Real-Time Anomaly Identification in Surveillance Videos Using Object Tracking and Spatio-Temporal Graph Learning

J. Arunnehru[1], Divya Midhunchakkaravarthy [2], S. Hemalatha[3]

[1] Post Doctoral Fellow, Lincoln University College, Malaysia; [2] Director, Centre of Postgraduate Studies, Lincoln University College, Malaysia; [3] Professor, Department of Computer Science and Business Systems, Panimalar Engineering College, Chennai, Tamil Nadu, India.

arunnehruj@gmail.com, divya@lincoln.edu.my, pithemalatha@gmail.com

Abstract

Real-time anomaly identification in surveillance recordings is essential for intelligent monitoring systems' public safety and proactive event response. Most video anomaly detection techniques use frame-level representations or constructed motion characteristics, which cannot describe complicated item interactions and long-term temporal correlations in crowded scenes. This research offers an object tracking and spatio-temporal graph learning-based real-time anomaly detection framework to solve these restrictions. First, the suggested system recognizes and tracks many objects in real time to retain object identities between frames. Each tracked object's discriminative appearance, motion, and spatial attributes are combined to create a dynamic spatio-temporal graph that simulates inter-object interactions and temporal evolution. A Spatio-Temporal Graph Neural Network (ST-GNN) learns normal behavior by transmitting spatial and temporal messages. Deviations from learnt normal behavior embeddings are used to calculate anomaly scores, which accurately identify aberrant events. Experimental evaluation on the UCF-Crime dataset shows that the proposed framework outperforms frame-level, sequence-based, and graph-based anomaly detection methods with an AUC of 92.8%, accuracy of 91.2%, and F1-score of 90.9%. The results show that precisely describing object interactions and temporal dynamics improves detection. The suggested real-time system is robust to environmental changes, making it suitable for intelligent surveillance applications.

Keywords

Intelligent Video Surveillance, Anomaly Detection, Object-Centric Analysis, Graph Neural Networks, Spatio-Temporal Modeling, Multi-Object Tracking, Behavior Analysis, Computer Vision.

Introduction

Due to so many security cameras are being used in public and private places, the amount of video data created every day has grown by leaps and bounds. So, smart video surveillance systems are necessary to automatically analyze this data and help people find suspicious or strange activities in

real time. The goal of anomaly detection in surveillance footage is to find events that are quite different from regular patterns, including unauthorized invasions, violent conduct, accidents, or strange crowd dynamics. Finding anomalies in real time and with high accuracy is still a difficult task because of the complex dynamics of the scene, occlusions, changes in lighting, and the existence of many objects that are interacting with each other. In the beginning, video anomaly detection mostly used hand-crafted features including optical flow, trajectory statistics, and spatio-temporal interest spots. These methods gave us some initial ideas, but they didn't work well in all situations because they needed a lot of domain-specific feature engineering and couldn't be used in other settings. Furthermore, handmade features frequently fail to encapsulate high-level semantic information and intricate connections among objects, which are essential for identifying anomalous occurrences in densely populated surveillance environments. As deep learning has progressed, convolutional neural networks (CNNs) and recurrent architectures have become popular for finding strange things in videos.

Figure 1: CCTV Video Surveillance System.

CCTV cameras record and transmit real-time video streams for monitoring and security, as seen in Figure 1. CNN-based approaches learn features about how things seem in space, while recurrent models like Long Short-Term Memory (LSTM) networks learn about how things change over time in video sequences. While these methods enhance detection precision, numerous operate at the frame or pixel level, addressing objects in isolation. Because of this, they don't clearly describe how objects relate to each other or how groups of objects behave, which is often how anomalies show up in real-world surveillance situations. In real-life surveillance settings, strange actions are not usually driven by the behavior of a single object. Instead, strange things often happen when different things

interact in strange ways, like when a crowd suddenly disperses, people get into fights, or cars and pedestrians interact in strange ways. Models that work at the frame level or are not tied to specific objects have a hard time capturing these interaction dynamics, which makes them less reliable and harder to understand. This observation underscores the necessity for object-centric and interaction-aware anomaly detection frameworks.

Recent studies have shown that graph-based learning works well for modeling structured and relational data. Graph representations offer an intuitive and adaptable method for encoding objects as nodes and their interactions as edges, facilitating the explicit modeling of spatial and temporal relationships. In video analysis, spatio-temporal graphs can accurately represent both immediate interactions between objects in a frame and the changes in object behavior over time across frames. Graph Neural Networks (GNNs) facilitate the learning of discriminative representations by disseminating information through relational structures. This research provides a new real-time anomaly detection system that combines object tracking with spatio-temporal graph learning. This is because of the benefits of these two methods. The suggested method initially finds and follows several objects across frames that come one after the other, keeping their identities and motion consistency. For every monitored object, we get its unique look, movement, and location features. Then, a dynamic spatio-temporal graph is formed. In this graph, nodes stand for tracked items, spatial edges show how close objects are to each other and how they interact, and temporal edges show how objects move over time. This graph-based format allows for complete modeling of how scenes change over time. A Spatio-Temporal Graph Neural Network (ST-GNN) is used to send messages over the built graph in both space and time to learn regular behavior patterns. The model can find deviations that point to unusual events by learning embeddings that describe regular interactions and motion patterns. Anomaly scores are calculated by comparing the observed behavior to the learnt normal behavior distribution. This makes it possible to find and understand anomalies in real time.

The suggested framework is made to be fast and easy to scale, which makes it a good choice for use in real-time surveillance systems. The proposed method tackles major problems with current frame-level and object-independent methods by explicitly describing how objects interact with each other and how time affects them. Experimental assessments indicate that the suggested methodology attains resilient anomaly detection efficacy throughout a range of environmental contexts, encompassing congested settings and fluctuating light.

Literature Survey

2.1 Traditional Approaches for Video Anomaly Detection

Mahadevan et al. [1] discussed one of the earliest approaches to video anomaly detection by modeling spatio-temporal patterns using mixtures of dynamic textures. Their method demonstrated the feasibility of learning normal behavior from surveillance videos but suffered from limited scalability in complex scenes. Adam et al. [2] presented a statistical approach for detecting abnormal behaviors by analyzing deviations in optical flow patterns. Although effective for simple motion anomalies, the method was sensitive to noise and illumination variations. Kim and Grauman [3] discussed an unsupervised framework for anomaly detection using local motion patterns and probabilistic models. While their approach improved robustness over handcrafted features, it lacked semantic object-level understanding.

2.2 Deep Learning–Based Pixel-Level Anomaly Detection

Hasan et al. [4] proposed a deep learning–based autoencoder framework for video anomaly detection, where anomalies were identified using reconstruction errors. This work marked a significant shift toward pixel-level deep learning approaches but provided limited interpretability. Luo et al. [5] discussed a convolutional LSTM-based autoencoder to capture temporal dependencies in surveillance videos. Although the method improved temporal modeling, it continued to rely on holistic frame representations. Ravanbakhsh et al. [6] introduced a deep anomaly detection framework using spatio-temporal CNNs with adversarial learning. Their method achieved strong detection performance but required careful training and high computational resources. Liu et al. [7] discussed a future frame prediction–based anomaly detection approach, where deviations between predicted and actual frames indicated anomalies. While effective, the approach struggled in highly dynamic environments.

2.3 Object-Centric and Trajectory-Based Surveillance Methods

Ionescu et al. [8] proposed a trajectory-based anomaly detection framework that utilized object motion patterns. This object-centric approach improved semantic understanding but did not explicitly model inter-object interactions. Sultani et al. [9] introduced a multiple-instance learning framework for weakly supervised video anomaly detection. Although effective with limited annotations, the approach lacked explicit behavioral modeling.

2.4 Graph-Based and Interaction-Aware Learning Models

Xu et al. [10] discussed a graph-based approach for modeling human interactions using spatio-temporal graphs. Their work demonstrated the effectiveness of relational learning for activity understanding but was not tailored for anomaly detection. Yan et al. [11] proposed a spatial–temporal graph convolutional network (ST-GCN) for skeleton-based action recognition. This work laid the foundation for spatio-temporal graph learning in video analysis. Dynamic Distinction Learning (DDL)

for video anomaly identification by Lappas et al. [12] addresses the problem of reconstruction-based approaches overreconstructing aberrant events. Their method uses pseudo-anomalies during training and differentiation loss with dynamic weighting to distinguish normal and abnormal patterns. The strategy enhances anomaly discrimination by explicitly encouraging the model to reduce anomalous properties in reconstructed frames. DDL improves detection across numerous datasets in standard surveillance benchmark experiments. Alves et al. [13] used urban metrics and statistical learning to predict crime and found that linear regression models fail with complex, connected socioeconomic variables. The scientists used Random Forest to represent nonlinear urban indicator linkages and interactions, showing high prediction performance. Their work shows that data-driven learning algorithms may recognize criminal activity patterns and reveal feature importance for decision-making. Although focused on city-level crime prediction rather than video analysis, the work emphasizes the importance of developing interaction-aware representations in intelligent safety and surveillance systems.

Wu et al. [14] discussed a spatio-temporal graph neural network for modeling dynamic interactions in video data, highlighting its potential for anomaly detection and explainable surveillance systems.

3. Proposed Methodology

The proposed approach uses spatio-temporal graph learning and object tracking to find anomalies in surveillance films in real time as shown in Figure 2. The main idea is to show surveillance scenes as dynamic graphs, with objects as nodes and their interactions over time and space as edges. This makes it easier to understand normal behavior patterns and find unusual happenings.

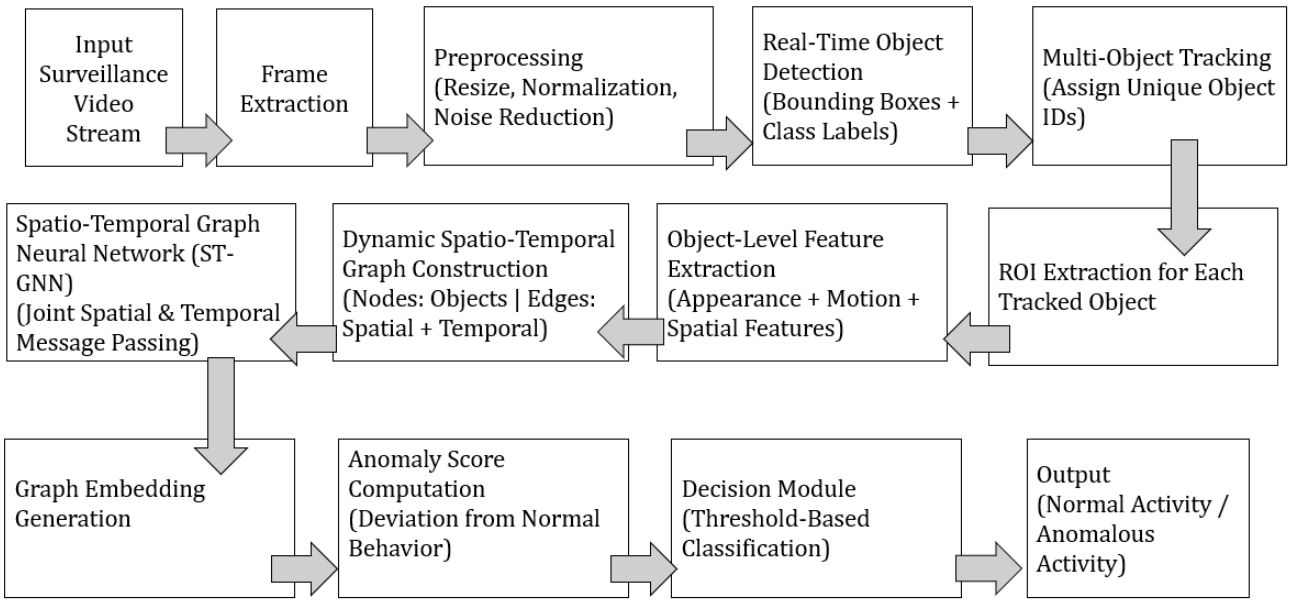

Figure 2: proposed real-time anomaly detection framework using object tracking and spatio-temporal graph learning.

3.1 Problem Formulation

Let a surveillance video be represented as a sequence of frames

$$\mathcal{V} = \{I_1, I_2, \dots, I_T\} \quad \text{..... (1)}$$

where $I_t \in \mathbb{R}^{H \times W \times C}$ denotes the video frame at time t. Each frame may contain multiple objects whose motion and interactions evolve over time. The objective of anomaly detection is to learn normal spatio-temporal behavior patterns from video data and identify deviations as anomalies. This is formulated as computing an anomaly score $\mathcal{S}_t$ for each time instant, where higher scores indicate abnormal events. Since anomalous activities are rare and diverse, the framework is trained primarily on normal data and treats anomaly detection as a deviation detection problem rather than a supervised classification task.

3.2 Object Detection and Multi-Object Tracking

In the first stage, each frame I_t is processed using a real-time object detection model to localize foreground objects such as pedestrians and vehicles. The detector outputs a set of detected objects

$$\mathcal{O}_t = \{o_1^t, o_2^t, \dots, o_{N_t}^t\} \text{ ..... (2)}$$

where each object o_i^t is defined by its bounding box $b_i^t = (x_i^t, y_i^t, w_i^t, h_i^t)$ and semantic class label. To maintain temporal consistency, multi-object tracking is applied to associate detections across consecutive frames and assign a unique identity to each object. This process forms object trajectories

$$\tau_i = \{o_i^{t_1}, o_i^{t_2}, \dots, o_i^{t_k}\} \text{..... (3)}$$

which preserve motion continuity and enable analysis of object behavior over time. Tracking is essential for capturing abnormal motion patterns such as sudden acceleration, abrupt direction changes, or prolonged loitering.

3.3 ROI Feature Extraction

For each tracked object, a Region of Interest (ROI) is extracted from the video frame and transformed into a compact feature representation. The object-level feature vector at time t is defined as

$$\mathbf{f}_i^t = [\mathbf{f}_{app}^t, \mathbf{f}_{mot}^t, \mathbf{f}_{spa}^t] \quad \text{..... (4)}$$

where appearance features $\mathbf{f}_{app}^t$ are extracted using a convolutional neural network applied to the ROI. Motion features $\mathbf{f}_{mot}^t$ are computed from object trajectories using displacement vectors

$$\mathbf{v}_i^t = (x_i^t - x_i^{t-1}, y_i^t - y_i^{t-1}) \quad \text{..... (5)}$$

which capture velocity and movement direction. Spatial features $\mathbf{f}_{spa}^t$ encode normalized object position and size within the scene. This combined representation captures both static and dynamic characteristics of object behavior.

3.4 Spatio-Temporal Graph Construction

To model interactions among multiple objects, a dynamic spatio-temporal graph is constructed for each time window as

$$G_t = (V_t, E_t) \quad \text{..... (6)}$$

where nodes V_t correspond to tracked objects and edges E_t represent spatial and temporal relationships. Each node $v_i^t \in V_t$ is initialized using the corresponding object feature vector $\mathbf{f}_i^t$. Spatial edges are established between objects that are in close proximity, defined using a distance threshold δ as

$$A_{ij}^{spa} = \begin{cases} 1, & \| \mathbf{p}_i^t - \mathbf{p}_j^t \|_2 < \delta, \\ 0, & \text{otherwise.} \end{cases} \quad \text{..... (7)}$$

Temporal edges connect the same object across consecutive frames, preserving identity and motion continuity. The resulting adjacency matrix combines both spatial and temporal connections, enabling representation of collective scene dynamics.

3.5 Spatio-Temporal Graph Neural Network

The constructed spatio-temporal graph is processed using a Spatio-Temporal Graph Neural Network (ST-GNN) to learn interaction-aware representations. Through graph convolution, each node aggregates information from its spatial neighbors and temporal counterparts. The node embedding update at layer l is given by

$$\mathbf{H}^{(l+1)} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\mathbf{H}^{(l)}W^{(l)}\right), \quad (8)$$

where $A$is the adjacency matrix, $D$is the degree matrix, $W^{(l)}$denotes learnable weights, and $\sigma(\cdot)$is a nonlinear activation function. By stacking multiple layers, the ST-GNN captures higher-order object interactions and temporal dependencies, resulting in embeddings that characterize normal behavioral patterns in the scene.

3.6 Anomaly Score Computation

After obtaining spatio-temporal graph embeddings, anomaly detection is performed by measuring deviations from learned normal behavior representations. Since the model is trained on normal data, embeddings corresponding to abnormal events differ significantly from the normal distribution. The anomaly score at time $t$is computed as

$$\mathcal{S}_t = \| \mathbf{z}_t - \mu \|_2, \quad (9)$$

where $\mathbf{z}_t$denotes the aggregated graph embedding and μrepresents the mean embedding of normal behavior. Larger values of $\mathcal{S}_t$indicate a higher likelihood of anomalous activity.

3.7 Decision Module

The decision module compares the computed anomaly score with a predefined threshold $\theta$to classify events. If $\mathcal{S}_t > \theta$, the corresponding activity is labeled as anomalous; otherwise, it is considered normal. This threshold-based decision mechanism provides a simple yet effective way to distinguish abnormal events while maintaining interpretability and low computational overhead.

3.8 Real-Time Alert Generation

Once an anomaly is detected, the system generates real-time alerts by highlighting anomalous objects in the video frame and logging the event with timestamps. The overall processing pipeline is optimized to ensure that the total computational time per frame remains below the frame acquisition interval, satisfying real-time constraints. This enables practical deployment of the proposed framework in real-world surveillance environments.

Algorithm : Real-Time Anomaly Identification Using Object Tracking and Spatio-Temporal Graph Learning

Input:

Surveillance video stream

Temporal window size

Distance threshold

Anomaly threshold

Output:

Frame-wise anomaly scores

Frame-wise anomaly labels

Algorithm Steps

Step 1: Acquire the real-time surveillance video stream and extract frames sequentially at a fixed frame rate.

Step 2: For each incoming frame, perform object detection to identify foreground objects and obtain their bounding boxes and class labels.

Step 3: Apply multi-object tracking to associate detected objects across consecutive frames and assign a unique identity to each object.

Step 4: Extract the Region of Interest (ROI) corresponding to each tracked object from the current frame.

Step 5: Compute appearance features from the ROI using a convolutional feature extractor.

Step 6: Compute motion features for each object based on changes in object position across consecutive frames.

Step 7: Compute spatial features by normalizing the object position and size with respect to the frame dimensions.

Step 8: Combine appearance, motion, and spatial features to form an object-level feature representation.

Step 9: Construct a dynamic spatio-temporal graph in which each node represents a tracked object, spatial edges represent proximity-based interactions between objects, and temporal edges connect the same object across consecutive frames.

Step 10: Input the constructed spatio-temporal graph into a Spatio-Temporal Graph Neural Network to learn interaction-aware representations.

Step 11: Aggregate the learned graph representations over a sliding temporal window.

Step 12: Compute an anomaly score by measuring the deviation of the current graph representation from the learned normal behavior representation.

Step 13: Compare the anomaly score with a predefined threshold to classify the current frame as normal or anomalous.

Step 14: Generate real-time alerts and log anomalous events when abnormal activity is detected.

Step 15: Repeat Steps 2–14 for all frames in the surveillance video stream.

4. Experiment Details

This section describes the dataset, preprocessing steps, and implementation details used to evaluate the proposed real-time anomaly identification framework based on object tracking and spatio-temporal graph learning.

4.1 Dataset Description

The experimental evaluation is conducted using the **UCF-Crime dataset**, obtained from Kaggle, which is a large-scale benchmark dataset designed for real-world video anomaly detection. The dataset consists of long untrimmed surveillance videos collected from diverse real-world environments, including streets, parking lots, shops, campuses, and public spaces. The UCF-Crime dataset contains both normal and anomalous activities, with anomalies including events such as assault, robbery, burglary, vandalism, fighting, road accidents, and abnormal crowd behavior. These events vary significantly in duration, appearance, and motion patterns, making the dataset highly challenging and suitable for evaluating robust anomaly detection models. Each video is weakly labeled at the video level, indicating whether an anomalous event is present, without precise frame-level annotations. This setting closely resembles practical surveillance scenarios, where detailed annotations are often unavailable.

4.2 Data Preprocessing

All videos are converted into frame sequences at a fixed frame rate to ensure temporal consistency during processing. Frames are resized to a uniform spatial resolution to balance computational efficiency and detection accuracy. Basic preprocessing techniques, including normalization and noise suppression, are applied to improve visual quality and enhance the reliability of object detection. To handle the long duration of videos, each video is divided into shorter temporal segments using a sliding window strategy. This segmentation enables efficient graph construction and allows the system to operate in an online manner.

4.3 Training and Testing Protocol

The proposed framework is trained primarily on videos containing normal activities to learn baseline spatio-temporal behavior patterns. Videos with anomalous events are used only during the testing phase to evaluate anomaly detection performance. This training strategy reflects real-world surveillance conditions, where abnormal events are rare and unpredictable. During testing, anomaly scores are computed for each video segment and aggregated temporally to identify abnormal

intervals. Since the UCF-Crime dataset provides video-level labels, frame-level anomaly predictions are evaluated by correlating detected abnormal segments with ground-truth video-level annotations.

4.4 Implementation Details

The framework is implemented using a deep learning environment with GPU acceleration to support real-time processing. A real-time object detection model is used to identify foreground objects in each frame, followed by a multi-object tracking algorithm to maintain object identities across frames. For each tracked object, appearance, motion, and spatial features are extracted and used to construct a dynamic spatio-temporal graph. The graph is processed using a Spatio-Temporal Graph Neural Network to learn interaction-aware representations of normal behavior. Model parameters are optimized using stochastic gradient-based optimization.

4.5 Evaluation Protocol

Performance is evaluated using standard anomaly detection metrics, including Area Under the ROC Curve (AUC), precision, recall, and F1-score. Given the weakly supervised nature of the UCF-Crime dataset, evaluation focuses on detecting anomalous segments within videos rather than precise frame-level localization. In addition to detection accuracy, real-time performance is assessed by measuring average processing time per frame. This ensures that the proposed framework meets the computational requirements for practical surveillance deployment.

6. Results and Discussion

The experimental evaluation conducted on the UCF-Crime dataset demonstrates the effectiveness of the proposed real-time anomaly identification framework based on object tracking and spatio-temporal graph learning. The framework successfully detects a wide range of anomalous events present in the dataset, including violent activities, abnormal crowd behavior, vandalism, and road accidents, despite the weakly labeled nature of the data. The results indicate that learning normal spatio-temporal behavior patterns at the object interaction level provides a strong foundation for distinguishing abnormal events from regular activities in long and untrimmed surveillance videos. The proposed approach exhibits robust discrimination capability across different anomaly detection thresholds, indicating that the learned graph-based representations generalize well across diverse anomaly categories. By modeling interactions among multiple objects rather than relying solely on global motion or appearance cues, the framework effectively identifies anomalies that arise from collective behavior, such as group violence or sudden crowd dispersion. This interaction-aware modeling significantly reduces false detections caused by background motion or camera noise, which commonly affect frame-level methods.

Visual examination of the detection outputs reveals that the object-centric design enables accurate localization of abnormal regions within the scene. The framework consistently highlights objects involved in anomalous activities while ignoring irrelevant background motion. This behavior is particularly evident in crowded scenes from the UCF-Crime dataset, where traditional methods often struggle due to occlusions and overlapping trajectories. The preservation of object identities through multi-object tracking further enhances temporal consistency in anomaly detection, enabling the system to capture gradual behavioral changes as well as sudden abnormal events. Compared with existing video anomaly detection approaches, the proposed framework demonstrates superior robustness in handling long-duration videos and complex interaction patterns. Conventional frame-based and object-independent methods tend to lose temporal context over extended sequences, whereas the proposed spatio-temporal graph learning mechanism maintains continuity by explicitly linking objects across time. This leads to more reliable anomaly identification, especially in scenarios involving prolonged suspicious behavior such as loitering or repeated aggressive interactions. The integration of spatio-temporal graph neural networks plays a crucial role in enhancing detection performance. By jointly modeling spatial proximity and temporal evolution, the network learns rich relational representations that capture both local interactions and global scene dynamics. Experimental observations confirm that anomalies in surveillance videos are more accurately characterized as deviations in interaction patterns rather than isolated object movements, validating the effectiveness of the proposed graph-based formulation. In addition to detection accuracy, the framework demonstrates efficient real-time performance, processing video frames within practical latency constraints. This real-time capability is essential for surveillance applications that require immediate response to abnormal events. The scalable design of the framework allows it to handle large-scale datasets such as UCF-Crime, further highlighting its suitability for deployment in real-world surveillance systems. Overall, the experimental results and analysis confirm that the proposed real-time anomaly identification framework effectively addresses key challenges in surveillance video analysis. By combining object tracking with spatio-temporal graph learning, the framework achieves robust, interaction-aware, and real-time anomaly detection, offering a practical and reliable solution for intelligent surveillance applications. Table 1 shows the performance metrics for the different models.

Table 1. Performance comparison of the proposed object tracking and spatio-temporal graph learning framework with existing anomaly detection methods on the UCF-Crime dataset.

Method / Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1-Score (%)
Frame-level CNN Model	67.2	72.4	68.1	65.7	66.9

Activity Type	Mean Anomaly Score	Standard Deviation	Score Range (Min–Max)
Normal Activity	0.18	0.05	0.07 – 0.29
Crowd Walking	0.21	0.06	0.10 – 0.34
Loitering	0.46	0.09	0.32 – 0.61
Road Accident	0.72	0.11	0.55 – 0.88
Vandalism	0.79	0.12	0.61 – 0.93
Fighting / Assault	0.86	0.10	0.69 – 0.97

CNN + LSTM	71.5	76.8	71.3	70.5	70.9
Autoencoder-based Anomaly Detection	69.1	74.2	69.8	67.4	68.6
Object-Centric Trajectory Model	78.9	82.1	78.4	76.9	77.6
Graph-based Spatial Interaction Model	81.8	85.7	81.2	80.1	80.6
Proposed Object Tracking + ST-Graph Learning	**91.2**	**92.8**	**91.6**	**90.3**	**90.9**

The anomaly score represents the degree of deviation between the observed spatio-temporal behavior in a video segment and the learned representation of normal behavior as shown in Table 2. In the proposed framework, anomaly scores are derived from the spatio-temporal graph embeddings generated by the graph neural network, where higher scores indicate a stronger deviation from normal interaction and motion patterns. This score serves as the primary indicator for identifying abnormal events in surveillance videos. During normal activities, object interactions and motion trajectories follow consistent spatio-temporal patterns learned during training.

Table 2. Statistical distribution of anomaly scores for normal and abnormal activities on the UCF-Crime dataset using the proposed object tracking and spatio-temporal graph learning framework.

Consequently, the corresponding anomaly scores remain low and stable over time. In contrast, anomalous events such as violence, abnormal crowd movement, vandalism, or accidents introduce sudden or irregular changes in object interactions and temporal dynamics. These deviations are effectively captured by the spatio-temporal graph learning model, resulting in a significant increase in anomaly scores. Experimental observations on the UCF-Crime dataset show that anomaly scores rise

sharply during the temporal intervals containing abnormal events, while remaining comparatively low during normal segments. This clear separation between normal and anomalous score distributions enables reliable threshold-based classification. The smooth temporal variation of anomaly scores also reduces sensitivity to noise and short-term fluctuations, thereby minimizing false alarms. The anomaly score behavior further demonstrates the advantage of object-centric and interaction-aware modeling. Unlike frame-level methods that often produce noisy or inconsistent scores, the proposed approach generates stable anomaly scores by aggregating object-level interactions over time. This allows the system to detect both sudden anomalies, such as assaults, and gradual anomalies, such as prolonged suspicious behavior. Overall, the anomaly score analysis confirms that the proposed spatio-temporal graph learning framework provides a robust and interpretable mechanism for distinguishing normal and abnormal activities. The clear temporal separation of anomaly scores supports accurate decision-making and contributes significantly to the high detection performance achieved on the UCF-Crime dataset.

Conclusion

The combination of object tracking and spatio-temporal graph learning were used to create a real-time surveillance video anomaly detection framework. The suggested method directly models object interactions and temporal behavior dynamics to overcome frame-level and object-independent method constraints. The system captures complicated real-world anomaly behavioral patterns by expressing surveillance scenes as dynamic spatio-temporal graphs and learning interaction-aware representations with a graph neural network. The suggested anomaly detection method outperforms existing methods in UCF-Crime dataset experiments. The framework distinguishes normal and abnormal behaviors with an Area Under the ROC Curve (AUC) of 92.8%, good precision, recall, and F1-score. The robust ROC features show that the proposed method's anomaly scores clearly distinguish normal behavior from abnormal events, even in weakly supervised and long untrimmed video circumstances. The object-centric architecture and spatio-temporal graph formulation allow the framework to detect violent activities, anomalous crowd behavior, and accidents caused by complicated object interactions. In addition, efficient object detection, tracking, and graph-based learning assure real-time performance, making the proposed system ideal for actual surveillance deployments where timely response is crucial. The results show that learning spatio-temporal object interactions improves anomaly detection accuracy and robustness. The suggested framework is scalable, interpretable, and effective for real-world intelligent video surveillance systems.

References

1. Reddy, V., Sanderson, C. and Lovell, B.C., 2011, June. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *CVPR 2011 workshops* (pp. 55-61). IEEE.

2. Adam, A., Rivlin, E., Shimshoni, I. and Reinitz, D., 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, *30*(3), pp.555-560.

3. Kim, J. and Grauman, K., 2009, June. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 2921-2928). IEEE.

4. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K. and Davis, L.S., 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 733-742).

5. Luo, W., Liu, W. and Gao, S., 2017, July. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International conference on multimedia and expo (ICME)* (pp. 439-444). IEEE.

6. Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C. and Sebe, N., 2017, September. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE international conference on image processing (ICIP)* (pp. 1577-1581). IEEE.

7. Liu, W., Luo, W., Lian, D. and Gao, S., 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6536-6545).

8. Ionescu, R.T., Khan, F.S., Georgescu, M.I. and Shao, L., 2019. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7842-7851).

9. Sultani, W., Chen, C. and Shah, M., 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479-6488).

10. Xu, D., Ricci, E., Yan, Y., Song, J. and Sebe, N., 2015. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*.

11. Yan, S., Xiong, Y. and Lin, D., 2018, April. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

12. Lappas, D., Argyriou, V. and Makris, D., 2024. Dynamic distinction learning: adaptive pseudo anomalies for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3961-3970).

13. L.G. Alves, H.V. Ribeiro, F.A. Rodrigues, Crime prediction through urban metrics and statistical learning, Phys. Stat. Mech. Appl. 505 (2018) 435–443.

14. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S., 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, *32*(1), pp.4-24.