Audio-Visual Anomaly Detection in Long Surveillance Videos Using Context-Aware Temporal Modeling

Logan Hansen
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
lhansen@colostate.edu

Liangying Ding
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
liangying.ding287@uab.edu

Bennett Crawford
Department of Computer Science, University of Houston, Houston, TX, USA.
bennett.crawford802@uh.edu

Tianyi Luo
Department of Computer Science, George Mason University, Fairfax, VA, USA.
tianyi.work@gmu.edu

Abstract

The proliferation of video surveillance systems has created an urgent need for automated anomaly detection methods capable of processing long-duration footage with high accuracy and low latency. While existing approaches have explored either visual or audio modalities independently, the fusion of audio-visual signals remains underexplored, particularly in the context of temporally extended video sequences where contextual dependencies are critical. This paper presents a comprehensive framework for audio-visual anomaly detection in long surveillance videos using context-aware temporal modeling. The proposed system architecture integrates a dual-stream encoder for synchronized audio and visual feature extraction, a hierarchical temporal memory module that captures both short-term and long-range dependencies, and a cross-modal attention mechanism that dynamically weights the contribution of each modality based on scene context. We discuss the structural trade-offs inherent in designing such a system, including the balance between temporal resolution and computational cost, the governance of data privacy during model training, and the infrastructure requirements for real-time deployment in edge and cloud environments. The paper further examines sustainability considerations, such as energy consumption during inference on large-scale camera networks, and fairness implications arising from biased training data distributions across different environmental conditions. Through a detailed analysis of deployment scenarios in smart city, industrial, and public transit contexts, we illustrate how context-aware temporal modeling can improve detection robustness while reducing false alarm rates. Finally, we outline policy recommendations for responsible deployment, including transparency in model decision-making and equitable performance across diverse demographic and geographic settings. This work contributes a systems-level

perspective that bridges algorithmic innovation with socio-technical governance, offering a roadmap for future research in multimodal surveillance analytics.

Keywords

audio-visual anomaly detection, surveillance video, temporal modeling, context awareness, multimodal fusion, large-scale systems, socio-technical infrastructure.

1. Introduction

The modern urban environment is increasingly instrumented with dense networks of surveillance cameras and microphones, generating vast quantities of continuous audiovisual data that far exceed human capacity for real-time monitoring. Anomaly detection in this context refers to the automatic identification of events or behaviors that deviate from expected patterns, including accidents, criminal activity, infrastructure failures, and public safety incidents. Traditional approaches have relied predominantly on visual data, employing convolutional neural networks and recurrent architectures to detect spatial and temporal irregularities [1, 2]. However, visual-only systems suffer from fundamental limitations, including occlusion, poor lighting, and narrow fields of view, which can be partially mitigated by incorporating audio streams that capture acoustic signatures of anomalous events such as screams, glass breaking, or engine failures [3, 4]. The fusion of these two modalities offers a more robust representation of the scene, but introduces significant challenges related to temporal alignment, modality imbalance, and computational overhead.

Long surveillance videos, defined here as recordings spanning hours or days, present additional difficulties because anomalies are rare events embedded within long periods of normal activity. Detecting such sparse events requires models that can maintain contextual awareness over extended time horizons, distinguishing between transient noise and genuine anomalies while avoiding catastrophic forgetting of earlier context [5, 6]. Context-aware temporal modeling addresses this need by incorporating mechanisms that selectively attend to relevant historical information, adaptively weighting the influence of past observations based on their relevance to the current frame. This paper argues that the integration of audio-visual fusion with context-aware temporal modeling represents a necessary evolution in surveillance analytics, moving beyond frame-level classification toward a more holistic understanding of scene dynamics.

The contributions of this work are threefold. First, we propose a system architecture that combines dual-stream encoding, hierarchical temporal memory, and cross-modal attention in a manner that is both computationally tractable and scalable to large camera networks. Second, we provide a thorough analysis of the structural trade-offs and governance challenges inherent in deploying such systems, including data privacy, algorithmic bias, and energy sustainability. Third, we offer forward-looking policy recommendations that balance the societal benefits of improved anomaly detection with the need for responsible oversight. By situating our technical contributions within a broader socio-technical framework, we aim to inform both researchers and practitioners who must navigate the complex landscape of large-scale surveillance infrastructure.

2. Related Work

Research in video anomaly detection has evolved significantly over the past decade, moving from handcrafted features to deep learning-based approaches. Early methods relied on optical flow and trajectory analysis to identify deviations from learned motion patterns [1]. The

advent of autoencoders and generative adversarial networks enabled reconstruction-based anomaly detection, where abnormal events are identified by high reconstruction error [2, 7]. These methods, while effective in constrained settings, struggle with the diversity of real-world scenes and the scarcity of labeled anomaly data. More recently, self-supervised learning paradigms have emerged, leveraging pretext tasks such as frame prediction, temporal order verification, and contrastive learning to learn representations without explicit anomaly labels [8, 9]. Despite these advances, most visual-only methods remain sensitive to environmental variations such as lighting changes, camera motion, and occlusions.

Audio-based anomaly detection has developed in parallel, with applications ranging from industrial machine monitoring to public safety. Convolutional and recurrent neural networks applied to spectrograms have shown strong performance in detecting acoustic events such as gunshots, breaking glass, and human distress calls [3, 10]. However, audio alone is often ambiguous, as similar acoustic signatures can arise from benign sources, leading to high false positive rates. The fusion of audio and visual modalities offers a path toward disambiguation, as the two streams provide complementary information that can be jointly reasoned over [4, 11]. Early fusion approaches concatenate features from both modalities at the input level, while late fusion combines decisions from separate classifiers. Intermediate fusion, where cross-modal interactions occur at multiple network layers, has demonstrated superior performance by enabling the model to learn shared representations [12].

Temporal modeling for long videos has been addressed through recurrent architectures such as long short-term memory networks and gated recurrent units, as well as through transformer-based models that employ self-attention over time [5, 6]. These methods face a fundamental trade-off between capturing long-range dependencies and maintaining computational feasibility. Hierarchical approaches, which process video at multiple temporal scales, have been proposed to mitigate this issue by first detecting candidate segments at coarse resolution and then refining analysis at finer granularity [13]. The challenge of maintaining context over extended durations is particularly acute in surveillance applications, where normal behavior patterns may shift slowly over time due to changes in lighting, crowd density, or human activity cycles. Context-aware mechanisms that adaptively modulate the temporal receptive field based on scene dynamics offer a promising direction for addressing this challenge.

3. System Architecture for Context-Aware Audio-Visual Anomaly Detection

The proposed system architecture is organized into three principal components: a dual-stream feature extraction module, a hierarchical temporal memory module, and a cross-modal attention fusion mechanism. Each component is designed to address specific challenges inherent in long surveillance video analysis while remaining amenable to distributed deployment across edge and cloud infrastructure.

The dual-stream feature extraction module processes audio and visual data through separate encoder networks that are pretrained on large-scale datasets and fine-tuned on domain-specific surveillance footage. The visual stream employs a 3D convolutional backbone that captures spatiotemporal features from short video clips, producing a sequence of feature vectors that encode both appearance and motion information [2, 8]. The audio stream operates on log-mel spectrograms derived from the synchronized audio track, using a residual network architecture that has been shown to be effective for acoustic event classification [3, 10]. A critical design consideration is the temporal alignment between the two streams, as audio and visual events may have different onset latencies and durations. The system incorporates a

learned alignment layer that estimates temporal offsets based on cross-correlation between the feature sequences, allowing the model to handle asynchronous events such as a visual impact followed by a delayed sound [4, 11].

The hierarchical temporal memory module is responsible for maintaining contextual information over long video durations. Rather than processing the entire video as a single sequence, which would be computationally prohibitive, the module operates at multiple temporal scales. At the finest scale, short-term memory buffers capture local patterns over windows of a few seconds, enabling the detection of transient anomalies such as a person falling or a vehicle suddenly braking [5]. At an intermediate scale, medium-term memory aggregates information over minutes, capturing patterns such as crowd flow dynamics or periodic machinery operations. At the coarsest scale, long-term memory stores compressed representations of hours-long activity patterns, enabling the model to recognize anomalies that develop gradually, such as a person loitering or a slow leak in a pipeline [6, 13]. The memory module employs a gating mechanism that determines which information to retain, update, or discard at each scale, preventing the accumulation of irrelevant context that could degrade detection performance. This hierarchical design balances the need for long-range context with the practical constraints of memory and computation, as the coarser scales operate at lower temporal resolution and thus require fewer resources.

The cross-modal attention fusion mechanism dynamically weights the contribution of audio and visual features based on the current scene context. In many surveillance scenarios, the informativeness of each modality varies significantly over time. For example, in a low-light environment, visual features may be unreliable while audio features remain salient, whereas in a noisy industrial setting, the opposite may be true. The attention mechanism computes a set of modality-specific relevance scores for each temporal segment, using a learned function that takes into account both the current feature vectors and the contextual information stored in the hierarchical memory [12, 14]. These scores are used to modulate the fusion of the two streams, producing a combined representation that emphasizes the more informative modality for each time step. This approach avoids the limitations of fixed fusion strategies, which either treat both modalities equally or rely on static weights that cannot adapt to changing conditions. The output of the fusion module is fed into a classification head that produces an anomaly score for each temporal segment, along with a confidence estimate that can be used for thresholding and decision-making.

4. Structural Trade-Offs and System-Level Considerations

The design of any large-scale surveillance analytics system involves a series of structural trade-offs that must be carefully balanced to achieve acceptable performance across diverse operating conditions. One of the most fundamental trade-offs is between temporal resolution and computational cost. High temporal resolution, achieved by processing video at the native frame rate and audio at high sampling frequencies, allows the detection of brief, transient anomalies but requires substantial compute resources for encoding and memory management [1, 5]. Conversely, downsampling the temporal input reduces computational load but risks missing short-duration events that may be critical for safety applications. The hierarchical temporal memory module mitigates this trade-off by allocating high resolution only to the finest temporal scale, while coarser scales operate at reduced resolution. However, this introduces a design parameter: the number of scales and the resolution at each scale must be chosen based on the expected duration of anomalies in the target deployment environment. For a smart city application where anomalies may range from a few seconds (e.g., a car crash)

to several minutes (e.g., a fire), a three-scale hierarchy with resolutions of one second, ten seconds, and one minute may be appropriate. For a manufacturing plant where equipment failures develop over hours, a different configuration would be needed.

Another critical trade-off involves the balance between model complexity and inference latency in real-time deployment. The dual-stream encoder with cross-modal attention introduces significant computational overhead compared to single-modality systems, particularly when deployed across hundreds or thousands of cameras in a city-wide network [15]. To address this, the architecture supports a tiered deployment strategy where edge devices perform lightweight preprocessing and initial anomaly screening, while more computationally intensive analysis is offloaded to cloud servers. Edge devices can execute the dual-stream encoder at reduced resolution and pass only segments with elevated anomaly scores to the cloud for full hierarchical analysis [16]. This approach reduces bandwidth requirements and central processing load, but introduces latency in the cloud path that may be unacceptable for time-critical events. The governance of this trade-off requires careful specification of service-level agreements that define maximum acceptable detection latency for different anomaly categories.

Sustainability is an increasingly important consideration for large-scale surveillance infrastructure, as the energy consumption of continuous video processing can be substantial. A single camera with on-device processing may consume tens of watts, and a city-wide network of thousands of cameras can represent a significant electrical load [17]. The proposed architecture contributes to energy efficiency through its hierarchical design, which allows the majority of camera feeds to be processed at coarse temporal resolution during periods of normal activity, with finer resolution activated only when anomalies are suspected. Additionally, the cross-modal attention mechanism can be configured to disable one modality when it is consistently uninformative, such as turning off audio processing in consistently quiet environments. These adaptive strategies reduce average power consumption without sacrificing detection performance during critical events. However, they require careful calibration to avoid missing anomalies that occur during low-resolution periods, and the energy savings must be weighed against the cost of more complex control logic.

5. Deployment Infrastructure and Governance

Deploying an audio-visual anomaly detection system at scale requires a robust infrastructure that spans edge devices, network connectivity, and centralized processing resources. Edge devices, typically camera units with embedded processors, must support real-time audio capture, visual encoding, and initial feature extraction. The computational capacity of these devices is limited, necessitating the use of lightweight neural network architectures that have been optimized through techniques such as quantization, pruning, and knowledge distillation [16, 18]. The selection of edge hardware involves a trade-off between processing power and cost, as more capable processors increase per-unit expense but reduce the need for cloud offloading. For large-scale deployments, a heterogeneous mix of edge devices may be optimal, with higher-capability units deployed at critical locations such as transit hubs and lower-capability units used in less sensitive areas.

Network infrastructure must support the transmission of feature vectors and anomaly scores from edge devices to central servers, while also accommodating control signals that adjust processing parameters based on changing conditions. Latency and bandwidth requirements vary depending on the deployment tier; edge-processed anomaly scores require minimal bandwidth but must be transmitted with low latency for real-time alerting, while cloud-

offloaded segments require higher bandwidth but can tolerate moderate delays [15, 19]. The design of the network topology must account for redundancy and failover to ensure continued operation during outages, as surveillance systems are often relied upon for critical safety functions. Governance of this infrastructure involves establishing protocols for data retention, access control, and audit logging, which must comply with local regulations regarding surveillance and privacy.

Data privacy is a paramount concern in any surveillance system, particularly when audio recordings are involved, as they may capture private conversations or sensitive personal information. The proposed architecture incorporates privacy-preserving techniques at multiple levels. At the edge, audio features can be extracted and transmitted without retaining raw audio waveforms, reducing the risk of privacy breaches [20]. The hierarchical memory module operates on feature representations rather than raw data, further limiting exposure. However, these measures are not foolproof, as feature vectors can potentially be inverted to reconstruct approximate versions of the original signals. Differential privacy mechanisms can be applied during training to limit the information that the model can memorize about individual data points, but this comes at the cost of reduced detection accuracy [21]. Governance frameworks must therefore define acceptable trade-offs between privacy protection and system performance, with input from legal experts, community stakeholders, and civil liberties organizations.

6. Fairness, Bias, and Robustness

The performance of audio-visual anomaly detection systems can vary significantly across different demographic groups and environmental conditions, raising concerns about fairness and equity. Training data for surveillance models is often collected from a limited set of locations and time periods, leading to underrepresentation of certain populations, lighting conditions, acoustic environments, and cultural behaviors [22]. For example, a model trained primarily on daytime footage from urban centers may perform poorly in rural areas, at night, or during cultural events where normal behaviors differ from the training distribution. The cross-modal attention mechanism can partially compensate for such biases by relying more heavily on the modality that is less affected by environmental variation, but this assumes that the model has learned appropriate modality weighting strategies during training. If the training data itself is biased, the model may learn to ignore audio in environments where it is actually informative, or vice versa.

Robustness to distribution shift is another critical concern, as surveillance systems are deployed in dynamic environments where conditions change over time. Seasonal variations, infrastructure modifications, and changes in human activity patterns can all cause the input distribution to drift away from the training distribution, leading to degradation in detection performance [23]. The hierarchical temporal memory module provides a degree of robustness by continuously updating its contextual representations based on recent observations, allowing the model to adapt to gradual changes. However, rapid shifts, such as those caused by a sudden power outage or a large public event, may overwhelm the adaptation mechanism. Continuous monitoring of model performance metrics, such as false positive and false negative rates, is necessary to detect distribution shift and trigger retraining or recalibration. Governance processes should establish thresholds for acceptable performance degradation and define procedures for model updates, including the collection of new training data from the deployment environment.

Addressing fairness requires proactive measures during dataset collection, model development, and deployment. Dataset curation should strive for demographic and environmental diversity, with explicit sampling strategies to ensure representation of underrepresented groups [22]. During model training, fairness constraints can be incorporated into the loss function to penalize performance disparities across predefined groups, though defining these groups in a surveillance context raises its own ethical challenges. Post-deployment, ongoing auditing of system performance across different locations, times, and population segments is essential to identify and correct disparities. These audits should be conducted by independent third parties with access to raw performance data, and the results should be made publicly available to build trust with affected communities [24]. The cost of such auditing must be factored into the overall system budget, as it represents a recurring operational expense.

7. Policy Implications and Forward-Looking Perspectives

The deployment of audio-visual anomaly detection systems at scale carries profound policy implications that extend beyond technical performance. One of the central tensions is between the societal benefits of improved public safety and the potential for surveillance to infringe on civil liberties. Automated systems can process vast amounts of data without fatigue, potentially enabling earlier detection of crimes and accidents, but they also create the possibility of mass surveillance that chills legitimate expressive activity and disproportionately impacts marginalized communities [25]. Policymakers must therefore establish clear guidelines for the permissible uses of such systems, including restrictions on data retention, requirements for human oversight of automated decisions, and mechanisms for appeal when individuals are falsely flagged as anomalous.

Transparency is a key principle for responsible deployment. Citizens should be informed when they are entering areas under automated surveillance, and they should have access to information about the capabilities and limitations of the systems being used. Model explainability techniques can help operators understand why a particular event was flagged as anomalous, enabling them to verify the system's reasoning and override incorrect decisions [14]. However, explainability is an active area of research, and current methods often produce explanations that are difficult for non-experts to interpret. Policy should therefore mandate a minimum standard of explainability for systems used in high-stakes contexts, while also investing in research to improve the interpretability of deep learning models.

Looking forward, the evolution of audio-visual anomaly detection will be shaped by advances in foundation models, self-supervised learning, and neuromorphic computing. Foundation models pretrained on massive multimodal datasets have the potential to reduce the need for task-specific labeled data, enabling faster adaptation to new deployment environments [26]. Self-supervised learning techniques that do not require anomaly labels during training can further lower the barrier to adoption, though they introduce new challenges in evaluating model performance [9]. Neuromorphic hardware, which mimics the event-driven processing of biological neural systems, offers the promise of ultra-low-power inference that could enable continuous processing on battery-powered edge devices [27]. These technological developments must be accompanied by parallel advances in governance frameworks that ensure these powerful tools are deployed in ways that respect human rights and promote social good.

8. Conclusion

This paper has presented a comprehensive framework for audio-visual anomaly detection in long surveillance videos, centered on a context-aware temporal modeling architecture that integrates dual-stream encoding, hierarchical memory, and cross-modal attention. We have examined the structural trade-offs inherent in designing such a system, including the balance between temporal resolution and computational cost, the infrastructure requirements for scalable deployment, and the sustainability implications of continuous processing. The discussion of fairness, bias, and robustness has highlighted the need for proactive measures to ensure equitable performance across diverse conditions and populations, while the policy analysis has underscored the importance of transparency, oversight, and community engagement. As surveillance technologies continue to advance, the research community must maintain a dual focus on algorithmic innovation and responsible governance, recognizing that the ultimate value of these systems lies not only in their technical capabilities but in their contribution to safer, more just, and more inclusive societies.

References

1. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6479–6488.

2. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 733–742.

3. Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279–283.

4. Ramachandran, P., & Sari, L. (2021). Multimodal anomaly detection for surveillance video using audio-visual fusion. IEEE Transactions on Information Forensics and Security, 16, 4120–4133.

5. Luo, W., Liu, W., & Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked RNN framework. Proceedings of the IEEE International Conference on Computer Vision, 341–349.

6. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1452–1464.

7. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. International Conference on Information Processing in Medical Imaging, 146–157.

8. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., & Wu, Y. (2016). Learning fine-grained image similarity with deep ranking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1386–1393.

9. Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 4037–4058.

10. Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. IEEE International Workshop on Machine Learning for Signal Processing, 1–6.

11. Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2016). ModDrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8), 1692–1706.

12. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. International Conference on Machine Learning, 2048–2057.

13. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

15. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

16. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations.

17. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

18. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

19. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646.

20. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

21. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the ACM Conference on Computer and Communications Security, 308–318.

22. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 77–91.

23. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). Dataset shift in machine learning. MIT Press.

24. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the Conference on Fairness, Accountability, and Transparency, 33–44.

25. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

26. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

27. Mead, C. (1990). Neuromorphic electronic systems. Proceedings of the IEEE, 78(10), 1629–1636.