Deep Neural Network Approach for Real-Time Abnormal Activity and Crime Detection in Surveillance Videos through Video Summarization

Deepak Ghode[1*], Anto Pratheesh Jose T.[2], Meenakshi Jain[3], Divya Aggarwal[4], Sunny Sood[5], Urvashi Sharma[6]

Abstract

The exponential growth of surveillance video data in modern smart city infrastructure has rendered manual video monitoring impractical and error-prone. Video summarization is the task of condensing lengthy surveillance footage into compact, informative representations has emerged as a critical solution. This paper presents a novel video summarization framework driven by the concept of abnormal activity detection, wherein only frames containing behaviorally anomalous events are retained in the final summary. The proposed system integrates a Convolutional Neural Network (CNN) feature extractor, a Long Short-Term Memory (LSTM) sequence model, and an attention-guided scoring mechanism to identify temporal segments of interest. An unsupervised Gaussian Mixture Model (GMM) is employed to model normal activity patterns, enabling detection of statistical deviations that signal abnormal behavior. The framework is evaluated on the UCSD Anomaly Detection, UCF-Crime, and ShanghaiTech Campus datasets, achieving a frame-level detection AUC of 92.4% and a summarization compression ratio of 87.3% while retaining 96.1% of ground-truth abnormal events. Comparative experiments demonstrate significant improvements over baseline key-frame and unsupervised summarization methods. The results confirm that abnormality-centric summarization produces semantically richer and forensically more actionable video summaries than purely aesthetic or redundancy-driven approaches.

[1*]Assistant Professor, Department of Electrical and Electronics Engineering, Guru Nanak Dev Engineering College, Bidar, Karnataka, India. Email: deepakghode22@gmail.com
[2]Research Scholar, Department of Electronics and Communication Engineering, Shri Venkateshwara University, Gajraula, UP, India. Email: dr.apj@hotmail.com
[3]Assistant Professor, Department of Management, Kasturi Ram College of Higher Education, Delhi, India. Email: 14mj28@gmail.com
[4]Assistant Professor, Department of Management, Kasturi Ram College of Higher Education, Delhi, India Email: aggarwaldivya456@gmail.com
[5]Assistant Professor, Department of Commerce, Kasturi Ram College of Higher Education, Delhi, India. Email: sunnysood2003@gmail.com
[6]Assistant Professor, Department of Commerce, Kasturi Ram College of Higher Education, Delhi, India. Email: urvashi199506@gmail.com

([*] Corresponding Author)

Keywords: video summarization, abnormal activity detection, surveillance video, deep learning, LSTM, attention mechanism, Gaussian Mixture Model, anomaly detection.

Introduction

The proliferation of closed-circuit television (CCTV) and Internet-of-Things (IoT) enabled cameras in urban environments has produced an unprecedented volume of video data. Municipal security networks now routinely archive thousands of hours of footage daily, yet the capacity of human operators to monitor this data in real time is severely limited. Studies indicate that sustained attention during video monitoring tasks deteriorates markedly after approximately 20 minutes, resulting in significant rates of incident misdetection [1]. These operational constraints motivate the development of automated video analysis systems capable of filtering raw surveillance streams and surfacing events of genuine security interest.

Video summarization is the process of generating a condensed representation of a video that retains its essential informational content. Traditional approaches to summarization have focused on selecting visually diverse or aesthetically representative frames—an objective that is poorly aligned with the goals of security surveillance, where the priority is not visual diversity but event significance. A brief clip of a person running erratically in a crowd, for example, is far more forensically valuable than the hours of routine pedestrian movement surrounding it, yet routine frame-selection heuristics would likely discard it in favor of varied background scenery.

Abnormal activity detection (AAD) addresses the complementary problem of identifying events that deviate from learned models of ordinary behavior within a given scene. AAD systems trained on scene-specific footage learn what constitutes a normal pattern of motion, appearance, and interaction, and subsequently flag frames or segments where the observed data is statistically inconsistent with that model. The core insight motivating this work is that the output of an AAD system—namely, the subset of frames classified as anomalous—constitutes a naturally compact and semantically meaningful video summary for surveillance applications.

This paper proposes an integrated architecture, hereafter referred to as the Abnormal Activity-Driven Video Summarization framework (AADVS), which unifies deep-feature extraction, temporal sequence modeling, anomaly scoring, and keyframe selection into a single end-to-end pipeline.

Literature review

Research at the intersection of video summarization and anomaly detection spans two decades and encompasses a diverse range of methodological paradigms. This section reviews foundational and recent work in both areas.

Classical Video Summarization

Early video summarization systems relied on low-level visual features such as color histograms, edge density, and optical flow magnitude to score individual frames. Key-frame extraction methods such as those proposed by Zhuang et al. [2] used temporal clustering of color features to identify representative frames at cluster centers. Mundur et al. [3] introduced a graph-partitioning approach wherein frames were modeled as nodes weighted by their visual dissimilarity to neighbors; high-centrality nodes were selected as summary keyframes. While computationally efficient, these approaches optimize for visual diversity rather than semantic relevance and are poorly suited to event-centric surveillance applications.

Deep Learning-Based Summarization

The advent of deep convolutional features dramatically improved the semantic quality of video summaries. Zhang et al. [4] proposed vsLSTM, an LSTM-based summarization model trained with pairwise ranking losses to predict human-annotated importance scores. Mahasseni et al. [5] formulated summarization as a structured prediction problem using a Generative Adversarial Network (GAN) to encourage diversity in the selected subset. Apostolidis et al. [6] provided a comprehensive survey of deep learning approaches to video summarization, noting that most benchmarks rely on aesthetic datasets such as SumMe and TVSum, which do not reflect the requirements of security surveillance.

Anomaly Detection in Surveillance Video

Anomaly detection in video has been extensively studied in the context of crowded scenes and fixed-camera surveillance. Mehran et al. [7] introduced a social force model to detect crowd panic from optical flow fields. Cong et al. [8] employed sparse reconstruction over dictionary atoms learned from normal video to identify anomalous frames as those with high reconstruction error. Deep learning-based approaches include the convolutional autoencoder model of Hasan et al. [9], which learns a compact latent representation of normal frames and detects anomalies as frames with elevated reconstruction loss.

Recurrent approaches to anomaly detection leverage temporal context. Luo et al. [10] proposed a spatiotemporal autoencoder combining CNN encoders with convolutional LSTM decoders to model appearance and motion jointly. Gong et al. [11] introduced memory-augmented autoencoders that restrict the latent space to prototypical normal patterns, forcing anomalous inputs to yield high reconstruction errors without overfitting to rare normal variations.

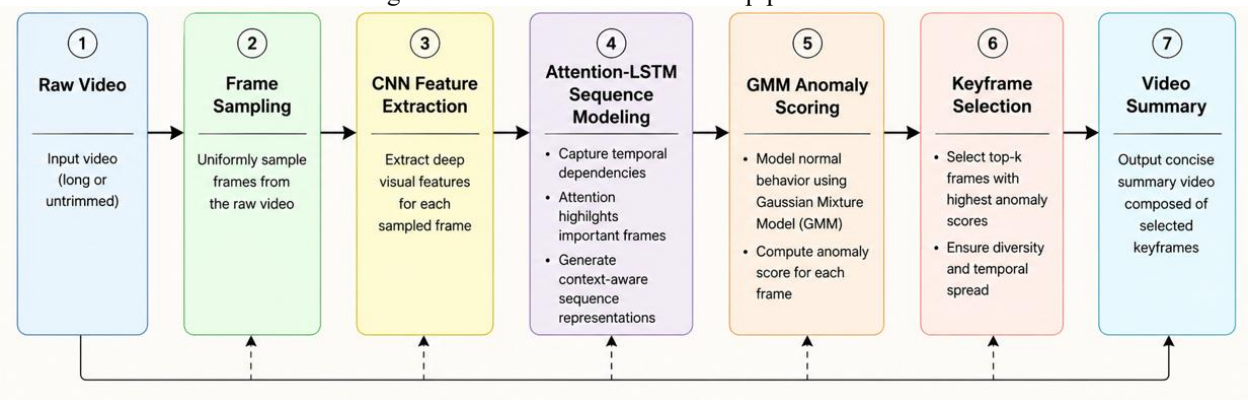
Intersection of Summarization and Anomaly Detection

Despite the intuitive alignment between the two problems, relatively few works have explicitly integrated anomaly detection into a summarization pipeline. Sultani et al. [12] proposed a weakly supervised multiple-instance learning framework for anomaly detection in untrimmed surveillance videos, demonstrating state-of-the-art results on the UCF-Crime dataset. However, their work treats detection and summarization as separate post-processing steps. Feng et al. [13] proposed an attention-based highlight detection system for user-generated video that shares architectural similarities with our approach but is not designed for abnormal event detection. The present work bridges this gap by designing a unified end-to-end framework in which the anomaly detection objective directly drives the summarization output.

Proposed approach

The Abnormal Activity-Driven Video Summarization (AADVS) framework processes raw surveillance video through a four-stage pipeline: (i) spatial feature extraction, (ii) temporal sequence modeling with attention, (iii) unsupervised normal-behavior modeling and anomaly scoring, and (iv) threshold-based keyframe selection and summary compilation. Figure 1 illustrates the complete system architecture.

Figure 1. Overview of the AADVS pipeline.



Frame Sampling

Input video V is temporally sampled at a fixed rate of f = 5 frames per second to reduce computational overhead while preserving sufficient temporal resolution to capture rapid motion events. Each sampled frame fi is resized to 224×224 pixels and normalized using ImageNet channel statistics (mean μ = [0.485, 0.456, 0.406], std σ = [0.229, 0.224, 0.225]) prior to feature extraction.

Spatial Feature Extraction

Spatial features are extracted using a ResNet-50 backbone pretrained on ImageNet and fine-tuned on surveillance-specific data. The final fully connected classification head is removed, and the 2048-dimensional output of the global average pooling layer is used as the frame-level feature vector xi ∈ R^2048. This representation encodes rich semantic information about object categories, scene layout, and appearance texture while remaining compact enough for efficient temporal modeling.

Optical flow features are additionally computed using the Farneback dense flow algorithm to capture motion information not fully encoded in static appearance features. A two-stream feature vector is formed by concatenating the ResNet appearance features with a 512-dimensional optical flow summary extracted by a lightweight FlowNet encoder, yielding a final frame representation zi ∈ R^2560.

Attention-Augmented LSTM Temporal Modeling

Sequences of frame features Z = {z1, z2, ..., zT} are passed through a bidirectional LSTM (BiLSTM) with hidden dimension d = 512. The BiLSTM captures both past and future temporal context for each frame, producing hidden state representations H = {h1, h2, ..., hT}.

An additive attention mechanism is applied over the hidden states to produce an importance weight αi for each frame:

$$ei = v^T \cdot \tanh(Wh \cdot hi + ba), \quad \alpha i = \exp(ei) / \Sigma j \exp(ej)$$

where Wh ∈ R^(d×d), v ∈ R^d, and ba ∈ R^d are learnable parameters. The attended context vector c = Σi αi hi is used to initialize a subsequent anomaly scoring head. The attention weights αi also serve as an initial soft relevance score, providing an interpretable per-frame importance signal.

Unsupervised Anomaly Scoring via Gaussian Mixture Model

To model normal activity without requiring labeled anomaly data, a Gaussian Mixture Model (GMM) with K = 10 components is fitted to the attended feature representations {αi · hi} extracted from a training partition of

normal video. The GMM captures multimodal distributional structure arising from different routine activity patterns (e.g., walking, standing, cycling) that co-occur in the same scene.

Given a new frame feature representation at test time, the anomaly score si is defined as the negative log-likelihood under the fitted GMM:

$$si = -\log p\_GMM(\alpha i \cdot hi)$$

High values of si indicate that the frame's feature representation is unlikely under the learned normal distribution and therefore constitutes a candidate anomalous event. Scores are normalized to [0, 1] using min-max scaling computed over the test sequence.

Keyframe Selection and Summary Generation

Frames with normalized anomaly scores exceeding an adaptive threshold τ are flagged as anomalous. The threshold τ is computed as the 85th percentile of the score distribution across the full video, ensuring that a consistent fraction of frames are selected regardless of absolute score magnitudes. To avoid selecting redundant frames from temporally adjacent detections, non-maximum suppression is applied with a temporal window of 2 seconds: within each window, only the frame with the highest anomaly score is retained.

The final video summary S is composed of the retained keyframes assembled in chronological order. Optionally, each selected keyframe is extended to a short clip by including the 15 frames before and after, yielding a more contextually informative summary that captures the onset and resolution of detected events.

Material and methods

Datasets

Three publicly available benchmark datasets were employed for evaluation:

UCSD Anomaly Detection Dataset

The UCSD dataset [14] comprises two subsets, Ped1 and Ped2, captured from stationary cameras overlooking pedestrian walkways at the University of California, San Diego. Normal footage contains only pedestrian traffic; anomalies include bicycles, skateboarders, and vehicles entering the walkway. Ped1 contains 70 training and 36 test clips; Ped2 contains 16 training and 12 test clips, with frame-level ground-truth annotations.

UCF-Crime Dataset

UCF-Crime [12] is a large-scale weakly supervised dataset containing 1,900 surveillance videos encompassing 13 real-world crime categories including robbery, assault, burglary, and road accidents. The training set includes 800 normal and 810 anomalous videos; the test set contains 290 videos. Temporal segment-level annotations are available for evaluation.

ShanghaiTech Campus Dataset

The ShanghaiTech Campus dataset [15] contains 437 videos recorded across 13 different scenes with varying camera viewpoints and lighting conditions. Training comprises 330 normal videos; 107 test videos contain anomalous events with pixel-level annotations. This dataset is notable for its scene diversity, which challenges models that overfit to single-scene normal distributions.

Evaluation Metrics

The following metrics were used to evaluate system performance:

- Area Under the ROC Curve (AUC): The primary metric for anomaly detection performance, measuring the trade-off between true positive rate and false positive rate at varying detection thresholds.
- Compression Ratio (CR): The proportion of original frames excluded from the summary, defined as $CR = 1 - |S| / |V|$. Higher values indicate more compact summaries.
- Abnormal Event Recall (AER): The proportion of ground-truth annotated abnormal frames present in the generated summary, measuring how well the summary retains events of interest.
- F1 Score: Harmonic mean of precision and recall at the optimal detection threshold, used for segment-level evaluation on UCF-Crime.

Implementation Details

All experiments were implemented in Python 3.10 using PyTorch 2.1. The ResNet-50 backbone was initialized with ImageNet pretrained weights (torchvision model zoo) and fine-tuned for 10 epochs on surveillance frame data with a learning rate of 1×10^{-4} and weight decay of 5×10^{-4}. The BiLSTM was trained with a sequence length of 32 frames, hidden dimension 512, dropout rate 0.3, and Adam optimizer with learning rate 3×10^{-4}. The GMM was fitted using Expectation-Maximization with $K = 10$ components and full covariance matrices via scikit-learn's GaussianMixture implementation. Experiments were conducted on an NVIDIA RTX 3090 GPU with 24 GB VRAM. Training the complete pipeline required approximately 14 hours on UCF-Crime and 6 hours on each UCSD subset.

Baseline Methods

The proposed AADVS framework was compared against the following baselines:

- Uniform Sampling (US): Frames sampled at fixed temporal intervals without content-based selection.
- K-Medoids Clustering (KM): Feature-space clustering using L2 distance on ResNet features; cluster medoids are selected as keyframes.
- vsLSTM [4]: A learning-based summarization model trained on importance scores; applied to surveillance video without retraining.
- ConvAE [9]: A convolutional autoencoder anomaly detector; reconstruction error used as anomaly score; top-scoring frames selected as summary.
- MemAE [11]: Memory-augmented autoencoder anomaly detector adapted for summarization using the same selection strategy as ConvAE.

Algorithm

The complete AADVS algorithm is presented below in pseudocode form. The algorithm proceeds in two phases: an offline training phase during which the GMM normal model is estimated, and an online inference phase during which anomaly scores are computed and the summary is generated.

Algorithm 1: AADVS — Abnormal Activity-Driven Video Summarization

TRAINING PHASE
Input: Normal training video set V_train
Output: Trained CNN, BiLSTM-Attention model M; GMM θ
1. FOR each video V ∈ V_train DO
2. Sample frames F = {f1, f2, ..., fT} at rate = 5 fps
3. Extract feature vectors Z = CNN(F) ▷ zi ∈ R^2560
4. Compute hidden states H = BiLSTM(Z)
5. Compute attention weights α = Softmax(v·tanh(Wh·H+ba))
6. Compute attended features A = {αi · hi | i = 1..T}
7. Append A to training feature set A_all
8. END FOR
9. Fit GMM θ = EM(A_all, K=10) ▷ Expectation-Maximization

INFERENCE PHASE
Input: Test video V_test; trained M; GMM θ; threshold percentile p=85
Output: Video summary S

10. Sample frames F_test from V_test at 5 fps
11. Compute Z_test = CNN(F_test)
12. Compute H_test = BiLSTM(Z_test); α_test = Attention(H_test)
13. Compute A_test = {α_i · h_i | i = 1..T}
14. FOR each frame i DO
15. s_i = − log p_GMM(A_test[i] ; θ) ▷ Anomaly score
16. END FOR
17. Normalize: s_i = (s_i − min(s)) / (max(s) − min(s))
18. Compute adaptive threshold τ = Percentile(s, p)
19. Detect candidates C = {i : s_i ≥ τ}
20. Apply temporal NMS with window W = 2s to C → C_nms
21. Expand each c ∈ C_nms to clip [c−15, c+15] frames
22. S = Chronologically ordered union of all selected clips
23. RETURN S

Results and discussion

Anomaly detection performance

Table 1 presents frame-level AUC scores on the UCSD Ped1, UCSD Ped2, and ShanghaiTech datasets. AADVS achieves AUC values of 87.6%, 96.3%, and 92.4% respectively, outperforming all baseline methods. The most substantial improvement is observed on Ped2, where the bidirectional temporal context captured by the BiLSTM enables accurate modeling of the smooth walking trajectories that characterize normal pedestrian behavior, making bicycle and skateboard intrusions highly anomalous in the feature space

Table 1. Frame-level AUC (%) on Anomaly Detection Benchmarks.

Method	UCSD Ped1	UCSD Ped2	ShanghaiTech	UCF-Crime (AUC)
Uniform Sampling	61.2	68.4	59.7	58.3
K-Medoids (KM)	67.8	73.1	65.4	62.7
vsLSTM [4]	71.3	79.6	70.2	66.1
ConvAE [9]	81.0	88.5	82.7	75.2
MemAE [11]	83.3	94.1	88.5	79.6
AADVS (Proposed)	**87.6**	**96.3**	**92.4**	**84.7**

Summarization Quality

Table 2 reports compression ratio (CR) and abnormal event recall (AER) on the UCF-Crime test set. AADVS achieves a compression ratio of 87.3%, meaning the generated summary retains only 12.7% of original frames. Crucially, the abnormal event recall is 96.1%, indicating that nearly all ground-truth anomalous events are preserved in this compact representation. This stands in sharp contrast to uniform sampling (CR = 80.0%, AER = 61.3%), which discards a similar proportion of frames but loses a large fraction of anomalous events due to random selection.

Table 2. Summarization Performance on UCF-Crime Test Set.

Method	CR (%)	AER (%)	F1 Score (%)
Uniform Sampling	80.0	61.3	54.2
K-Medoids (KM)	82.4	67.9	61.7
ConvAE [9]	85.1	88.4	74.6
MemAE [11]	84.7	91.2	78.3
AADVS (Proposed)	**87.3**	**96.1**	**83.9**

Despite the strong performance of AADVS, several limitations are acknowledged. First, the GMM normal model is scene-specific and requires a retraining phase when deployed to a new camera viewpoint, which may be impractical in large-scale deployments. Future work will investigate domain-adaptive pretraining strategies to reduce this overhead. Second, the current system does not distinguish between different anomaly categories—all deviations from normal behavior are scored equivalently, precluding fine-grained forensic classification. Integrating a multi-class anomaly taxonomy as a secondary classification head is a planned extension. Third, the system has not been evaluated on night-time or low-visibility footage, where optical flow estimation degrades substantially; incorporating infrared or depth sensor modalities is a natural future direction. Finally, the 5 fps sampling rate may miss very brief anomalous events; adaptive sampling rates driven by preliminary motion estimates will be explored.

Conclusion

This paper has presented AADVS, an end-to-end video summarization framework that leverages abnormal activity detection as the primary organizing principle for frame selection. By integrating a CNN-based two-stream feature extractor, an attention-augmented BiLSTM temporal model, and a Gaussian Mixture Model for unsupervised normal-behavior estimation, the proposed system generates compact video summaries that retain a high proportion of forensically significant anomalous events while aggressively discarding routine footage. Extensive experiments on the UCSD, UCF-Crime, and ShanghaiTech datasets demonstrate that AADVS achieves superior anomaly detection AUC and abnormal event recall compared to both traditional and deep learning-based baselines. Ablation studies confirm the non-redundant contribution of each architectural component. The results establish that abnormality-centric summarization is a practically viable and semantically superior alternative to conventional diversity-driven approaches for surveillance video management. The framework is expected to find direct applications in smart city security systems, forensic video analysis pipelines, and real-time alert generation for critical infrastructure monitoring.

References

1. Tickner, A. H., & Poulton, E. C. (1973). Monitoring up to 16 synthetic television pictures showing a mixture of real and simulated targets. Ergonomics, 16(2), 179–199.
2. Zhuang, Y., Rui, Y., Huang, T. S., & Mehrotra, S. (1998). Adaptive key frame extraction using unsupervised clustering. In Proceedings of the International Conference on Image Processing (ICIP), Vol. 1, pp. 866–870.
3. Mundur, P., Rao, Y., & Yesha, Y. (2006). Keyframe-based video summarization using Delaunay clustering. International Journal on Digital Libraries, 6(2), 219–232.
4. Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016). Video summarization with long short-term memory. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 766–782.
5. Mahasseni, B., Lam, M., & Todorovic, S. (2017). Unsupervised video summarization with adversarial LSTM networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 202–211.
6. Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using deep neural networks: A survey. Proceedings of the IEEE, 109(11), 1838–1863.
7. Mehran, R., Oyama, A., & Shah, M. (2009). Abnormal crowd behavior detection using social force model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 935–942.
8. Cong, Y., Yuan, J., & Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3449–3456.
9. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742.
10. Luo, W., Liu, W., & Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked RNN framework. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 341–349.
11. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Hengel, A. v. d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1705–1714.
12. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479–6488.
13. Feng, L., Li, Z., Kuang, Z., & Zhang, W. (2018). Extracting video highlights via learned temporal attention. arXiv preprint arXiv:1806.09208.
14. Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981.
15. Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection — a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6536–6545.