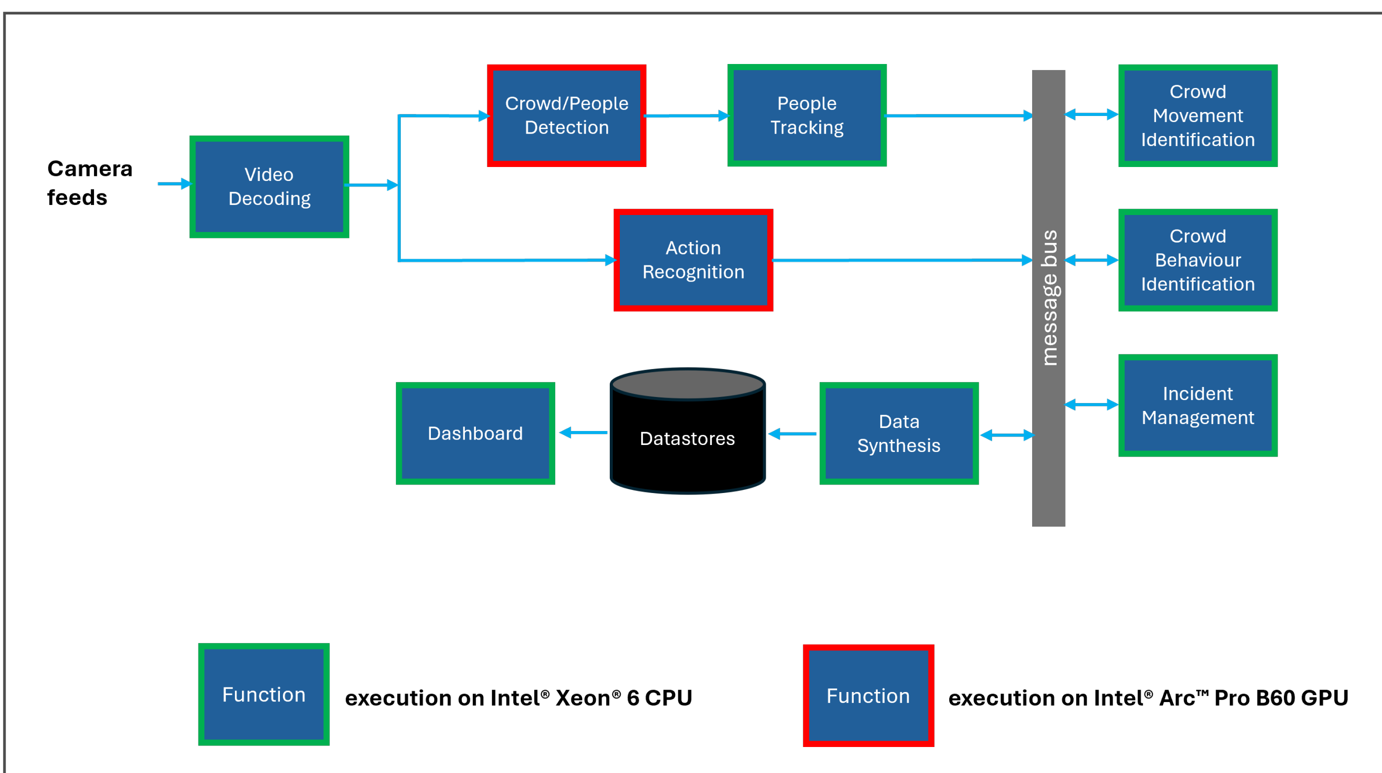

Figure 2. Functional architecture of the Crowd Intelligence application.

Supermicro SYS-212B-N2T Edge Server Configuration	
CPU	Intel® Xeon® 6731P (32 cores)
Memory	512GB DDR5 (Dual Channel)
GPU	2 x Intel® Arc™ Pro B60 Cards (with 24GB GDDR6 memory)
OS	Ubuntu 24.04

Figure 3. Configuration of the edge server for deploying the Crowd Intelligence application.

Architecture

The edge server that hosts the computer vision application is located in the venue's on-site control room, ensuring proximity to management teams and minimizing latency for critical crowd control and safety operations. Equipped with both CPU and GPU components, the edge server receives continuous live video streams from a network of wireless cameras positioned at strategically identified hotspots throughout the venue.

Figure 2 diagrams the core functional components of the computer vision application. Also delineated is the mapping of functions to CPU and GPU resources on the edge server platform.

The application's core functionality centers on a sophisticated multi-model video analytics pipeline designed to process and analyze real-time surveillance feeds. The pipeline enables simultaneous processing of multiple video streams through a series of integrated analytical components.

Each incoming video feed undergoes systematic decoding before being processed through specialized artificial intelligence models. The pipeline incorporates two primary AI components: crowd/people detection using RF-DETR Small model, and action recognition using MViT model. Both models are based on Transformer architecture, representing state-of-the-art approaches in computer vision processing. In the pipeline, the two models work in concert along with people tracking algorithm to provide comprehensive situational awareness across the monitored venue. Inference outputs are relayed to downstream functional modules via a centralized message bus architecture. This design ensures efficient data flow and enables real-time coordination between system components. The key operational functions include crowd movement pattern analysis, behavioral identification algorithms, and automated incident detection and handling protocols. Derived data and results are stored in a database, with a front-end dashboard providing operators with viewing and control functionality.

Implementation

The application runs on a Supermicro's SYS-212B-N2T rackmount server in a compact 2U chassis, with front I/O access and support for full-height full-length PCIe Gen5 slots, well-suited for Edge AI deployment. As shown in Figure 3, the server features a single processor from the Intel® Xeon® 6 processor family and is equipped with two Intel® Arc™ Pro B60 GPU cards. The application first processes incoming video feeds through software decoding on the Intel Xeon 6 processor. AI inference workloads — execution of ViT models — are offloaded to the Intel Arc Pro B60 GPUs for optimal performance. The rest of the application runs on the CPU, including person tracking, result synthesis, operator dashboard, as well as all the functional blocks on the message bus.

As depicted in Figure 4, the ViT models used in the application are optimized and converted to Intermediate Representation (IR) format using the OpenVINO™ Model Converter tool. During operation, the IR formatted models are loaded and executed on the Intel Arc Pro GPU through the OpenVINO runtime environment, ensuring efficient utilization of available hardware acceleration capabilities for real-time video analytics.

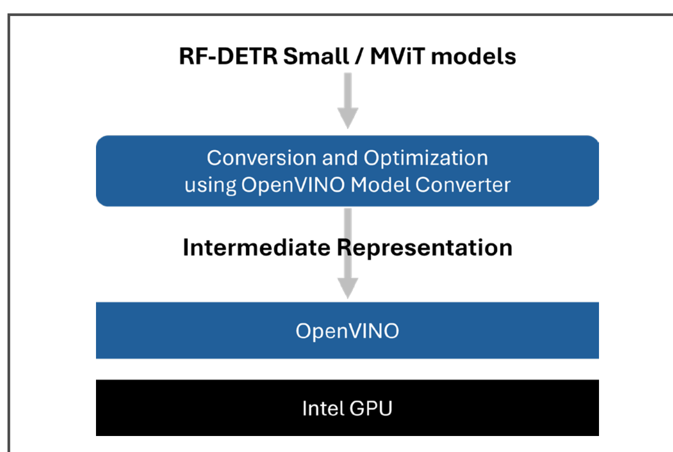

Figure 4. Using OpenVINO™ Toolkit for model conversion and execution.

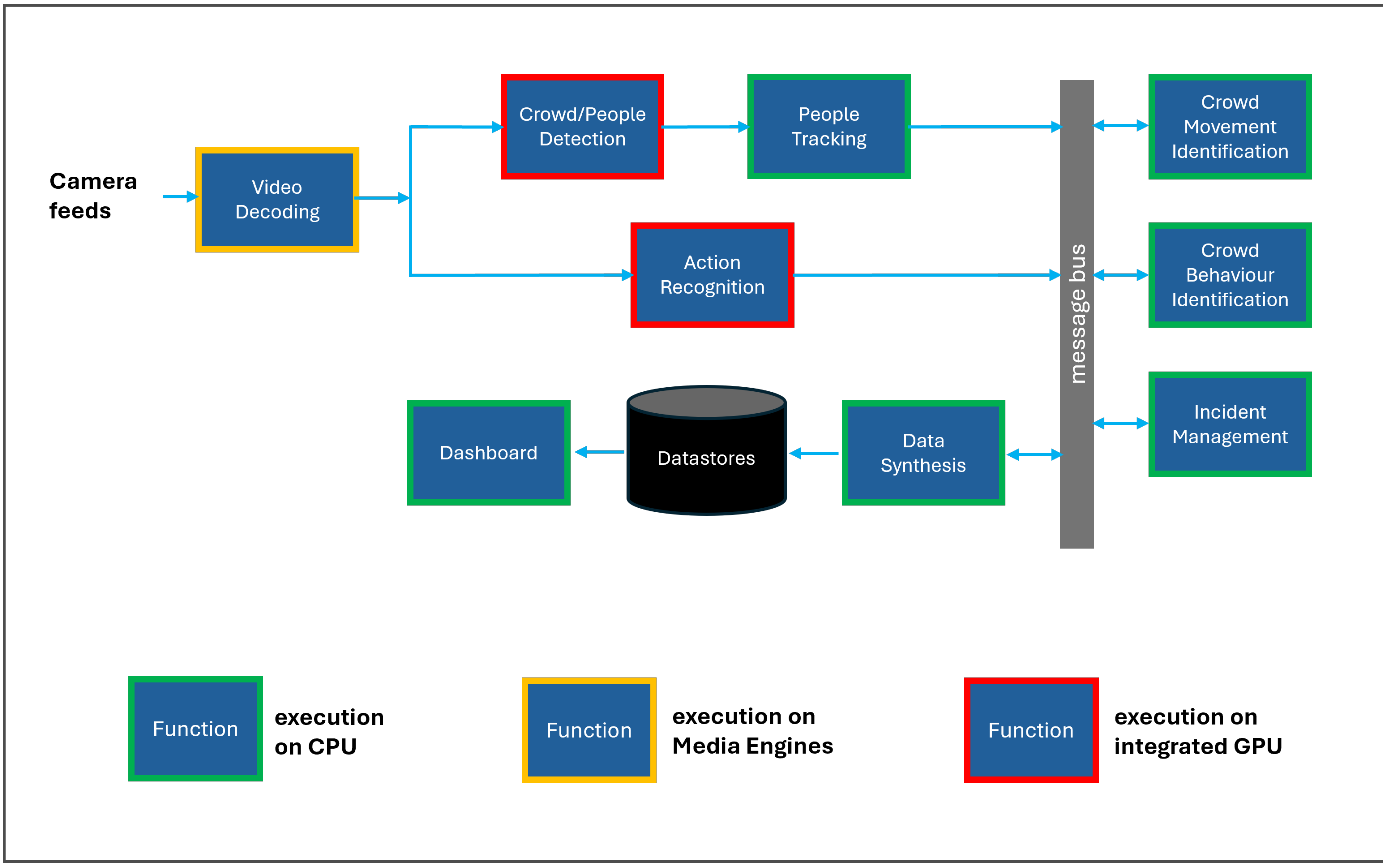

Figure 5. Optimizing workload placement on Intel Core™ Ultra processor.

Performance

Deloitte developed a prototype implementation of the solution and evaluated its performance, with a focus on processing efficiency at both the individual model and integrated system levels.

Their measured data indicated good throughput performance for each ViT model when executed individually on a single Intel Arc Pro B60 GPU. Processing rates of tens of frames per second were achieved for both RF-DETR Small and MViT - each comprising more than ~30 million parameters. It validated the GPU's capability to handle large-scale AI models efficiently.

The system-level evaluation examined processing time across the entire application pipeline as a function of input video stream density. Deloitte's results revealed that when equipped with one Intel Arc Pro B60 GPU, the edge server could process more than 10 concurrent video streams with latency bounded within the use case's tolerance. It confirmed the performance scalability of the Intel Xeon 6 CPU and Intel Arc Pro B60 GPU combination for deploying multi-stream video analytics solutions in edge scenarios where stream density and rapid processing are critical.

Scaling

While the On-Site Crowd Intelligence application was tested and proven on an Intel Xeon 6 processor-based edge server with Intel Arc Pro GPUs, its implementation architecture supports flexible scaling of AI inference workloads across diverse hardware configurations with minimal software modifications. In other words, the application pipeline can be deployed on a wide range of Intel based devices, from small factor embedded computer to high-performance edge server. This is made possible through OpenVINO which provides a unified model execution runtime that operates across Intel client and server processors and accelerators.

On client platforms, OpenVINO can harness multiple compute engines, including CPU, GPU, and NPU. For example, as shown in Figure 5, if the On-Site Crowd Intelligence application is deployed on an Intel® Core™ Ultra platform, video decoding can be handled by the media engines on the integrated GPU, and the AI Inference functions can be offloaded to the integrated GPU, while using CPU resources exclusively for control and data processing.

On Intel Xeon 6 processor-based server platforms without discrete GPUs, OpenVINO can execute AI models using the CPU alone with good performance by leveraging its Advanced Matrix Extensions (AMX) capability.

For performance scale-up, client or server systems can incorporate one or more Intel Arc Pro GPU cards to accelerate AI workloads.

Conclusion

This Deloitte and Intel collaboration demonstrates the practical deployment of Vision Transformer models on edge hardware for real-time crowd safety and venue experience applications. The solution, tested on a new-generation Supermicro edge server, proves that modern Intel based hardware and AI Edge software components can support sophisticated AI workloads on-premises.

Performance results highlight the Intel Arc Pro B60 GPU's performance for AI inference, achieving high throughput in execution of Vision Transformer models and enabling high stream density at the application level. The scalable OpenVINO-enabled software stack provides deployment flexibility across diverse Intel platforms, from power-efficient client systems to high-performance edge servers with discrete GPUs.

This implementation bridges advanced AI capabilities with practical edge deployment, establishing a foundation for broader Vision Transformer adoption in critical safety and crowd control applications, and enabling similar innovations across many sectors such as manufacturing, retail, transportation, sports and entertainment, where real-time edge intelligence delivers transformative value.

Learn More

Intel® Xeon® 6 Processor

Intel® Arc™ Pro B-series GPU

Supermicro Edge AI Servers

[1]As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see http://deloitte.com/us/about for a detailed description of the company's legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

0826/KL/DJA/PDF Please Recycle 369371-001US