Revolutionizing Video Surveillance: AI-Powered Anomaly Detection

Nishant Deheriya
MTech Student
NRI Group Bhopal
MP, INDIA 462021

Devendra Kumar Bajpai, PhD
HOD CSE NIRT
NRI Group Bhopal
MP, INDIA 462021

P.K. Sharma, PhD
Principal NIRT
NRI Group Bhopal
MP, INDIA 462021

ABSTRACT

The widespread deployment of surveillance cameras in public spaces such as airports, roadways, and financial institutions has led to the generation of massive volumes of video data. However, a significant portion of this footage is reviewed only after an incident occurs, making manual monitoring both inefficient and error-prone. Automated anomaly detection in video surveillance has thus emerged as a critical area of research, aiming to enhance security by identifying abnormal activities in real-time.

This study proposes a novel deep learning-based framework for anomaly detection in video streams by integrating an Inflated 3D Convolution Network (I3D-ResNet50) with deep Multiple Instance Learning (MIL). The model treats video sequences as collections of instances, where normal and anomalous segments are classified as negative and positive instances, respectively. Each video snippet is individually assessed using a fully connected neural network (NN) to compute an anomaly score, enabling precise identification of abnormal activities. To enhance feature extraction and generalization, the I3D-ResNet50 model is applied after performing 10-crop augmentations on the UCF-101 dataset, which comprises approximately 50 GB of video data spanning 15 types of anomalous events such as fighting, theft, and abuse, alongside normal activities.

Extensive experiments demonstrate the effectiveness of our approach, achieving an Area Under the Curve (AUC) score of 83.85% within just 10,000 iterations, significantly outperforming conventional anomaly detection techniques. The proposed model exhibits robustness in detecting subtle and complex anomalies in dynamic environments, making it well-suited for real-time surveillance applications. By reducing the dependency on manual monitoring and improving anomaly detection accuracy, this system has the potential to enhance public safety and security across various domains, including transportation hubs, banking institutions, and smart cities.

General Terms

Machine Learning, Computer Vision, Surveillance Systems, Deep Learning, Anomaly Detection, Artificial Intelligence, Pattern Recognition, Neural Networks, Security and Privacy

Keywords

Anomaly Detection, Multiple Instance Learning, Deep Learning, Video Surveillance, I3D-ResNet50, CNN, LSTM.

1. INTRODUCTION

The ongoing digital transformation, driven by rapid advancements in Artificial Intelligence (AI), continues to reshape various aspects of life, influencing fields such as Computer Vision (CV), smart agriculture, physics, drug discovery, social network analysis, and security. Among these applications, video surveillance plays a crucial role in ensuring public safety by monitoring real-world environments. However, the overwhelming volume of video data generated by security cameras poses significant challenges in manual monitoring, as human operators must continuously analyse footage to detect suspicious activities. This process is not only time-consuming but also prone to inaccuracies due to human fatigue and cognitive limitations. Anomalies in video surveillance refer to events that deviate from normal patterns, such as fighting, theft, or robbery. Anomaly detection (AD) in this domain has gained increasing attention due to its critical role in security and behavioural analysis. Despite significant progress in CV-based anomaly detection, existing methods face several limitations. Traditional approaches, such as rule-based systems and sparse coding techniques, often suffer from high false alarm rates due to their reliance on predefined normal event dictionaries. Moreover, the scarcity of annotated datasets, low-resolution footage from street cameras, and the high intra-class variability of anomalies further complicate the accurate identification of suspicious events.

Several challenges hinder the development of robust AD models for video surveillance:

1. **Data Imbalance**: Anomalous events constitute only a small fraction of the overall video content, making it difficult to train models effectively while filtering out large amounts of redundant data. This imbalance impacts classifier performance and computational efficiency.
2. **High Variability of Anomalies**: Unlike predefined objects in object detection tasks, anomalies exhibit diverse patterns and contexts, making it challenging to define discriminative features across different scenarios.
3. **Complexity of Video Data**: Unlike images, videos contain temporal dependencies that must be accounted for during analysis. While spatial features such as RGB histograms can capture static information, effective AD models must incorporate motion dynamics and sequential patterns to improve detection accuracy.

To address these challenges, we propose an advanced weakly supervised learning framework that integrates deep **Multiple Instance Learning (MIL)** with **I3D-ResNet50** to enhance anomaly detection in video streams. Our approach introduces a novel ranking function that refines MIL-based classification, ensuring more effective separation of normal and anomalous video instances. By leveraging the UCF-Crime dataset, which contains a diverse range of security-related incidents, we demonstrate that our method outperforms state-of-the-art approaches in terms of **Area Under Curve (AUC), Receiver Operating Characteristic (ROC) curve, and False Alarm (FA) rates**.

2. RELATED WORK

Anomaly detection (AD) in surveillance videos has been an extensively researched topic, primarily due to the inherent complexity of visual characteristics and the variations observed across different video classes. Existing approaches for AD in surveillance networks can be categorized into three primary methodologies: **sparse coding-based detection, unsupervised learning methods, and weakly supervised learning approaches**.

Sparse Coding-Based Anomaly Detection

Sparse coding-based anomaly detection techniques rely on learning a global dictionary from low-level feature representations of normal video frames. These methods are trained exclusively on normal videos, constructing a dictionary that encapsulates regular patterns. During the testing phase, video segments with high reconstruction errors are classified as anomalies, as they do not conform to the learned normal patterns. However, these approaches suffer from **high false alarm rates** since normal video scenes can exhibit significant variations, making it challenging to represent all normal patterns within a single dictionary.

Unsupervised Anomaly Detection

Unsupervised AD techniques, particularly those employing generative models, attempt to reconstruct normal video samples while minimizing reconstruction errors. These methods assume that abnormal instances, being absent from the training set, will exhibit **higher reconstruction errors** when processed by the model. While effective in certain scenarios, generative models often **overfit to normal data**, leading to poor generalization when encountering real-world anomalies. Moreover, the lack of a precise definition of "abnormality" within these methods can result in **ambiguity between normal and anomalous events**, limiting their effectiveness.

Weakly Supervised Anomaly Detection

Weakly supervised anomaly detection has demonstrated **significant improvements over unsupervised methods**, particularly by leveraging **video-level annotations** instead of frame-level labels. Since annotating anomalies at the frame level is costly and labour-intensive, recent research has focused on utilizing weakly labelled datasets to improve AD models.

Sultani et al. introduced a **deep anomaly ranking model** based on **Multiple Instance Learning (MIL)**, which assigns high anomaly scores to abnormal video segments. Their approach used **Deep Convolutional 3D (C3D) features** extracted from **UCF-Crime**, a large-scale dataset containing **128 hours of video footage** with various security-related incidents. Their MIL-based ranking framework achieved an **AUC of 75.41%**, surpassing conventional techniques.

Ullah et al. proposed an anomaly detection model that integrates **Convolutional Neural Networks (CNNs) and Bi-Directional Long-Short-Term Memory (BDLSTM)** networks. The CNN extracts spatial features, while the BDLSTM processes temporal dependencies, enabling improved classification of normal and abnormal events. Their method outperformed previous techniques, achieving **accuracy improvements of 3.41% and 8.09%** on the **UCF-Crime and UCF-Crime2Local datasets**, respectively.

Zhong et al. developed a **cascade model** employing a **Graph Convolutional Neural Network (GCNN)** to refine noisy labels before applying **optical flow-based motion prediction**. Additionally, they introduced a **pseudo-anomaly evaluation metric** to assess the generalization ability of AD models. Their approach achieved **state-of-the-art performance**, with **AUC scores of 88.9% on the Avenue dataset, 82.6% on Ped1, 97.7% on Ped2, and 70.7% on ShanghaiTech datasets**.

Tian et al. introduced the **Robust Temporal Feature Magnitude (RTFM) model**, which classifies normal and anomalous video snippets using weak labels. Their method utilizes **I3D or C3D** for feature extraction, followed by training a snippet classifier. When tested on **UCF-Crime, UCSD-Peds, and XD-Violence datasets**, the RTFM model demonstrated **higher detection accuracy and improved sensitivity to subtle anomalies** compared to conventional methods.

Yan et al. proposed **Spatiotemporal Collaboration-based Segmentation (STC-Seg)**, a weakly supervised framework designed for video segmentation. Their method integrates **optical flow and unsupervised depth estimation** to generate pseudo-labels for training deep networks. Additionally, they introduced a **puzzle loss function** to enhance mask generation, facilitating efficient **end-to-end training with box-level annotations**. Their model was validated on **KITTI MOTS and YouTube visualization datasets**, demonstrating adaptability in both image-level and video-level instance segmentation tasks.

Overall, recent advancements in weakly supervised AD, particularly those leveraging deep learning models such as **CNNs, LSTMs, and GCNNs**, have significantly enhanced anomaly detection performance. However, challenges such as **false alarms, high variability in anomalous patterns, and the trade-off between detection accuracy and computational efficiency** remain open research areas.

3. PROPOSED METHODOLOGY

This section outlines the proposed methodology for anomaly detection in surveillance videos, which consists of three distinct phases:

1. **Video Preprocessing**
2. **Feature Extraction and Anomaly Score Generation**
3. **Multiple Instance Learning (MIL)**

1. Video Preprocessing

This phase prepares raw surveillance videos for feature extraction and model training. In this study, we use the **UCF-Crime dataset** [11], which consists of videos with **fixed parameters** of **240 × 320 pixels resolution** and a **frame rate of 30 frames per second**.

To enable efficient processing and model training:

- Each training video is divided into **32 non-overlapping temporal snippets**, ensuring that each snippet captures a distinct time window within the video.
- These snippets are grouped into **positive and negative packets** based on the presence or absence of anomalies:
  - **Positive Packet (P_a):** Contains at least one anomaly. The snippets in this packet are denoted as **(a_1, a_2, a_3, ..., a_k)**, where *k* represents the number of snippets.
  - **Negative Packet (P_n):** Comprises normal video snippets, denoted as **(n_1, n_2, n_3, ...,**

n_k), ensuring that no anomalies are present.

This approach facilitates efficient learning by allowing the model to differentiate between normal and abnormal video segments.

2. Feature Extraction and Anomaly Score Computation

To effectively represent the spatial and temporal properties of video snippets, we utilize **I3D-ResNet50**, a **pretrained deep learning model** trained on the **Kinetics dataset** [12]. This model extends **2D convolutional networks (2D ConvNets) to three dimensions (3D ConvNets)**, a process referred to as **2D-ConvNet inflation**, as proposed in [13].

Feature Extraction Process:

1. **Pretrained I3D-ResNet50 Model**
    - Extracts **spatiotemporal features** from each video snippet.
    - Uses **3D convolutional layers** to capture motion dynamics.
    - **Outperforms traditional 2D-CNNs** in video analysis.
2. **Feature Augmentation**
    - **10-crop augmentation** is applied to enhance robustness.
    - Each video snippet is represented using **16-frame clips** averaged to derive a final feature vector.
3. **Feature Normalization and Dimension Reduction**
    - Features undergo **L2 normalization** to maintain scale consistency.
    - Extracted feature vectors have a fixed dimensionality of **2048D**.
    - A **three-layer Fully Connected Neural Network (FCNN)** processes these features:
        - **Layer 1:** 256 neurons
        - **Layer 2:** 16 neurons
        - **Layer 3:** 1 neuron (final anomaly score output)

Each snippet is assigned an **anomaly score** between **0 and 1**, where a higher score indicates a higher probability of anomalous behaviour.

3. Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) is employed to train the anomaly classifier in a **weakly supervised** manner. Unlike traditional supervised learning, where **frame-level anomaly labels** are required, MIL only relies on **video-level annotations**, significantly reducing annotation effort.

MIL-Based Training Approach:

- **Key Challenge:** Anomalous events occur at unknown time points within a video, making precise snippet-level labelling impractical.
- **Solution:** We treat each video as a **bag (packet)** of snippets and assume that at least one snippet in an anomalous video contains an anomaly.
- The **ranking loss function** is applied to train the model using only **video-level labels**.

The objective function optimizes anomaly detection by considering the highest-scoring snippet within each packet:

$$\min \frac{1}{m}\sum_{j=1}^{m}\overbrace{\max(0,1-Y_{P_j}(\max_{i\in P_j}(w.\phi(x_i))-b))}^{A}+\frac{1}{2}\|w\|^2 \qquad (1)$$

where:

- $Y\rho j$ is the label of each packet,
- $\emptyset(\chi)$ represents the feature extraction function,
- b is the bias term,
- m is the number of training samples,
- W is the weight parameter

4. Novel Deep Multiple Instance Learning (DMIL)

Anomaly detection (AD) is often formulated as an **outlier detection problem** rather than a strict classification task. The fundamental assumption is that **abnormal snippets should have higher anomaly scores than normal snippets**. To enforce this principle, a **ranking loss** function is applied:

$$f(I_a) > f(I_n)$$

where:

- Ia represents an anomalous video,
- In denotes a normal video,
- $\int$ (I) is the predicted anomaly score, ranging from **0 to 1**.

5. Handling False Alarms in MIL Training

To improve reliability, false alarm scenarios are explicitly addressed:

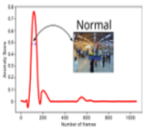
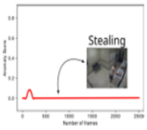
1. **False Anomaly Alerts (Type I Error):**
    - Occurs when the model **incorrectly classifies normal events as anomalies**.
2. **False Normal Alerts (Type II Error):**
    - Occurs when the model **fails to detect actual anomalies**.

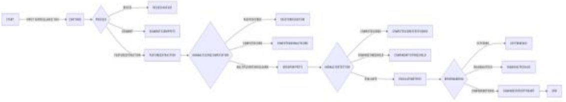
To mitigate these errors, the **ranking loss function** is further refined. The objective function ensures that:

- The **highest-ranking snippet** from an anomaly packet is assigned a **higher score** than the highest-ranking snippet from a normal packet.
- Within an anomalous packet, the **most anomalous snippet** receives a higher score than the least anomalous snippet.

This optimization minimizes both **false positives and false negatives**, improving the model's robustness.

Table 1 Examples of false warning cases

No	Cases	Description	Example
1	False warning	Predicts that a normal event will turn out to be abnormal.	
2	False warning	predicts that an abnormal event will turn out to be normal.	



Flow Chart

4. RESULTS AND DISCUSSION

Numerous Anomaly detection (AD) is a challenging task, and several standard datasets are available for evaluating different approaches. To assess the effectiveness of the proposed method, extensive experiments were conducted using the **UCF-Crime** dataset, a widely recognized benchmark for anomaly detection. Developed by Sultani et al. [11], this dataset provides a diverse range of real-world anomalies, including **robbery, fighting, and burglary**, making it suitable for evaluating anomaly detection models in various scenarios. The dataset consists of **128 hours of video**, divided into **1610 training videos** and **290 testing clips**, as detailed in Table 2. Unlike some other AD datasets, UCF-Crime includes only **video-level labels**, adding to the challenge by eliminating frame-level supervision.

Compared to other anomaly detection datasets, **UCF-Crime is significantly larger and more complex**, featuring extensive intra-class variations, diverse backgrounds, varying camera angles, and different lighting conditions. These factors contribute to the difficulty of learning robust anomaly detection patterns.

Training Phase

To prepare the data for training, **each video is segmented into 32 unique, non-overlapping snippets**, with each snippet treated as an independent case within a packet. The choice of **32 snippets per video** was determined experimentally to ensure a balanced representation of both normal and abnormal events.

Feature extraction is performed using the **I3D-ResNet50** model, which processes video frames of **240 × 320 pixels** at **30 frames per second**. This model, pretrained on the **Kinetics dataset**, is particularly effective in capturing **spatiotemporal features** due to its ability to leverage 3D convolutions. To enhance robustness, **10-crop augmentation** is applied, and the extracted I3D features are computed for each **16-frame video snippet**. These features undergo **L2 normalization** to ensure scale invariance.

Each video snippet is represented as a **2048-dimensional feature vector**, which is then passed through a **three-layer Fully Connected Neural Network (FCNN)**. The architecture of the FCNN is as follows:

- **First layer**: 256 units
- **Second layer**: 16 units
- **Final layer**: 1 unit (outputs an anomaly score)

To prevent overfitting, a **dropout regularization rate of 30%** is applied between the FCNN layers. The **ReLU activation function** is used in both the first and last FCNN layers to introduce non-linearity and improve learning efficiency.

During training, each mini-batch consists of **30 positive and 30 negative packets**, which are randomly sampled to ensure a balanced learning process.

Testing Phase

In the testing phase, each video is resized to **240 × 320 pixels** at **30 fps** and divided into **32 non-overlapping snippets**, consistent with the training procedure. The extracted feature representations are passed through the trained **FCNN**, which generates an **anomaly score** for each video snippet. These anomaly scores indicate the likelihood of a segment containing anomalous activity, as illustrated in **Figure 1**.

The effectiveness of the proposed approach is further analyzed by computing performance metrics such as **Area Under the Curve (AUC)** and **Frame-level Precision-Recall**, which will be discussed in the following subsections.

Table 2 UCF-Crime dataset details

	Total	Train (85%)	Test (15%)
Anomaly	950	810 (85%)	140 (15%)
Normal	950	800 (84%)	150 (16%)
Total	**1900**	**1610**	**290**

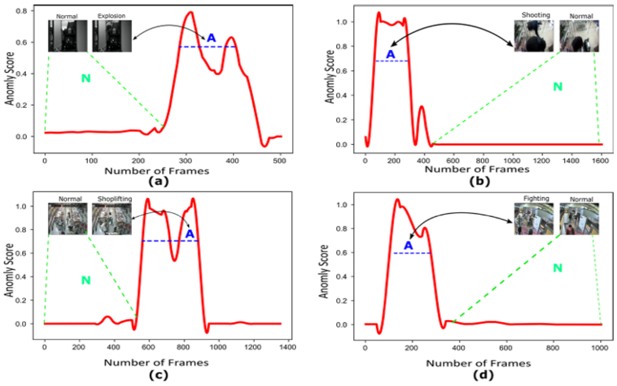


Figure 1: Visualization of testing results on UCF-Crime

5. CONCLUSION

This paper presents a **deep learning-based approach for anomaly detection (AD) in video surveillance**, leveraging a **Fully Convolutional Neural Network (FCNN)** in conjunction with **Multiple Instance Learning (MIL)** to improve anomaly identification. The proposed method processes surveillance videos at **240 × 320 resolution** and **30 frames per second**, dividing them into **32 non-overlapping snippets**. The **I3D-ResNet50 model** extracts spatiotemporal features, which are then passed through the **FCNN to compute anomaly scores**. An innovative **ranking function** enhances the learning process under **weakly supervised conditions**, leading to improved performance in real-world anomaly detection scenarios.

Extensive evaluations conducted on the **UCF-Crime dataset** demonstrate a **7.44% improvement** over the second-best approach in the dataset, highlighting the **effectiveness and robustness** of the proposed model. The results confirm that this methodology significantly enhances the accuracy of anomaly detection across diverse environments, including **varying lighting conditions, perspectives, and complex scene distributions**.

Future Scope :

While the proposed model shows **notable improvements**, several enhancements can be explored to further **refine performance and generalizability**:

1. **Meta-Heuristic Optimization:** Future work can integrate **meta-heuristic algorithms**, such as **Genetic Algorithms (GA), Particle Swarm Optimization (PSO), or Ant Colony Optimization (ACO)**, to **optimize the ranking function and fine-tune hyperparameters**, leading to improved model efficiency.

2. **Multi-Modal Learning:** Incorporating **audio-visual fusion techniques** can enhance anomaly detection by leveraging **acoustic cues** alongside visual information, improving the system's ability to detect suspicious activities such as gunshots, screams, or breaking glass.

3. **Adaptive Real-Time Processing:** Deploying the model in **real-time surveillance systems** with an **adaptive learning mechanism** would allow it to evolve dynamically based on **new data and environmental changes**, reducing false positives and improving detection rates.

4. **Cross-Dataset Generalization:** Expanding evaluations across **multiple benchmark datasets**, such as **XD-Violence, ShanghaiTech, or Avenue Dataset**, would ensure broader applicability and robustness in detecting diverse types of anomalies.

5. **Explainability and Interpretability:** Implementing **explainable AI (XAI) techniques**, such as **Grad-CAM or SHAP**, can provide better insights into **which features contribute most to anomaly classification**, increasing trust and usability for law enforcement and security agencies.

6. **Edge and Cloud-Based Deployment:** Optimizing the framework for **edge computing** (IoT-based surveillance cameras) and **cloud-based platforms** would facilitate scalable, real-time anomaly detection in **smart cities, airports, and industrial security systems**.

6. ACKNOWLEDGMENTS

Authors thank management and Department of NIRT, College of Engineering, Bhopal India for support and assistance to carry out this work.

7. REFERENCES

[1] M. T. Subha Devi, M. Dhanalakshmi, S. A, S. ML and L. N, "Anomaly Detection in Video Surveillance," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-3, doi: 10.1109/I2CT61223.2024.10543949.

[2] Vajda, D.L., Do, T.V., Bérczes, T. et al. Machine learning-based real-time anomaly detection using data pre-processing in the telemetry of server farms. Sci Rep 14, 23288 (2024). https://doi.org/10.1038/s41598-024-72982-z.

[3] Proceedings of the IEEE International Conference on Computer Vision, 2017, v. 2017-October, page number. 341-349.DOI: http://dx.doi.org/10.1109/ICCV.2017.45.

[4] Edmund Fosu Agyemang.Anomaly detection using unsupervised machine learning algorithms: A simulation study.Scientific African Volume 26, December 2024, ScienceDirect.

[5] Hamza Karim, Keval Doshi, Yasin Yilmaz;Real-Time Weakly Supervised Video Anomaly Detection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6848-6856.

[6] Wang, T., Liu, Z., & Liu, L. (2024). Investigating a three-dimensional convolution recognition model for acoustic emission signal analysis during uniaxial compression failure of coal. Geomatics, Natural Hazards and Risk, 15(1).https://doi.org/10.1080/19475705.2024.2322483.

[7] Ullah W, Ullah A, Haq IU, Muhammad K, Sajjad M, Baik SW (2021) Cnn features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimedia Tools and Applications 80(11):16979–16995.

[8] Zhong Y, Chen X, Jiang J, Ren F. A cascade reconstruction model with generalization ability evaluation for anomaly detection in videos. Pattern Recogn 2022. 122:108336.

[9] Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, pp. 4975–4986.

[10] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE International Conference on Computer Vision 2013,pp. 2720–2727.

[11] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479–6488.

[12] Zisserman A, Carreira J, Simonyan K, Kay W, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, et al (2017) The kinetics human action video dataset.

[13] 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.

[14] Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. In: CVPR 2011, pp. 3313–3320. IEEE.

[15] Dubey S, Boragule A, Jeon M (2019) 3d resnet with ranking loss function for abnormal activity detection in videos. In: 2019 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 1–6. IEEE.