

Edge-Optimized AI Surveillance Pipeline with Active Voice Alerting for High-Value Showrooms

Dr. Kishore¹ Vamshi C Krishna², Akash S³, Satwik M⁴, Shrilakshmi C Badiger⁵

¹Dept. of ECE, K S School of Engineering and Management, Bengaluru, India

²Dept. of ECE, K S School of Engineering and Management, Bengaluru, India

³Dept. of ECE, K S School of Engineering and Management, Bengaluru, India

⁴Dept. of ECE, K S School of Engineering and Management, Bengaluru, India

⁵Dept. of ECE, K S School of Engineering and Management, Bengaluru, India

Abstract - Securing commercial environments such as malls and luxury retail showrooms requires highly advanced smart surveillance technologies. Historically, these setups have offloaded video analytics to remote cloud servers, a method fraught with drawbacks including severe latency, high bandwidth consumption, slow response times, and significant privacy vulnerabilities. This research introduces a novel Edge-Optimized AI Surveillance Pipeline equipped with Active Voice Alerting, ensuring that all intelligent monitoring processes are handled locally on edge hardware. By synthesizing lightweight artificial intelligence with real-time Human Activity Recognition (HAR), the system actively detects potential theft and suspicious behavior within high-value spaces. To maximize detection accuracy, we fuse visual data from RGB cameras with kinetic data from IMU sensors. We subject the underlying deep learning model to INT8 quantization, which secures stable inference without overloading the embedded hardware. Moving beyond the passive recording of conventional systems, this pipeline triggers immediate, real-time voice warnings the moment an anomaly is detected. Operating entirely independent of the cloud, this local inference pipeline inherently protects user privacy and reduces external network reliance.

Key Words: Active voice alerting, Edge AI, Human activity recognition, Multimodal sensing.

1. INTRODUCTION

The landscape of security monitoring has been fundamentally transformed by the rapid integration of embedded systems, edge computing, and artificial intelligence (AI). High-value commercial venues demand proactive surveillance that can identify and intercept suspicious actions prior to escalation. Standard closed-circuit television (CCTV) networks primarily function as passive recording devices and heavily depend on cloud-based analytics, introducing unwanted processing latency and a strict reliance on continuous internet connectivity. Furthermore, streaming continuous video data to remote servers exhausts available bandwidth and raises operational costs, while simultaneously exposing potentially sensitive customer footage to external network vulnerabilities. Consequently, decentralized edge-based surveillance has become a highly practical solution for uninterrupted, real-time monitoring. By enabling data processing directly on the local embedded device, edge AI eliminates the need to transmit every frame to the cloud. This localized strategy cuts down

latency, ensures system reliability, and facilitates instantaneous decision-making during critical security events.

Breakthroughs in model compression have empowered these intelligent vision systems to function on hardware constrained by limited computational power. In this project, we introduce a tailored Edge-Optimized AI Surveillance Pipeline featuring Active Voice Alerting, engineered for luxury retail security. By aggregating spatial data from Red-Green-Blue (RGB) video feeds with kinetic data from Inertial Measurement Unit (IMU) sensors, we executed highly accurate multimodal Human Activity Recognition (HAR). To guarantee hardware efficiency and rapid execution, we compressed the core AI model via INT8 quantization and executed it through a custom C++ inference engine. Beyond passive detection, the framework proactively issues automated audio warnings upon detecting anomalies. Ultimately, we deliver a low-latency, highly scalable, and privacy-centric security solution designed for real-world deployment.

2. LITERATURE REVIEW

Previous research explores various facets of surveillance optimization. B. Solmaz et al. focused on utilizing optical flow methods to analyze stability and detect unusual behaviors in densely populated environments, proving that motion pattern observation is highly practical [1]. To address network latency, E. Kim et al. utilized software-defined networking (SDN) and NETCONF protocols to ensure deterministic, low-delay communication [2]. X. Zhang et al. tackled storage constraints by developing a video coding method that separates dynamic foreground objects from static backgrounds via motion estimation [3]. In the realm of the Internet of Things (IoT), F. Esparza-Segundo et al. demonstrated the viability of low-cost surveillance by combining an ESP32-CAM with Telegram to transmit wireless image notifications [4]. M. R. Palav et al. deployed machine learning across IoT frameworks to process smart device data for real-time anomaly detection [5]. S. Kanagamalliga et al. built a low-delay speech synthesis and transcription system using Twilio, illustrating how rapid voice processing bolsters alert mechanisms [6].

Advancements in HAR increasingly rely on multimodal data. C. L. Fabian et al. determined that lightweight algorithms deliver high accuracy suitable for edge environments [7]. G.

A. Martinez-Mascorro et al. successfully deployed 3D Convolutional Neural Networks (3D CNNs) to capture temporal and spatial video features, outperforming traditional analysis methods [8]. Addressing data scarcity, L. Wang et al. engineered a cross-modal few-shot learning system that fuses audio and video data [9], while A. Ghadiya et al. improved anomaly detection using attention mechanisms in a weakly supervised setup [10]. M. Rani et al. utilized a hybrid Vision Transformer and MobileNetV1 architecture to identify activities in low-light edge deployments [11], and J. Chen et al. combined visual and WiFi sensing data through lightweight attention modules [12].

Deploying these robust models on resource-constrained hardware requires stringent optimization. B. Jia et al. optimized resource allocation within fog computing networks to reduce service delays [13]. M. O. Ozcan et al. created a remote debugging framework for troubleshooting applications natively on edge devices [14]. X. Yang et al. drastically reduced computational complexity through INT8 quantization and pruning [15]. D. Chenna et al. performed layer-level quantization analysis for ONNX models [16], while M. Singh et al. adapted quantization dynamically to preserve neural network performance [17]. A. Castagnetti et al. further reduced energy consumption via hardware-aware neural architecture searches [18]. A. S. Affaan Safi et al. achieved accurate detection of loitering using a YOLOv8 and Gated Recurrent Unit (GRU) framework [19], and B. N. Rao et al. engineered a flexible movable camera system using a Raspberry Pi [20].

While this literature highlights massive strides in AI anomaly detection, many systems lack an active deterrence mechanism or lean heavily on cloud infrastructure. Our proposed system resolves this by unifying automated voice alerts, INT8 quantization, and multimodal sensor fusion into a single, localized edge framework.

3. PROPOSED METHODOLOGY

The core methodology revolves around a decentralized, edge-based architecture engineered to identify human anomalies instantaneously. We designed the system to function through four distinct components: the Data Acquisition Module, Edge AI Processing Module, Human Activity Recognition Module, and Active Voice Alerting Module. Initially, the Data Acquisition Module observes the showroom environment by continuously capturing high-definition RGB video streams directly from the surveillance camera. Concurrently, IMU sensors harvest critical kinetic information, measuring parameters like spatial orientation and acceleration. We route this multimodal data to the Edge AI Processing Module. Here, the data undergoes preprocessing and feature extraction before being evaluated by a custom-built C++ inference engine. To accommodate the limited computational capacity of the embedded hardware, we compressed the underlying deep learning model utilizing INT8 precision quantization, which significantly cuts memory consumption without sacrificing analytical accuracy.

4. SYSTEM ARCHITECTURE

Following initial processing, the Human Activity Recognition Module interprets the fused RGB and IMU data to detect targeted behavioral anomalies. The algorithms scan for specific theft-related gestures, unauthorized access, suspicious object concealment, and prolonged loitering. If the system registers an anomaly, it immediately triggers the Active Voice Alerting Module. This module broadcasts an automated audio warning through connected speakers, providing an active deterrence mechanism that directly warns the suspect and simultaneously alerts nearby staff. Because we execute every phase of this pipeline locally on the edge node, the system bypasses cloud servers entirely, ensuring minimal network congestion, protected user privacy, and real-time responsiveness.

5. RESULTS AND DISCUSSION

Testing the finalized surveillance framework on dedicated edge hardware confirmed its computational efficiency and real-time monitoring capabilities. We observed that applying INT8 optimization drastically reduced the necessary processing power and memory footprint when contrasted against standard floating-point models. The fusion of single-input RGB video with IMU sensor data yielded a superior understanding of physical behaviors. This multimodal approach effectively minimized false positives and successfully identified object concealment, loitering, and abnormal hand patterns with high reliability. Upon detecting an anomaly, the voice alert module consistently issued instantaneous audio warnings, significantly elevating the security awareness of the environment. Furthermore, the decentralized nature of the system completely eliminated reliance on cloud networks, resulting in enhanced privacy preservation and ultra-low latency. The entire framework sustained smooth, real-time frame rates while maximizing the efficiency of the underlying edge hardware.

6. CONCLUSIONS

This project underscores the immense practical advantages of combining artificial intelligence with edge computing for secure, low-latency surveillance. The integration of IMU data and RGB streams vastly outperforms traditional CCTV setups in anomaly detection, while processing data at the edge inherently solves the communication delays and privacy vulnerabilities associated with external cloud dependency. Furthermore, the addition of automated voice warnings elevates the system from a passive observer to an active security participant. Future extensions of this work could integrate adaptive learning systems, incorporate larger multimodal datasets, and utilize advanced transformer architectures to deploy this framework across wider, more complex commercial environments beyond retail showrooms.

ACKNOWLEDGEMENT

The authors extend their sincere gratitude to the Department of Electronics and Communication Engineering at K S School

of Engineering and Management for providing the facilities and technical guidance required to complete this research.

REFERENCES

1. B. Solmaz et al., "Identifying behaviors in crowd scenes using stability analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.
2. E. Kim et al., "Ultra-low latency deterministic networking using SDN and NETCONF," *IEEE Commun. Mag.*, 2018.
3. X. Zhang et al., "Efficient surveillance video coding using background reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, 2012.
4. F. Esparza-Segundo et al., "ESP32-CAM based smart surveillance with Telegram integration," 2025.
5. M. R. Palav et al., "AI-IoT framework for real-time customer behavior analysis," 2024.
6. S. Kanagamalliga et al., "Real-time telephonic speech transcription and synthesis system," 2024.
7. C. L. Fabian et al., "Machine learning approaches for multimodal Human Activity Recognition," 2024.
8. G. A. Martinez-Mascorro et al., "Suspicious activity detection using 3D CNNs," 2020.
9. L. Wang et al., "Multimodal few-shot learning for behavioral analysis," 2024.
10. A. Ghadiya et al., "Cross-modal fusion for weakly supervised video anomaly detection," 2024.
11. M. Rani et al., "Hybrid MobileNet and Vision Transformer model for Human Activity Recognition," 2025.
12. J. Chen et al., "WiMix: WiFi and vision-based multimodal Human Activity Recognition," 2023.
13. B. Jia et al., "Resource allocation in fog computing networks," 2018.
14. M. O. Ozcan et al., "Remote debugging framework for edge applications," 2019.
15. X. Yang et al., "INT8 quantization and pruning optimization for lightweight neural networks," 2025.
16. D. Chenna et al., "Quantization Analysis Tool for ONNX edge deployment," 2025.
17. M. Singh et al., "Q-Net Compressor for adaptive quantization," 2024.
18. A. Castagnetti et al., "Hardware-Aware Neural Architecture Search for Edge AI," 2022.
19. A. S. Affaan Safi et al., "YOLOv8 and GRU based smart surveillance framework," 2026.
20. B. N. Rao et al., "Raspberry Pi based movable surveillance camera system," 2020.