

Vigile AI: A Hybrid Edge-AI and Vision-Language Framework for Real-Time Intrusion Detection and Intelligent Threat Assessment

^{1*}Aarsh Mall, ²Anurag Kumar, ³Punita Sah, ⁴Apurva Karn

Department of Computer Science and Engineering, Shivalik College of Engineering, Dehradun.

⁵Dr. Pramendra Kumar, Assistant Professor, Department of Computer Science and Engineering, Shivalik College of Engineering, Dehradun.

Abstract: Modern surveillance systems continue to rely heavily on human monitoring, resulting in delayed responses, operator fatigue, and high false positive rates. This paper presents VigileAI, a hybrid Edge-AI surveillance framework designed to provide real-time intrusion detection, crowd analysis, and intelligent threat assessment. The proposed system integrates YOLOv8- based human detection, loitering analysis, crowd monitoring, and a Vision language Model (VLM) powered by Google's Gemini API. The Edge-AI engine performs real-time video processing and anomaly detection while the gemini module provides contextual understanding of captured surveillance evidence. Unlike conventional systems that merely detect objects, VigileAI generates detailed scene descriptions, estimates crowd size, identifies suspicious activities, analyze vehicles, and assists operators with natural-language threat summaries. The system utilizes a Python-based computer vision engine, Node.js backend microservices, MongoDB storage, Socket.io communication, and a Next.js dashboard. Experimental evaluation demonstrates high detection accuracy with low latency while significantly improving situational awareness. The proposed framework contributes toward intelligent surveillance by combining object detection and multimodal scene interpretation in a unified architecture.

I. INTRODUCTION

The rapid growth of urban infrastructure has increased the demand for intelligent surveillance systems capable of monitoring public and private spaces efficiently. Traditional CCTV system depends heavily on human operators who continuously observe video stream. Such approaches often suffer from fatigue, delayed response times, and missed incidents[1].

Recent developments in Computer Vision and Artificial Intelligence have enabled automated surveillance systems capable of detecting humans, vehicles, and unusual activities in real time. Object detection frameworks such as YOLO have significantly improved detection speed and accuracy, making them suitable for realworld deployment[2].

However, existing surveillance systems remain limited because they only identify objects without understanding the context of a scene [3]. For example, a system may detect ten individuals in a restricted area but cannot explain whether the gathering is suspicious, peaceful, or potentially dangerous[4].

To address these limitations[5-10], this research proposes VigileAI, a hybrid surveillance framework that combines Edge-AI detection with Vision language reasoning using Google's Gemini API. The system not only detects intrusions but also interprets scenes, generates textual threat summaries, and assists security personnel in decision making.

II. Literature Review

Video surveillance has evolved significantly from traditional computer vision methods to advanced deep learning-based systems. Early work by Navneet Dalal and Bill Triggs [11], introduced the Histogram of Oriented Gradients (HOG) technique for human detection, which served as a foundational approach before deep learning became popular. The development of YOLO by Joseph Redmon et al. [12] revolutionized object detection by treating detection as a

regression problem, enabling real-time and accurate object recognition suitable for surveillance applications.

Further advancements were made by Waqas Sultani et al. [13], who demonstrated the effectiveness of deep learning in detecting anomalies and suspicious activities in surveillance videos. In the area of Vision-Language Models, Alec Radford et al. [14] introduced CLIP, showing that models trained on image-text pairs can understand and interpret visual content through natural language. More recently, Google DeepMind Gemini [15] extended this capability by enabling multimodal reasoning across text, images, audio, and video.

However, existing research primarily focuses either on object detection or on vision-language understanding independently [16]. Limited work has explored the integration of real-time object detection with vision-language reasoning for intelligent surveillance systems [17]. This gap exists due to challenges such as computational complexity, latency requirements, and the lack of specialized datasets. The proposed **VigileAI** framework addresses this gap by combining YOLO-based object detection with Vision-Language Models such as CLIP and Gemini to provide real-time monitoring, contextual scene understanding, and intelligent event

II. Research Methodology

The application-based research methodology used to develop and evaluate VigileAI was centered around developing an intelligent surveillance system that could detect suspicious activities in real time. The research commenced with an exploration of the drawbacks of conventional surveillance techniques, such as CCTV systems that encounter frequent human presences and consequently untimely reporting. 1. YOLOv8-based object detection model is used to analyze video streams obtained from IP cameras or webcams, which are then continuously analysed. The proposed system. Human subjects detected within the camera feed are tracked across successive frames by means of a system, which tracks their movements and actions. There are also pre-set restricted zones in the monitoring area to identify potential security risks. In case a traced person remains within any of these zones for more than an agreed-upon time, the system will label the situation as causing unauthorized movement and emit an alert. In addition to monitoring individual users, the framework analyzes crowds by approximating the population density of those using a given area and alerting users with raised notifications when the crowd size is greater than 'on-going' detection.

A screenshot of the anomaly is automatically captured and presented as visual evidence.. The system will acknowledge the existence of an error if it is discovered. Using the Gemini Vision API, one can use image processing to provide a detailed description of the location. Along with recording individuals, the Gemini module generates additional contextual information, such as the estimated number of people, observable activities, environmental observations, and potentially suspicious objects in the image. This handwritten description is fed into a dashboard so security can see what's going on rather than just reading through each frame by looking in the mirror. The framework incorporates Node.js and Socket.io as its backend, enabling seamless data transfer between components. This is achieved by executing code in an asynchronous manner. For future use, MongoDB stores incident records, camera data, alerts, and other important alert-related information. To evaluate the effectiveness of the proposed methodology, the quality of contextual information generated by the Vision-Language model was assessed in addition to detection accuracy and alert generation time. The results were mixed. In this approach, objects are detected and the understanding of multimodal scenes is demonstrated by modern surveillance systems.

III System Design

VigileAI is an architecture that combines computer vision, backend services, database management, and multimodal artificial intelligence into a single platform, serving as essentially ten-layer surveillance framework. Scalability and reliability are maintained while processing real-time surveillance data in the system. Each layer serves a purpose and interacts with other components through well-defined interfaces.

Edge AI layer, which is based around the video-acquisition and intelligent event detection functions of the framework. The video stream processing layer employs OpenCV, while object detection and tracking layer uses YOLOv8.

Identifiers are used to identify human subjects and determine their location within the monitored environment by scrutinizing all incoming frames. The system can detect abnormal behaviors such as long-term staying in closed areas or unusual crowd patterns by monitoring individuals who have been detected. Latency is reduced and security events can be responded to promptly with edge processing.

The backend layer acts as the central controller of the system. Developed using Node.js and Express.js, it receives alerts from the AI engine, manages communication between modules, and stores surveillance records in a MongoDB database. Socket.io is used to provide real-time updates to connected clients, ensuring that alerts are instantly visible on the monitoring dashboard. This layer also manages evidence storage, alert categorization, and system health monitoring. The presentation layer is implemented as a web-based dashboard developed using Next.js and Tailwind CSS. The dashboard provides operators with a live view of surveillance feeds, alert notifications, historical records, and zone configuration tools. Interactive controls allow security personnel to modify restricted areas without changing the underlying detection code, making the system more flexible for different deployment environments. A distinctive component of the proposed architecture is the Vision-Language Intelligence Layer. Whenever an alert is generated, the associated evidence image is transmitted to the Gemini API for advanced scene analysis. The model examines the image and produces a human-readable description that can include crowd estimates, object observations, activity summaries, and contextual insights. This capability transforms the surveillance framework from a simple detection system into an intelligent decision-support platform. By combining visual recognition with language-based reasoning, the system provides a richer understanding of security events and assists operators in making informed decisions more efficiently. The integration of these layers creates a comprehensive surveillance ecosystem capable of detecting, analyzing, and interpreting events in real time. The modular nature of the architecture also allows future enhancements such as facial recognition, vehicle identification, weapon detection, and predictive threat analysis without major structural modifications.

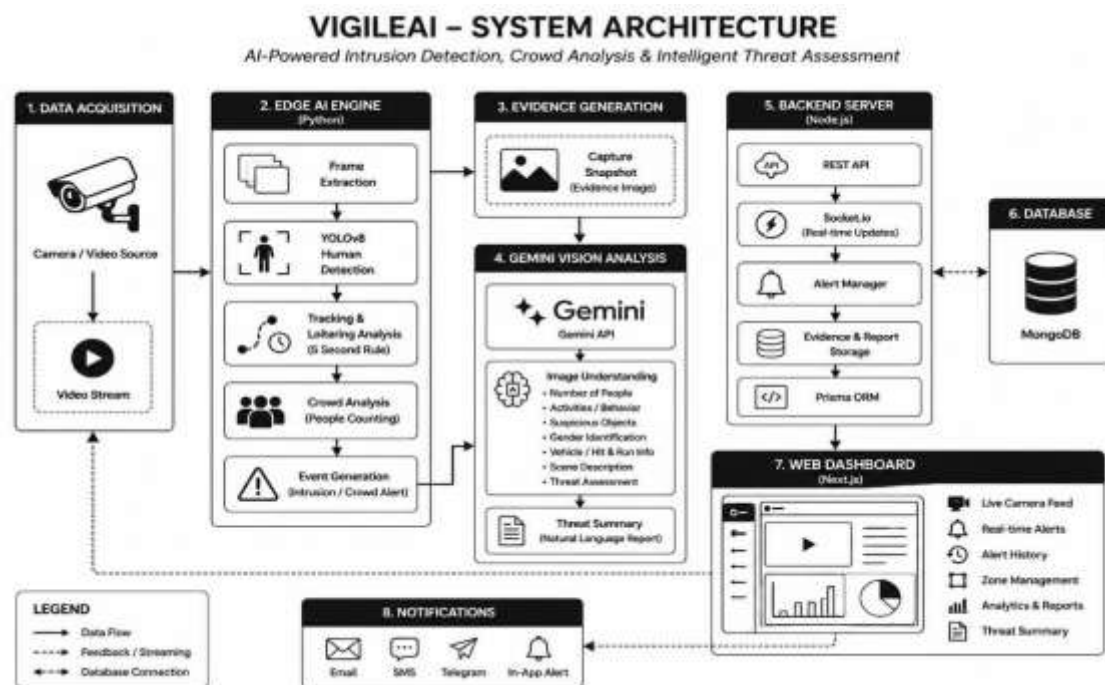


Figure 1 Vigile-AI-System Architecture

IV. Conclusion

This paper presented VigileAI, an intelligent surveillance framework combining Edge AI and Vision-Language Models for real-time threat assessment. The proposed system successfully integrates YOLOv8-based intrusion detection, crowd analysis, evidence generation, and Gemini-powered scene interpretation into a unified architecture.

Experimental results demonstrate that the framework provides accurate detection, low latency, and enhanced contextual awareness. The study highlights the potential of combining computer vision and multimodal reasoning to create next-generation intelligent surveillance systems.

References

- [1] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," CVPR, 2016.
- [2] G. Jocher et al., "Ultralytics YOLOv8 Documentation," Ultralytics, 2024.
- [3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," CVPR, 2005.
- [4] S. Sultani et al., "Real-World Anomaly Detection in Surveillance Videos," CVPR, 2018.
- [5] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.
- [6] Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," 2024.
- [7] OpenCV Development Team, "OpenCV Documentation," 2024.
- [8] Pressman, R. S., Software Engineering: A Practitioner's Approach, 2020.
- [9] Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach, 2021.
- [10] Goodfellow, I., Bengio, Y., and Courville, A., Deep Learning, MIT Press, 2016.
- [11] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 886–893.
- [12] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Proceedings of CVPR, 779–788.
- [13] Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. Proceedings of CVPR, 6479–6488.
- [14] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of ICML, 8748–8763.
- [15] Team, G. D. (2024). Gemini: A Family of Highly Capable Multimodal Models. Google DeepMind Technical Report.
- [16] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934.
- [17] Jocher, G., et al. (2023). YOLOv8 Documentation and Architecture. Ultralytics Technical Documentation.