Secure Video Surveillance via Hierarchical Multi-Stream Motion Encoding: A HY-Himmel Technical Report Extension

Pavel Riley
School of Computing, Clemson University, Clemson, SC, USA.
riley1997@clemson.edu

Claudio Bailey
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
claudio.work@ucf.edu

Tarun C. Krishnan
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
tarunkrishnan83@uab.edu

Mikko D. Allen
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
allen1997@missouri.edu

Abstract

The proliferation of video surveillance systems in public and private spaces has created an urgent demand for secure, efficient, and scalable motion analysis frameworks. Traditional single-stream motion encoding methods often suffer from limited temporal resolution, high computational overhead, and vulnerability to adversarial perturbations, thereby compromising both real-time performance and trustworthiness. This paper presents an extended analysis of the HY-Himmel hierarchical interleaved multi-stream motion encoding architecture, focusing on its system-level implications for secure video surveillance. Rather than detailing low-level algorithmic innovations, we examine the architectural trade-offs among accuracy, latency, memory usage, and energy consumption that arise from the hierarchical decomposition of motion into multiple temporal streams. We discuss deployment considerations including edge-cloud partitioning, bandwidth constraints, hardware acceleration, and real-time throughput requirements. The robustness of the architecture is evaluated in the context of adversarial attacks, lighting variations, and occlusions, drawing on recent empirical studies of video model resilience. Furthermore, we address critical socio-technical issues such as demographic bias in motion-based recognition, privacy preservation in public surveillance, and the need for transparent governance frameworks. By integrating technical design decisions with policy and fairness considerations, this paper provides a holistic view of how hierarchical multi-stream motion encoding can be operationalized in secure, responsible video surveillance systems. Our analysis reveals that while hierarchical interleaving offers substantial gains in temporal modeling fidelity and compression efficiency, its success depends heavily on careful calibration of stream granularity, inter-stream fusion strategies, and the adoption of privacy-preserving data handling protocols. We conclude with

recommendations for future research directions that balance performance, robustness, and ethical accountability.

Keywords

video surveillance, motion encoding, hierarchical architecture, multi-stream processing, system security, fairness, edge deployment, adversarial robustness, privacy governance.

1. Introduction

Modern video surveillance infrastructure spans urban centers, critical infrastructure, transportation hubs, and private enterprises, generating vast quantities of motion data that require near-instantaneous analysis. The task of automatically detecting, tracking, and interpreting human activities from these streams has motivated a extensive body of research in computer vision and machine learning. Early approaches relied on handcrafted features and optical flow computation, but the advent of deep learning, particularly three-dimensional convolutional networks and video transformers, has dramatically improved recognition accuracy [1][2][3]. However, these advances have simultaneously introduced new challenges related to computational cost, latency, and security. Contemporary surveillance systems must operate under stringent real-time constraints while also being resilient to adversarial manipulation, environmental variability, and data privacy concerns.

The HY-Himmel technical report [8] proposed a novel hierarchical interleaved multi-stream motion encoding framework that addresses several of these challenges by decomposing video input into multiple temporal streams that capture motion information at different granularities and then interleaving them hierarchically. This architecture, inspired by prior multi-stream designs such as SlowFast networks and temporal segment networks [5][3], extends the concept by introducing a systematic interleaving mechanism that preserves both fine-grained motion cues and long-range temporal dependencies. While the original report focuses primarily on the algorithmic architecture and benchmark performance, the present paper seeks to extend that analysis by examining the system-level implications of deploying such a framework in real-world, secure video surveillance contexts. We move beyond accuracy metrics to explore trade-offs in computational efficiency, robustness, fairness, and governance.

The motivation for this extended analysis stems from the observation that many state-of-the-art video understanding models are designed and evaluated in controlled laboratory settings, far removed from the heterogeneity and adversarial conditions of operational surveillance environments. For example, high-resolution video feeds from outdoor cameras often encounter severe illumination changes, occlusion, and sensor noise, while indoor deployments must contend with cluttered background and partial views. Moreover, the increasing use of video surveillance for law enforcement and access control raises serious ethical questions about algorithmic bias and privacy infringement [13][14]. A system that performs well on curated benchmarks may fail spectacularly when faced with demographic imbalance or targeted adversarial patches. Thus, any motion encoding architecture intended for secure deployment must be scrutinized not only for its ability to recognize actions but also for its robustness, fairness, and compliance with emerging regulatory frameworks.

The paper is structured as follows. Section 2 situates our work within the broader landscape of video surveillance and motion encoding research. Section 3 provides a high-level description of the hierarchical multi-stream motion encoding architecture, emphasizing its structural innovations. Section 4 analyzes the key design trade-offs that arise from the hierarchical and multi-stream nature of the approach, including latency, memory, and accuracy interactions.

Section 5 addresses deployment and infrastructure considerations, focusing on edge-cloud partitioning, hardware acceleration, and scalability. Section 6 examines robustness to adversarial perturbations and environmental variations, as well as sustainability through energy-efficient computation. Section 7 turns to fairness and policy implications, discussing demographic bias, privacy risks, and governance models. Finally, Section 8 concludes with a summary of findings and directions for future research.

2. Related Work

Video surveillance systems have evolved from simple motion detection based on frame differencing to sophisticated deep learning pipelines capable of recognizing complex human activities. A foundational contribution was the introduction of two-stream convolutional networks, which separate spatial and temporal information by processing RGB frames and optical flow independently [2]. This design principle later expanded into multi-stream architectures that incorporate additional modalities such as depth, infrared, or audio. Meanwhile, temporal segment networks [3] proposed a sparse sampling strategy to model long-range temporal structure without overwhelming computational resources. More recent transformer-based video models, such as ViViT [6] and TimeSformer [7], apply self-attention across space and time, achieving state-of-the-art results on large-scale benchmarks. However, these models often require enormous computational budgets, making them unsuitable for real-time edge deployment.

The SlowFast network [5] introduced a two-stream design with different frame rates, where a slow pathway captures fine spatial semantics and a fast pathway encodes coarse temporal motion. This hierarchical approach to temporal granularity directly inspired the multi-stream philosophy of the HY-Himmel architecture [8]. Other relevant work includes the use of 3D convolutional networks like C3D [4], which learn spatiotemporal features jointly, but at a high computational cost. Efficient video understanding has also been pursued through model compression techniques such as pruning, quantization, and knowledge distillation [17][18], which are critical for on-device deployment.

Security and robustness in video surveillance have received increasing attention. Adversarial attacks can mislead classifiers by adding imperceptible perturbations to video frames, and defenses must account for the temporal dimension [12]. Additionally, real-world surveillance systems must handle occlusions, non-rigid motion, and variable lighting conditions. Studies have shown that deep video models can be brittle to such distributional shifts [16]. From a fairness perspective, commercial recognition systems have been demonstrated to exhibit significant accuracy disparities across demographic groups, particularly for gender and skin tone [14]. These findings underscore the need for rigorous auditing and mitigation strategies in surveillance deployments. Privacy-preserving techniques, such as anonymization through blurring or generative obfuscation, are also active research areas [15][20].

3. Hierarchical Multi-Stream Motion Encoding Architecture

The central design of the HY-Himmel framework [8] is the decomposition of a video sequence into multiple temporal streams, each operating at a distinct frame rate or sampling interval, and the hierarchical interleaving of these streams to produce a compact yet comprehensive motion representation. Unlike conventional multi-stream approaches that independently process each stream and fuse their outputs at the decision level, the HY-Himmel architecture introduces interleaving layers that integrate information across streams at multiple hierarchical stages. This enables the model to capture motions occurring at

different time scales—ranging from rapid micro-movements of limbs to slower body gestures—while maintaining a unified feature map.

Each stream in the hierarchy is parameterized by a temporal stride and a spatial resolution. Lower streams employ high temporal resolution (e.g., every frame) to register fine-grained motion, while higher streams operate at coarser temporal strides (e.g., every four or eight frames) to aggregate longer-range motion patterns. The streams are then interleaved using a learned fusion module that selectively combines temporal features from neighboring levels. This interleaving is hierarchical: early fusion occurs between adjacent low-level streams, and later stages fuse these intermediate representations with higher-level streams. The result is a multi-resolution temporal representation that can be passed to downstream classifiers or detectors.

One of the key architectural advantages is the reduction in computational redundancy. By processing high-resolution spatial data only in the slow stream and using the fast streams primarily for motion cues, the model achieves a favorable accuracy-efficiency trade-off compared to uniform spatiotemporal models [5][8]. Furthermore, the interleaved design allows for efficient parallelization across streams, which is beneficial for hardware accelerators. The hierarchical structure also facilitates early exit strategies: if a surveillance application requires only coarse action detection (e.g., presence of a person), the model can terminate computation at an intermediate level without processing all streams fully, thereby saving energy and reducing latency.

4. Structural Trade-Offs and System Design Considerations

Deploying a hierarchical multi-stream motion encoding system involves navigating a complex landscape of trade-offs among accuracy, latency, memory footprint, and energy consumption. One fundamental trade-off is the number of streams versus the computational budget. Increasing the number of streams enhances temporal granularity but multiplies the memory and compute requirements. The HY-Himmel architecture [8] calibrates this by using a small number of streams (typically three to five) with carefully chosen strides, but the optimal configuration depends on the specific surveillance scenario. For instance, wide-area outdoor monitoring may benefit from more streams to capture both pedestrian and vehicular motions, while indoor access control may require fewer streams focused on fine hand gestures.

The interleaving frequency is another critical parameter. Frequent interleaving enhances information flow between streams but introduces synchronization overhead and potential latency accumulation. In real-time surveillance systems, latency must remain below a few hundred milliseconds to enable timely alerts. Therefore, practitioners must decide whether to fuse streams after each temporal block or only at the end of the feature hierarchy. The HY-Himmel report demonstrates that hierarchical interleaving with moderate fusion depth yields the best balance on benchmark datasets [8]. However, our analysis suggests that this balance is sensitive to the video frame rate; high-frame-rate cameras (e.g., 120 fps) benefit from deeper interleaving, whereas standard 30 fps footage may require shallower fusion to avoid redundancies.

Memory constraints further complicate deployment on edge devices with limited RAM and flash storage. The multi-stream architecture stores intermediate feature maps for each stream and each interleaving level, potentially leading to a large memory footprint. Techniques such as feature quantization and compressed activation storage have been proposed [17][19], but they come with accuracy penalties. An alternative is to use the hierarchical structure to

discard unnecessary streams at runtime based on scene complexity; for example, if no motion is detected in a low-level stream, its features can be pruned. Such adaptive resource management is an active research area and has not been fully explored in the context of the HY-Himmel framework.

Energy consumption is increasingly important for battery-powered surveillance cameras. As noted by several studies, the power budget for embedded vision systems is typically on the order of a few watts [20]. Multi-stream architectures incur additional energy cost due to parallel computation and data movement. Hierarchical interleaving may reduce energy by allowing earlier termination, but the overhead of dynamic scheduling and the need for a always-active fast stream can offset these gains. We hypothesize that an energy-optimized variant would involve asymmetric processing: the fast streams run on low-power specialized accelerators, while the slow stream uses a more powerful but sparingly activated neural core.

5. Deployment and Infrastructure

The practical deployment of a hierarchical multi-stream motion encoding system for video surveillance requires careful integration with existing infrastructure, including camera networks, edge computing devices, cloud services, and communication channels. A common architecture involves on-camera edge processors that perform initial motion encoding and send compact features to a central server for analysis. The HY-Himmel framework [8] is well-suited to this paradigm because the hierarchical encoding produces compressed representations at each level, enabling progressive transmission. For example, a low-resolution motion vector can be transmitted first for quick detection, followed by higher-resolution features for detailed classification if needed.

Bandwidth is a primary constraint. Transmitting raw video over wireless networks is often infeasible due to limited throughput and high cost. The multi-stream approach can mitigate this by transmitting only motion-encoded features, which are orders of magnitude smaller than raw frames. However, the interleaving stage requires that all streams be available simultaneously at the fusion point, which may demand synchronization over the network. Edge-cloud partitioning can alleviate this: low-level fast streams are processed entirely on the edge, and only their fused outputs are sent to the cloud for higher-level reasoning. This hybrid strategy reduces bandwidth while preserving accuracy.

Hardware acceleration is another critical factor. The parallel streams of the HY-Himmel architecture map naturally to multi-core CPUs or GPUs, but for low-power embedded deployment, custom accelerators such as neural processing units (NPUs) with multi-stream dataflow support are needed. Recent FPGA implementations of multi-stream video processors have demonstrated significant energy savings [20]. The hierarchical interleaving modules, which involve element-wise additions and attention-like operations, benefit from systolic array architectures commonly found in modern TPUs. Nevertheless, the lack of unified software middleware for dynamic stream management poses a barrier to widespread adoption. Open-source frameworks like TensorFlow Lite and ONNX Runtime are gradually adding support for multi-stream models, but production-grade surveillance systems often require custom schedulers.

Scalability is a concern when thousands of cameras are deployed in a city-wide surveillance network. The hierarchical decomposition allows each camera to independently generate motion features, which are then aggregated by a central coordinator. However, the coordinator must handle the fusion of heterogeneous streams from different viewpoints,

lighting conditions, and camera models. The HY-Himmel framework, as described in the technical report [8], assumes controlled conditions; extending it to large-scale, heterogeneous settings requires domain adaptation and robust normalization techniques. Additionally, network latency becomes stochastic, and the synchronization of interleaved streams may introduce jitter. A robust deployment would incorporate buffering and dynamic time-warping algorithms to align the streams despite variable network delays.

6. Sustainability and Robustness

Sustainability in video surveillance encompasses both environmental and economic dimensions. The energy consumed by large-scale surveillance systems is substantial, contributing to carbon emissions and operational costs. Hierarchical multi-stream motion encoding can contribute to sustainability by reducing computational load compared to uniform high-resolution processing. Nevertheless, the overhead of multiple streams and interleaving may increase overall energy if not carefully tuned. Research on energy-efficient design [17] suggests that pruning unnecessary channels in the fast streams and using lower-precision arithmetic can halve energy consumption with minimal accuracy loss. Additionally, the ability to dynamically activate only relevant streams based on scene motion can further improve energy efficiency.

Robustness refers to the system's ability to maintain performance under diverse and challenging conditions. Video surveillance systems routinely encounter low-light environments, rain, fog, and camera shake. The hierarchical multi-stream architecture [8] inherently captures motion at multiple temporal scales, which can help disambiguate slow, subtle movements from noise. For instance, in low-light conditions where individual frames are noisy, the fast stream may produce spurious motion signals, but the slow stream can smooth them out over longer durations. However, the interleaving mechanism may inadvertently amplify noise if the fusion weights are not learned to be robust. Recent studies on adversarial robustness of video models [12][16] show that multi-stream designs are generally more resistant to patch-based attacks because the attacker must simultaneously perturb multiple streams; yet they remain vulnerable to universal perturbations that fool the fusion module. We recommend incorporating adversarial training and input sanitization into the deployment pipeline.

Another aspect of robustness is temporal consistency. Many long video sequences contain repetitive motion patterns or long stationary periods. The hierarchical representation can compress these sequences efficiently by representing static scenes with only the slow stream and discarding fast streams when motion is absent. This adaptive behavior not only saves compute but also avoids false alarms due to static objects. However, it requires a reliable motion trigger, which itself can be attacked. Therefore, the trigger mechanism must be hardened, perhaps using a lightweight background subtraction model.

7. Fairness and Policy Implications

The deployment of video surveillance systems powered by hierarchical motion encoding raises profound fairness and policy concerns. Research on algorithmic fairness has demonstrated that many commercial vision systems exhibit biased performance across demographic groups [14]. For motion-based analysis, bias can arise from differences in gait, body proportions, clothing, and cultural gestures. For example, a system trained predominantly on Western subjects may misclassify specific hand gestures used in other cultures. The multi-stream architecture, by capturing motion at multiple temporal scales,

might mitigate some of these biases by relying on broader motion patterns rather than subtle details, but it could also amplify biases if certain streams are more sensitive to demographic attributes.

Privacy is an equally pressing issue. Video surveillance inherently involves the capture of identifiable human subjects. The hierarchical encoding approach offers a potential privacy benefit: by transmitting only motion features rather than raw video, it reduces the risk of exposing private visual details. However, recent research has shown that motion features can be reverse-engineered to reconstruct body pose and even identity [15]. Thus, the system must incorporate privacy-preserving mechanisms such as differential privacy, anonymization, or on-device processing that never leaves the edge. Policies governing data retention, access control, and audit trails are necessary to ensure accountability. The European Union's General Data Protection Regulation and similar frameworks require that video data be processed only for specified legitimate purposes and with appropriate technical safeguards.

Governance models for large-scale surveillance networks must involve multiple stakeholders: public safety agencies, civil liberties organizations, technology providers, and affected communities. Transparency about the system's capabilities and limitations, including known biases, is essential. The HY-Himmel architecture, by its hierarchical nature, could be designed to provide interpretable outputs at each level, such as saliency maps indicating which temporal features contributed to a decision. This interpretability can support audits and help detect when the system makes erroneous or biased predictions. We advocate for the integration of fairness metrics and adversarial testing into the development lifecycle, as well as periodic re-evaluation after deployment.

8. Conclusion

This paper has provided an extended system-level analysis of the HY-Himmel hierarchical interleaved multi-stream motion encoding architecture for secure video surveillance. By examining the structural trade-offs among accuracy, latency, memory, and energy, we have shown that the framework offers promising efficiency gains but requires careful calibration for each deployment context. The hierarchical design enables flexible deployment across edge and cloud infrastructures, though challenges remain in bandwidth management, hardware acceleration, and synchronization. Robustness to environmental variability and adversarial attacks can be improved through adaptive stream pruning and adversarial training, but further research is needed to understand vulnerabilities in the fusion mechanism. Fairness and privacy considerations demand that the system be transparent, auditable, and designed with safeguards against bias and data misuse. Future work should focus on developing standardized benchmarks for system-level metrics, exploring dynamic stream configuration policies, and integrating governance protocols directly into the architecture. As video surveillance becomes ubiquitous, the responsible deployment of hierarchical motion encoding systems will depend not only on algorithmic innovation but also on a deep commitment to ethical design and societal accountability.

References

1. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).

2. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (Vol. 27).

3. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision (pp. 20–36). Springer.

4. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4489–4497).

5. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6202–6211).

6. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6836–6846).

7. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning (pp. 813–823). PMLR.

8. 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.

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

10. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255).

11. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (Vol. 25).

12. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226).

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

15. Winkler, T., & Rinner, B. (2020). Security and privacy in video surveillance: A survey. ACM Computing Surveys, 53(5), 1–40.

16. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness and accuracy: A computational trade-off. In Proceedings of the International Conference on Machine Learning (pp. 6311–6321). PMLR.

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

18. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., … & He, K. (2019). Accurate, large minibatch SGD: Training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677.

19. Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). XNOR-Net: ImageNet classification using binary convolutional neural networks. In European Conference on Computer Vision (pp. 525–542). Springer.

20. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., … & Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.