© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722969

A Lightweight Asynchronous Fusion Framework for Multimodal Surveillance Anomaly Detection

DR. DENSY JOHN VADAKKAN[1], NIKHIL SAIESHWAR[2]
[1] Senior Assistant Professor, School of Science and Computer Studies, CMR University, Bengaluru, India
[2] Master of Science in Information Technology (Data Science), CMR University, Bengaluru, India

Abstract- Multimodal anomaly detection has gained significant attention in modern surveillance systems due to its ability to integrate information from multiple data sources and improve situational awareness. Existing approaches commonly combine modalities such as video, audio, thermal imagery, and sensor data to enhance anomaly detection performance. However, many current fusion techniques rely on strict temporal synchronization and assume that heterogeneous data streams are temporally aligned. In practical surveillance environments, modalities often operate at different sampling frequencies and generate data asynchronously, creating temporal inconsistencies not adequately addressed by many conventional fusion approaches. Furthermore, existing asynchronous learning methods frequently employ computationally intensive architectures that may limit their applicability in resource-constrained surveillance settings. This paper reviews recent developments in surveillance anomaly detection, multimodal fusion, temporal alignment, and asynchronous learning. Based on the limitations identified in existing literature, a lightweight asynchronous fusion framework is proposed for multimodal surveillance anomaly detection. The framework utilizes independent feature extraction, temporal windowing, and asynchronous fusion to facilitate the integration of heterogeneous surveillance data streams without requiring strict temporal synchronization. By emphasizing simplicity, flexibility, and practical deployment considerations, the proposed framework provides a conceptual foundation for future research on lightweight asynchronous multimodal surveillance systems.

Keywords—Multimodal Anomaly Detection, Surveillance Systems, Asynchronous Fusion, Temporal Misalignment, Data Fusion, Deep Learning, Sensor Fusion, Video Analytics

I. INTRODUCTION

The rapid growth of surveillance technologies has led to the widespread deployment of monitoring systems in smart cities, transportation networks, public spaces, industrial facilities, and critical infrastructure. These systems continuously generate large volumes of data from multiple sources, including surveillance cameras, environmental sensors, motion detectors, and other monitoring devices. As the scale and complexity of surveillance environments increase, manual monitoring becomes impractical, creating a growing need for automated anomaly detection systems capable of identifying unusual events and suspicious activities in real time.

Recent advances in deep learning and artificial intelligence have significantly improved anomaly detection capabilities within surveillance applications. In particular, multimodal anomaly detection has attracted considerable attention because it enables the integration of complementary information from multiple data sources. By combining modalities such as video, audio, thermal imagery, and sensor data, multimodal systems can achieve improved contextual awareness and more robust anomaly detection compared to single-modality approaches.

Effectively integrating heterogeneous surveillance modalities is still far from a solved problem, even if there is gain. Most fusion techniques in the literature under study assume that the underlying data streams are temporally synchronized and sampled at comparable rates, but real surveillance implementations hardly cooperate with such assumption. Video is typically captured continuously at high frame rates, while sensor readings usually arrive periodically, or only when a specific event triggers them. The resulting timing mismatches complicate the fusion process and can weaken the reliability of systems built around strict synchronization.

More researchers are studying asynchronous multimodal learning to solve the problem of data misalignment. However, many existing methods use complex and computationally expensive models. They are also mainly designed for autonomous driving, healthcare, or general irregular time-series data, not for surveillance. As a result, there is still a need for simple, lightweight methods that are specifically designed for surveillance systems.

The goal of this paper is to examine how temporal misalignment affects multimodal surveillance anomaly detection, and to review the approaches that have been used to address it. Building on the gaps identified in that review, we propose a lightweight asynchronous fusion framework aimed at integrating heterogeneous surveillance streams more flexibly and practically. It brings together independent feature extraction, temporal windowing, and asynchronous fusion, and is intended for settings where modalities sample at different rates and arrive asynchronously.

II. LITERATURE REVIEW

A. *Surveillance Anomaly Detection*

Anomaly detection has become an essential component of modern surveillance systems due to its ability to automatically identify unusual events and suspicious activities. Traditional surveillance approaches relied heavily on manual monitoring and handcrafted feature extraction techniques, which often struggled to handle complex real-world environments. Recent advances in deep learning have significantly improved anomaly detection performance through the use of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks.

Rezaee et al. [1] presented a comprehensive survey of deep learning-based crowd anomaly detection methods for distributed video surveillance systems, highlighting the effectiveness of deep neural networks in capturing complex spatial and temporal patterns. Similarly, Ramachandra et al. [2] reviewed a wide range of video anomaly detection techniques and demonstrated the growing dominance of deep learning approaches in surveillance applications. Although these studies achieved promising results, they primarily focused on video-only environments and did not consider the challenges associated with heterogeneous multimodal data streams.

B. *Multimodal Fusion Approaches*

To improve contextual understanding and detection accuracy, researchers have increasingly explored multimodal anomaly detection frameworks that combine information from multiple sources such as video, audio, thermal imagery, and sensor data. Jadhav et al. [3] investigated deep learning approaches for multimodal sensor data analysis and demonstrated that combining heterogeneous modalities can improve anomaly detection performance compared to single-modality systems.

Several studies have proposed fusion-based surveillance frameworks. For example, abnormal activity detection using RGB imagery, thermal imagery, and Channel State Information (CSI) [4] demonstrated the benefits of combining complementary data sources. Similarly, recurrent neural network-based approaches utilizing spatio-temporal audio-visual information [5] have shown improved performance in identifying anomalous activities. While these methods successfully exploit multimodal information, many rely on synchronized inputs and assume temporal alignment between modalities.

C. *Temporal Synchronization and Alignment*

Temporal synchronization plays a critical role in multimodal learning systems. Most conventional fusion techniques assume that data streams from different modalities are temporally aligned and sampled at comparable frequencies. However, real-world surveillance environments often contain heterogeneous sensors that generate information at varying rates and intervals.

Research on temporal alignment in multimodal learning [7] has highlighted the difficulties associated with integrating asynchronous data streams. Similarly, recent studies on irregular multivariate time-series anomaly detection, such as SFAFormer [10], have demonstrated that variations in sampling frequency can significantly affect model performance. These findings indicate that temporal inconsistency remains a practical challenge for multimodal surveillance systems, particularly when multiple heterogeneous modalities operate simultaneously.

D. Asynchronous Fusion Methods

To address synchronization challenges, recent research has begun exploring asynchronous multimodal fusion techniques. Approaches such as asynchronous multimodal video sequence fusion [8] and sampling-frequency-aware transformer architectures [10] have shown promising results in handling temporally misaligned data streams. Furthermore, multimodal asynchronous hybrid networks

[9] have demonstrated the potential of asynchronous learning for real-time anomaly detection applications. Despite these advancements, many existing asynchronous fusion approaches rely on computationally intensive architectures, including transformer-based and graph-based models. Additionally, a significant portion of the current research focuses on domains such as autonomous driving, medical imaging, and irregular time-series analysis rather than surveillance systems. More recent multimodal fusion studies have continued this trend: video-audio fusion for complex-environment anomaly detection [13], sensor-level multimodal fusion for anomaly detection in agricultural IoT settings [14], and attention-enhanced multimodal fusion for weakly supervised anomaly recognition [15] all report accuracy gains from combining modalities, yet none explicitly model asynchronous arrival rates between streams. As a result, the development of lightweight and surveillance-oriented asynchronous fusion frameworks remains an important research opportunity.

As shown in Table 1, existing studies either assume synchronized multimodal inputs or employ computationally intensive architectures. Lightweight asynchronous fusion frameworks specifically designed for surveillance applications remain relatively underexplored.

TABLE I. Summary of Existing Approaches

Reference	Focus Area	Key Contribution	Limitation
Rezaee et al. [1]	Video Surveillance	Deep learning anomaly detection survey	Video-only focus
Ramachandra et al. [2]	Video Anomaly Detection	Survey of deep learning approaches	No multimodal integration
Jadhav et al. [3]	Multimodal Sensors	Sensor fusion for anomaly detection	No asynchronous fusion
RGB-Thermal-CSI [4]	Surveillance Fusion	Multimodal activity detection	Requires synchronization
Audio-Visual RNN [5]	Audio-Video Anomaly Detection	Multimodal anomaly detection	Assumes aligned streams
SFAFormer [10]	Irregular Time Series	Frequency-aware anomaly detection	High computational complexity
Async Hybrid Network [9]	Real-Time Async Detection	Handles asynchronous inputs	Focused on autonomous driving

III. RESEARCH GAP AND MOTIVATION

The literature review demonstrates that significant progress has been made in surveillance anomaly detection through the adoption of deep learning and multimodal fusion techniques. By combining information from multiple sources such as video, audio, thermal imagery, and sensor data, existing approaches have improved anomaly detection accuracy and contextual understanding [3][4][5][6]. However, many of these methods rely on synchronized data streams and assume temporal alignment between modalities during the fusion process.

In practical surveillance environments, heterogeneous data sources often operate at different sampling frequencies and generate information asynchronously. For example, surveillance video streams may produce data continuously, while sensor readings are generated periodically or only when specific events occur. Such differences can introduce temporal inconsistencies that complicate multimodal fusion and may reduce the effectiveness of conventional synchronization-based approaches.

Although recent studies have explored asynchronous multimodal learning [8][9][10], many of the proposed solutions employ computationally intensive architectures, including transformer-based and graph-based models. Furthermore, a considerable portion of the existing research focuses on domains such as autonomous driving, healthcare, and irregular time-series analysis rather than surveillance applications. Consequently, the development of lightweight asynchronous fusion approaches specifically designed for multimodal surveillance anomaly detection remains relatively underexplored.

Motivated by these observations, this paper proposes a lightweight asynchronous fusion framework that aims to support the integration of heterogeneous surveillance data streams without requiring strict temporal synchronization. The proposed framework seeks to provide a simple, flexible, and surveillance-oriented approach for addressing temporal misalignment challenges in multimodal anomaly detection systems.

IV. PROPOSED FRAMEWORK

A. Framework Overview

To address the challenges associated with temporally misaligned multimodal surveillance data, this paper proposes a lightweight asynchronous fusion framework. The framework is designed to integrate heterogeneous surveillance modalities without requiring strict temporal synchronization. Unlike conventional multimodal systems that depend on aligned data streams, the proposed approach processes each modality independently and performs fusion at the temporal window level.

The framework consists of five major components: video feature extraction, sensor feature extraction, temporal windowing, asynchronous fusion, and anomaly decision-making. Figure 1 illustrates the overall architecture of the proposed framework.

B. Video Feature Extraction

Video is the richest source of information in a surveillance feed, carrying both spatial detail and temporal context about what is happening in the scene. In this framework, video is processed on its own track, extracting features tied to motion patterns, object interactions, and behavioural cues. These features then serve as the primary visual representation of the scene and are passed forward to the temporal windowing stage.

C. Sensor Feature Extraction

Sensor supplies a complementary layer of framework. Depending on the deployment, this could mean motion sensors, environmental sensors, or other event-triggered devices. The framework extracts features from these streams while keeping their original timing intact, which work with sensors that sample at rates quite different from the video feed.

D. Temporal Windowing

The framework groups incoming video and sensor features into fixed time intervals. Combining observations this way, instead of pairing them one-to-one, gives the framework to absorb differences in sampling rate and to tolerate events that arrive late from a given modality.

E. Asynchronous Fusion

Following temporal windowing, features extracted from different modalities are combined using an asynchronous fusion mechanism. Instead of requiring one-to-one correspondence between observations, the fusion process integrates all relevant information available within a given temporal window. This approach allows the framework to handle temporal inconsistencies while maintaining a lightweight architecture suitable for surveillance applications.

F. Anomaly Decision Module

The fused representation generated by the asynchronous fusion stage is analyzed by an anomaly decision module. Based on the combined information from video and sensor streams, the module determines whether the observed activity corresponds to normal behaviour or a potential anomaly. The output may subsequently be used to generate alerts, support operator decision-making, or trigger further investigation within a surveillance environment.

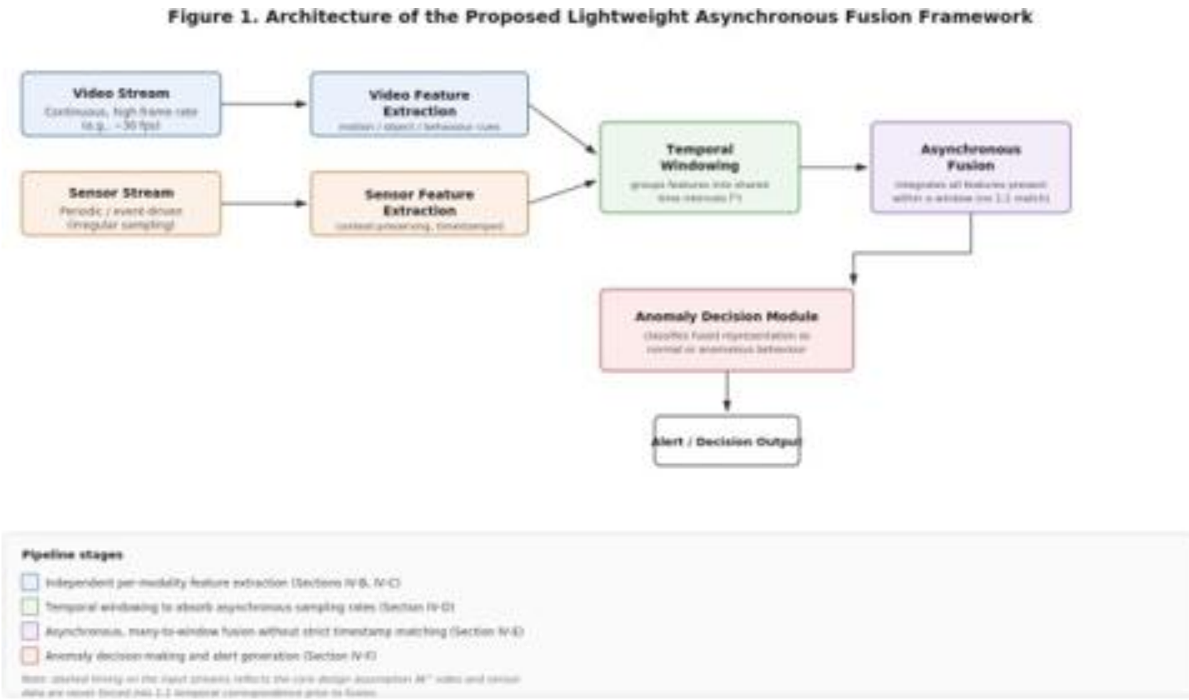

Figure 1. Architecture of the Proposed Lightweight Asynchronous Fusion Framework

G. Mathematical Formalization of the Framework

To complement the architectural description above, this subsection formalizes the framework's core operations. Let xVi denote the i-th raw video observation with timestamp

tVi, and let xS denote the j-th sensor observation with timestamp tS .

$vi = fV(xVi),\ i = 1, ..., n$ (1)

$sj = fS(xS),\ j = 1, ..., m$ (2)

where fV and fS are the independent video and sensor encoders of Sections IV-B and IV-C, producing feature vectors $vi \in \mathbb{R}dV$ and $sj \in \mathbb{R}dS$ respectively. No constraint is placed on n and m being equal, reflecting the differing sampling rates of the two modalities.

$Vk = \{ vi : tVi \in [k\Delta t, (k+1)\Delta t) \},\ Sk = \{ sj : tS \in [k\Delta t, (k+1)\Delta t) \}$ (3)

where Δt is the fixed window width (Section IV-D) and k indexes consecutive, non-overlapping time windows. Unlike synchronization-based methods, Vk and Sk may contain any number of observations, including zero, which allows the framework to tolerate missing or delayed modalities within a window.

$\bar{v}k = (1/|Vk|)\ \Sigma\ vi,\ \bar{s}k = (1/|Sk|)\ \Sigma\ sj$ (4)

$zk = Wf\ [\ \bar{v}k\ ;\ \bar{s}k\ ] + bf$ (5)

$\hat{y}k = \sigma(Wo\ zk + bo)$, anomaly if $\hat{y}k \geq \tau$ (6)

Equation (5) implements the asynchronous fusion step of Section IV-E as a concatenation-and-projection operator, where Wf and bf are learnable parameters. If a window contains no observations for a modality, the corresponding average is replaced by a zero vector, so the operator remains well-defined for empty Vk or Sk, formalizing the tolerance to missing or delayed modalities described qualitatively in Section IV-E. In Equation (6), σ denotes the sigmoid function and $\tau \in (0,1)$ is a decision threshold, so that the anomaly decision module of Section IV-F reduces to a simple thresholder classifier over the fused window representation.

V. DISCUSSION

A. Advantages of the Proposed Framework

The proposed framework offers several advantages for multimodal surveillance anomaly detection. The biggest advantage of the framework is how it handles temporal misalignment: fusing data within fixed windows, instead of demanding exact timestamp synchronization, lets the framework adapt more naturally to real deployments where modalities sample at different rates.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722969

It is also deliberately lightweight. Rather than heavier architectures, the design favors simplicity and flexibility, which matters in surveillance settings where computational resources are often limited.

Adding new modalities, whether audio, thermal sensors, network logs, or other surveillance-related data, shouldn't require reworking the overall architecture.

B. Limitations

Despite its potential advantages, the proposed framework has several limitations. As a conceptual framework, the study does not include experimental validation or performance benchmarking. Consequently, the effectiveness of the proposed asynchronous fusion strategy remains to be evaluated using real-world surveillance datasets.

In addition, the selection of temporal window sizes may influence the quality of information fusion. Windows that are too large may reduce temporal precision, while windows that are too small may not adequately capture relationships between asynchronous events. Further investigation is required to determine optimal windowing strategies for different surveillance scenarios.

C. Practical Applications

The proposed framework may be applicable in a variety of surveillance environments where heterogeneous data streams are generated asynchronously. The framework could be applied wherever surveillance data streams arrive asynchronously; smart city monitoring, public transportation, campus security systems, industrial facility monitoring, and critical infrastructure protection are all reasonable candidates.

Being able to combine video and sensor information without insisting on strict synchronization could make future multimodal surveillance systems more robust and easier to adapt to messy, real-world conditions.

VI. FUTURE RESEARCH DIRECTIONS

What is presented here is a conceptual starting point rather than a finished system, and there is plenty of room to build on it.

The most immediate next step is implementation: testing the framework against real surveillance datasets that actually contain heterogeneous, asynchronous streams, and measuring how the fusion holds up under different degrees of temporal inconsistency and sensor heterogeneity.

The fixed-width windowing used here is also a reasonable starting point, but probably not the best long-term choice. Windows that adjust dynamically based on data characteristics or event frequency could strike a better balance between temporal precision and fusion robustness.

There is also room to push the fusion mechanism itself further. Attention-based or transformer-based components could be layered in without abandoning the framework's lightweight goals, and might help it capture more complex relationships between modalities that arrive out of sync.

Finally, the framework was designed with video and sensor data in mind, but nothing about it rules out other modalities. Audio streams, thermal imagery, and network event logs are natural candidates, and folding them in could sharpen contextual awareness and anomaly detection in more complex environments.

VII. CONCLUSION

Multimodal anomaly detection has emerged as an important research area for modern surveillance systems due to its ability to combine information from multiple data sources and improve situational awareness. However, many existing approaches rely on strict temporal synchronization between modalities, making them less suitable for real-world environments where heterogeneous data streams often operate at different sampling frequencies and exhibit asynchronous behaviour.

This paper reviewed existing research in surveillance anomaly detection, multimodal fusion, temporal alignment, and asynchronous learning. Based on the identified limitations of current approaches, a lightweight asynchronous fusion framework was proposed to address temporal misalignment challenges in multimodal surveillance systems. The framework utilizes independent feature extraction, temporal windowing, and asynchronous fusion to support the integration of heterogeneous surveillance modalities without requiring strict temporal synchronization.

Although the proposed framework is conceptual and has not been experimentally validated, it provides a practical foundation for future research on lightweight asynchronous multimodal surveillance systems. Future studies may focus on implementation, evaluation using real-world datasets, and the incorporation of more advanced fusion techniques to further improve anomaly detection performance in complex surveillance environments.

REFERENCES

[1] K. Rezaee, et al., "A Survey on Deep Learning-Based Real-Time Crowd Anomaly Detection for Secure Distributed Video Surveillance," *Complex & Intelligent Systems*, 2022.

[2] B. Ramachandra, et al., "Video Anomaly Detection—A Survey," *IEEE Access*, 2020.

[3] A. Jadhav, et al., "Deep Learning Approaches for Multi-Modal Sensor Data Analysis and Abnormality Detection," *IEEE Sensors Journal*, 2022.

[4] J. Liu, et al., "Fusing RGB and Thermal Imagery with Channel State Information for Abnormal Activity Detection," *IEEE Transactions on Multimedia*, 2023.

[5] M. Ozkan, et al., "A Recurrent Neural Network for Multimodal Anomaly Detection Using Spatio-Temporal Audio-Visual Data," *Pattern Recognition*, 2023.

[6] S. Chen, et al., "Multimodal Anomaly Detection for Urban Safety: A Real-World Implementation in Large-Scale Surveillance Systems," *IEEE Transactions on ITS*, 2024.

[7] P. Shvetsov, et al., "Temporal Alignment in Multimodal Learning," *arXiv preprint*, 2023.

[8] Y. Wang, et al., "Asynchronous Multimodal Video Sequence Fusion via Learning Modality-Exclusive and -Agnostic Representations," *CVPR*, 2023.

[9] H. Zhang, et al., "When Every Millisecond Counts: Real-Time Anomaly Detection via the Multimodal Asynchronous Hybrid Network," *IEEE Transactions on Neural Networks*, 2024.

[10] L. Li, et al., "SFAFormer: Sampling Frequency-Aware Transformer Specialized for Unsupervised Anomaly Detection in Irregular Multivariate Time Series," *NeurIPS*, 2024.

[11] R. Patel, et al., "Real-Time Video Anomaly Detection: A Lightweight Approach," *IEEE ICIP*, 2023.

[12] T. Nguyen, et al., "Multimodal and Multiscale Feature Fusion for Weakly Supervised Video Anomaly Detection," *ECCV*, 2024.

[13] Y. Wang, Y. Zhao, Y. Huo, and Y. Lu, "Multimodal Anomaly Detection in Complex Environments Using Video and Audio Fusion," *Scientific Reports*, vol. 15, Art. no. 16291, 2025. Crossref

[14] Z. Zhou, "Multimodal Fusion Anomaly Detection Model for Agricultural Wireless Sensors," *Engineering Reports*, vol. 6, e13021, 2024. Crossref

[15] J. Shin, A. S. M. Miah, Y. Kaneko, N. Hassan, H.-S. Lee, and S.-W. Jang, "Multimodal Attention-Enhanced Feature Fusion-Based Weakly Supervised Anomaly Violence Detection," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 129-140, 2025. Crossref