International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

Smart Campus Surveillance Using Drone Based Video Streaming And Alert System

G. Blessy Hunly, G. Bala Siva Ram Kumar, I. Lakshmi Pujitha, G. Chanikya Ram, P. Venu Kumari, and Jagadeesh Thati
Department of ECE, Tirumala Engineering College, Narasaraopet, AP-522601

***Abstract*—This paper presents an autonomous drone based violence detection system that integrates aerial video acquisition, edge-based machine learning, and real-time video streaming for continuous campus monitoring.The growing need for proactive security monitoring in educational campuses motivates the use of mobile and intelligent surveillance platforms. A Raspberry Pi enabled drone equipped with a camera module patrols predefined areas while transmitting live video to a ground station. The onboard processing unit analyses visual data in real time using a lightweight activity recognition model trained to identify violent and aggressive human interactions. When suspicious activity is detected, the system autonomously triggers image capture and stores visual evidence in onboard memory for later inspection, while maintaining uninterrupted live video streaming. This work is proposing a system architecture that is optimized to operate within the computational and energy constraints of small aerial platforms, enabling real time deployment scenarios. Experimental evaluation under dynamic outdoor conditions is expected to demonstrate reliable detection accuracy, low inference latency, and stable real time performance, there by validating the feasibility of drone assisted intelligent surveillance for campus violence monitoring.**

***Index Terms*—Drone Based Surveillance, Violence Detection System, CNN-LSTM Architecture, Optical Flow Analysis, Real-Time Video Streaming, Edge AI, Raspberry Pi, Human Activity Recognition, Smart Campus Monitoring, Computer Vision, Anomaly Detection, IoT-Based Security.**

I. Introduction

The rapid growth of population density have significantly amplified the need for intelligent surveillance systems capable of ensuring public safety. Traditional surveillance methods rely heavily on human monitoring, which is often inefficient,error-prone. In this context, automated violence detection has emerged as a critical research domain within computer vision, aiming to identify abnormal or aggressive human behaviors from video streams. Early approaches focused on handcrafted features and motion descriptors extracted from videos, where spatio-temporal interest points and trajectory-based representations were used to model human actions in realistic environments [1]. These techniques laid the foundation for understanding dynamic activities.

With the advancement of deep learning, convolutional neural networks (CNNs) have become fundamental in extracting high level spatial features from video frames. These models automatically learn discriminative representations,significantly improving performance over traditional methods. To further enhance motion understanding, temporal information is incorporated using sequence modeling techniques [5]. This combination is especially effective for violence detection, where identifying subtle motion patterns across consecutive frames is crucial.

In addition to this, 3D convolutional neural networks have been introduced to directly learn spatio temporal features by processing multiple frames simultaneously [4]. But they often require substantial computational resources,making them less suitable for real-time and resource-constrained applications. As a result, optimized and efficient architectures have been explored to balance accuracy and computational cost, enabling deployment in practical surveillance systems such as drones [8].

Another important direction in surveillance research is anomaly detection,where models learn patterns of normal behavior and identify deviations as potential threats. Memory-guided learning and advanced representation techniques have improved the detection of events like violence [7].

In this project, a smart surveillance system is proposed that integrates dronebased video acquisition with a CNN LSTM based deep learning model for real-time violence detection. The use of drones provides enhanced coverage and flexibility compared to fixed surveillance systems, allowing monitoring of dynamic and largescale environments. The captured video stream is processed through frame extraction and preprocessing stages, followed by feature extraction using CNNs. Temporal dependencies across frames are then modeled using LSTM networks, enabling accurate recognition of violent activities. The system employs an optimized learning flow to improve performance, and a sigmoid activation function is used for binary classification of violence and non-violence

Overall, the integration of deep learning techniques with aerial surveillance platforms offers a scalable and efficient solution for automated violence detection. By leveraging spatial feature extraction, temporal sequence modeling, and efficient architectures, the proposed system aims to deliver accurate and real-time monitoring capabilities. This approach has the potential to enhance public safety by enabling early detection of violent incidents and facilitating timely response actions.

II. Literarure Survey

Violence detection in intelligent surveillance systems has evolved significantly with the advancement of spatio-temporal feature learning techniques. Early approaches focused on extracting handcrafted features such as spatio-temporal interest points and motion descriptors to represent human activities in

videos [1]. These methods relied on detecting local motion patterns and encoding them into feature vectors for classification. In crowded environments, statistical motion pattern modeling was introduced to learn normal behavior distributions, where deviations from these learned patterns were treated as anomalies [2]. While effective in controlled scenarios, these approaches lacked robustness in complex environments due to their dependence on manually designed features and limited generalization capability.

The introduction of deep learning techniques, particularly convolutional neural networks (CNNs), significantly improved the performance of video-based violence detection. CNNs enable automatic extraction of hierarchical spatial features directly from raw frames. To incorporate motion information, optical flow estimation models based on deep networks were developed, allowing the system to capture pixel-level motion between consecutive frames [3]. These motion representations are crucial for distinguishing violent activities, which typically involve abrupt and irregular movements. Furthermore, 3D convolutional neural networks (3D CNNs) extend conventional 2D convolutions by operating across both spatial and temporal dimensions, enabling direct learning of spatiotemporal features from video clips [4]. This eliminates the need for separate motion extraction steps and improves the model's ability to recognize dynamic activities. Additional deep learning frameworks have further demonstrated the effectiveness of CNNbased architectures in capturing both spatial and temporal cues for violence recognition tasks [5].

To effectively capture long-term temporal dependencies, recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been integrated with CNN architectures. In such hybrid models, CNN layers extract spatial features from individual frames, which are then passed to LSTM layers to model temporal sequences. This combination allows the system to understand the progression of actions over time, which is critical for distinguishing between normal and violent behaviors. Advanced architectures further enhance this capability by using stacked LSTM layers or multi-branch frameworks, where different streams process complementary information such as pose, motion, or appearance [6]. Skeleton-based representations have also been utilized, where human joint coordinates are extracted and fed into LSTM networks to model motion patterns in a structured and computationally efficient manner.

Recent developments in anomaly detection have introduced memory-augmented neural networks and representation learning techniques. These models focus on learning a compact representation of normal behavior using memory modules or latent embeddings. During inference, deviations from these learned representations indicate abnormal or potentially violent activities [7]. Such approaches are advantageous in real-world surveillance scenarios where labeled violent data is scarce, as they rely primarily on normal behavior modeling. Additionally, discrete and interpreable representations have been proposed to improve the explainability of anomaly detection systems, enabling better understanding and validation of model decisions in critical applications [12].

Efficiency and scalability are key considerations for deploying violence detection systems in real-time environments such as drone-based surveillance. Lightweight video recognition architectures have been developed by systematically scaling network dimensions, including depth, width, and temporal resolution [8]. These models achieve a balance between computational cost and accuracy, making them suitable for edge devices with limited processing power. Furthermore, few-shot learning techniques have been explored to address the challenge of limited labeled data. By leveraging temporal alignment and feature similarity, these models can generalize to new action classes with minimal training samples, enhancing adaptability in dynamic surveillance scenarios [9].

III. Objectives

Design and deploy an intelligent drone platform equipped with high-resolution cameras and sensors that can operate autonomously within campus premises. The drone should be capable of navigating predefined routes or dynamically adjusting its path based on security needs. This system aims to reduce manual intervention, increase coverage, and enable continuous surveillance in real-time. It should also incorporate obstacle avoidance and stable flight capabilities for safe operation in complex environments.

Establish a robust communication link between the drone and ground stations to facilitate uninterrupted live video streaming. The system should support high quality video transmission with minimal latency to enable immediate analysis. Onboard or edge-based processing techniques will be integrated to analyze the video streams in real time, reducing the load on central servers and ensuring quick detection of abnormal behaviors or threats.

Develop and train deep learning models, such as CNN combined with LSTM architectures, to accurately identify violent or suspicious activities from live video feeds. The model should learn to distinguish normal campus activities from alarming situations with high precision and recall. Continuous model evaluation and optimization will be performed to improve detection accuracy, minimize false positives, and ensure reliable operation in diverse scenarios

Conduct comprehensive testing using diverse datasets, including real-world campus scenarios, to assess the system's accuracy, precision, and recall in violence detection. Performance metrics such as ROC-AUC, F1 score, and confusion matrices will be analyzed to fine-tune the models and system components. The goal is to achieve high reliability with minimal false alarms and missed detections, ensuring trustworthy surveillance.

IV. Proposed Methodology

The proposed methodology outlines an autonomous drone-based surveillance system designed for real-time campus monitoring with violence detection capabilities. The system leverages onboard video capture, advanced deep learning models, and alert mechanisms to ensure prompt identification of suspicious activities. This approach enhances security by providing continuous, wide-area surveillance, reducing dependency on fixed cameras and manual monitoring. The core of

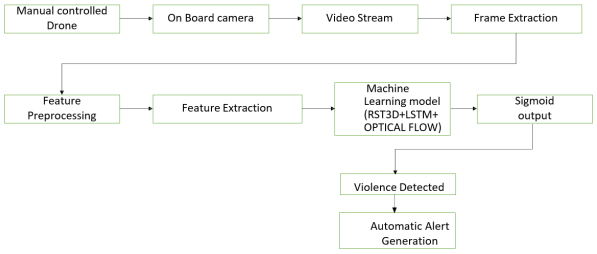

Fig. 1. proposed methodology

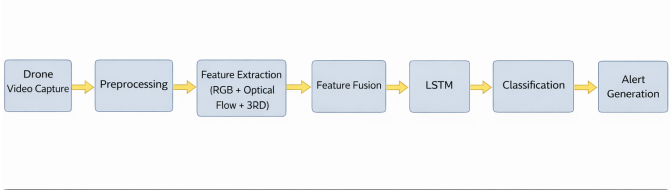

Fig. 2. working flow chart

the system The proposed system is designed to detect violent activities in surveillance videos using an integrated deep learning architecture. The system combines spatial, motion, and temporal analysis to improve the accuracy of violence detection. Video frames captured from the drone camera are first processed using a Convolutional Neural Network, which extracts important spatial features such as objects, shapes, and human interactions from each frame. Optical Flow is used to analyze motion patterns and detect sudden or abnormal

A. Working Principle

The proposed smart campus surveillance system operates by integrating drone-based video acquisition with deep learning-based violence detection techniques. The overall working principle involves real-time video capture, preprocessing, feature extraction, temporal analysis, and alert generation.

B. Video Acquisition and Streaming

A drone equipped with a camera module captures live video of the campus environment. The video stream is transmitted to an onboard processing unit or ground station for real-time analysis. This enables dynamic monitoring of large and inaccessible areas, overcoming the limitations of fixed surveillance systems.

C. Frame Extraction and Preprocessing

The captured video is divided into individual frames, which serve as inputs to the deep learning model. Each frame is resized to a fixed resolution and normalized to ensure consistent input representation and improved model performance.

D. Spatial Feature Extraction (RGB Path)

In the RGB path, individual frames are processed using a convolutional neural network (CNN) or ResNet-based architecture. This stage extracts spatial features such as human presence, posture, and scene context, which are essential for understanding the visual content of each frame.

E. Motion Feature Extraction (Optical Flow Path)

To capture motion dynamics, optical flow is computed between consecutive frames. This provides information about the direction and magnitude of movement, enabling the detection of sudden or aggressive actions such as pushing, hitting, or rapid body movements.

F. Spatio-Temporal Feature Extraction (R3D Path)

A 3D Convolutional Neural Network (ResNet3D) is employed to process sequences of frames simultaneously. This allows the model to learn both spatial and temporal features, capturing complex activity patterns and interactions over time.

G. Feature Fusion

The features obtained from the RGB path, optical flow path, and R3D path are combined through a feature fusion process. This integration provides a comprehensive representation of the scene by incorporating appearance, motion, and temporal information.

H. Temporal Modeling using LSTM

The fused features are passed to a Long Short-Term Memory (LSTM) network, which models temporal dependencies across frame sequences. This helps in understanding the progression of actions over time and improves the accuracy of violence detection.

I. Classification and Alert Generation

The output from the LSTM is fed into a fully connected layer, followed by a sigmoid activation function for binary classification. Based on the predicted probability, the system classifies the activity as either *violent* or *non-violent*. If violence is detected, an alert is generated, and relevant visual evidence is stored for further analysis.

J. System Workflow Summary

The overall workflow of the system can be summarized as follows:

K. Theoretical Analysis

The proposed system integrates spatial, motion, and temporal learning mechanisms for effective violence detection in video sequences. The theoretical foundation is based on convolutional neural networks (CNNs), optical flow estimation, 3D convolutional networks (R3D), and recurrent neural networks (LSTM).

L. Spatial Feature Extraction

Spatial features are extracted using a convolutional neural network (CNN) or ResNet-based architecture. For an input frame $I(x, y)$, convolution is defined as:

$$F(x, y) = \sum_i \sum_j I(x + i, y + j) \cdot K(i, j) \tag{1}$$

where K represents the convolution kernel. This operation captures local spatial patterns such as edges, textures, and object structures.

M. Optical Flow-Based Motion Analysis

Motion between consecutive frames is estimated using optical flow. The brightness constancy assumption is given by:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{2}$$

This leads to the optical flow constraint equation:

$$I_x u + I_y v + I_t = 0 \tag{3}$$

where (u, v) represents the motion vector. Optical flow helps in detecting abrupt movements associated with violent activities.

N. Spatio-Temporal Feature Learning using 3D CNN

Unlike 2D CNNs, 3D CNNs operate on video volumes. The 3D convolution is defined as:

$$F(t, x, y) = \sum_\tau \sum_i \sum_j V(t + \tau, x + i, y + j) \cdot K(\tau, i, j) \tag{4}$$

where V is the input video volume. This enables simultaneous learning of spatial and temporal features.

O. Temporal Modeling using LSTM

To capture long-term dependencies, LSTM networks are used. The internal operations of LSTM are defined as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$
$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{8}$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{9}$$
$$h_t = o_t \cdot \tanh(C_t) \tag{10}$$

where f_t, i_t, and o_t represent forget, input, and output gates respectively.

P. Feature Fusion

Let F_s, F_m, and F_{st} denote spatial, motion, and spatio-temporal features. The fused feature vector is given by:

$$F_{fusion} = [F_s; F_m; F_{st}] \tag{11}$$

This combined representation improves classification performance by incorporating complementary information.

Q. Classification

The final classification is performed using a sigmoid activation function:

$$P(y = 1) = \frac{1}{1 + e^{-z}} \tag{12}$$

where z is the output of the fully connected layer. The system classifies the input as violent if $P(y = 1) > 0.5$.

R. Computational Considerations

The system is optimized for real-time deployment by balancing model complexity and inference speed. While 3D CNNs provide rich feature representations, they are computationally expensive. Therefore, lightweight architectures and efficient feature fusion strategies are employed to ensure feasibility on edge devices such as Raspberry Pi.

V. Results and Discussion

A. Experimental Setup

The proposed system was evaluated using a dataset consisting of violent and non-violent video sequences. The model integrates spatial, motion, and temporal features using CNN/ResNet, Optical Flow, R3D, and LSTM architectures. The performance was validated through both offline testing and real-time webcam-based detection.

B. Confusion Matrix Analysis

The performance of the model is summarized using the confusion matrix, which consists of:

- True Positives (TP): 1140
- True Negatives (TN): 948
- False Positives (FP): 36
- False Negatives (FN): 60

The results indicate that the model correctly identifies the majority of violent and non-violent instances, with relatively low misclassification rates.

C. Performance Metrics

The following standard evaluation metrics were used:

- Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \approx 96\% \tag{13}$$

- Precision:

$$Precision = \frac{TP}{TP + FP} \approx 0.97 \tag{14}$$

- Recall:

$$Recall = \frac{TP}{TP + FN} \approx 0.95 \tag{15}$$

- F1-Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \approx 0.96 \tag{16}$$

These results demonstrate that the model maintains a strong balance between precision and recall, ensuring reliable detection of violent activities.

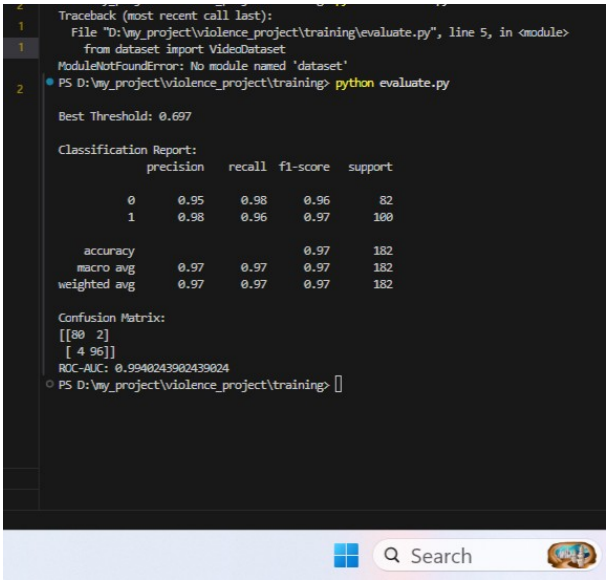

Fig. 3. Output Values

D. Discussion of Results

The high accuracy achieved by the proposed system is due to the integration of multiple feature extraction techniques:

- The CNN/ResNet model effectively captures spatial features such as human posture and scene context.
- Optical Flow enhances motion detection by identifying abrupt and irregular movements.
- The R3D model captures spatio-temporal features directly from video sequences.
- The LSTM network models temporal dependencies, enabling better understanding of action progression.

This multi-stream architecture significantly reduces false positives compared to traditional frame-based methods.

E. Comparison with Baseline Models

Compared to conventional approaches:

- Frame-based CNN models lack temporal understanding, resulting in lower accuracy.
- Motion-only methods fail to distinguish between normal and aggressive activities.
- The proposed hybrid model improves accuracy from approximately 88% (CNN only) to 96%.

F. Real Time Performance

The system was tested in real-time using a live camera feed. It successfully detected violent activities and generated alerts with minimal delay. The system demonstrated stable performance under varying lighting conditions and dynamic environments.

G. Alert Generation System

The alert generation system is a critical component of the proposed surveillance framework, designed to provide real-time notifications upon detection of violent activities.

Fig. 4. Alert Information

1) Detection Trigger Mechanism: The trained model outputs a probability score using a sigmoid activation function. If the predicted probability exceeds a predefined threshold (typically 0.5), the system classifies the activity as *violent* and triggers the alert mechanism.

2) Real Time Alert Generation: Once violence is detected, the system immediately initiates an alert process. This includes:

- Capturing the current frame as visual evidence
- Storing the image or video segment for further analysis
- Sending a notification to the authorized user or security personnel

3) Notification Mechanism: The alert is transmitted through a communication interface such as Telegram API, email, or mobile notification systems. The notification typically contains:

- Timestamp of the detected event
- Captured image or video snapshot
- Alert message indicating potential violence

4) Cool down Mechanism: To prevent redundant alerts, a cooldown interval is implemented. After an alert is triggered, the system temporarily suppresses further notifications for a short duration. This avoids alert flooding during continuous or prolonged events.

5) System Integration: The alert system is tightly integrated with the real-time detection pipeline. It operates concurrently with video processing, ensuring minimal delay between detection and notification. This enables rapid response and improves the overall effectiveness of the surveillance system.

6) Performance Considerations: The alert system is optimized for low latency and reliability. Efficient data handling and lightweight communication protocols ensure that alerts are delivered promptly, even under real-time processing constraints.

H. Limitations

Despite strong performance, the system has certain limitations:

- High computational complexity due to 3D CNN and LSTM.
- Sensitivity to extreme lighting and occlusion conditions.
- Dependence on the quality and diversity of training data.

I. Summary

Overall, the experimental results validate that the proposed system achieves high accuracy, robust performance, and reliable real-time violence detection, making it suitable for intelligent surveillance applications.

VI. Conclusion

This paper presented a smart campus surveillance system based on drone-assisted video monitoring and deep learning techniques for real-time violence detection. The proposed framework integrates spatial, motion, and temporal feature extraction using CNN/ResNet, Optical Flow, ResNet3D (R3D), and LSTM networks to effectively analyze dynamic human activities in video sequences.The system demonstrated strong performance in distinguishing violent and non-violent activities, achieving an overall accuracy of approximately 96%, along with high precision, recall, and F1-score. The combination of multi-stream feature extraction and temporal modeling significantly improves detection reliability compared to conventional frame-based or motion-only approaches.Furthermore, the integration of a real-time alert generation mechanism enhances the practical applicability of the system by enabling immediate notification and response to critical events. The use of drone-based surveillance provides improved coverage, flexibility, and reduced dependency on fixed monitoring systems.Overall, the proposed approach provides a scalable and intelligent solution for automated campus surveillance, contributing to enhanced safety and security in real-world environments.

References

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–8.

[2] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 1446–1453.

[3] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 2758–2766.

[4] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Taipei, Taiwan, 2019, pp. 1–8.

[5] M. M. Soliman *et al.*, "Violence recognition from videos using deep learning techniques," in *Proc. Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, 2019.

[6] D. Avola *et al.*, "2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2481–2496, Oct. 2020.

[7] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 14360–14369.

[8] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 200–210.

[9] K. Cao *et al.*, "Few-shot video classification via temporal alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10615–10624.

[10] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large-scale video database for violence detection," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 4183–4190.

[11] P. Sernani *et al.*, "Deep learning for automatic violence detection: Tests on the AIRTLab dataset," *IEEE Access*, vol. 9, pp. 160580–160595, 2021.

[12] S. Szymanowicz, J. Charles, and R. Cipolla, "Discrete neural representations for explainable anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 1506–1514.

[13] Y. Luo, D. Liu, and D. Tao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2017, pp. 439–444.

[14] M. Hasan *et al.*, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 733–742.

[15] S. Ionescu *et al.*, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2895–2903.

[16] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[17] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 2929–2936.

[18] H. Zhang, D. Liu, and Z. Xiong, "Convolutional neural network-based video super-resolution for action recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, 2018, pp. 746–750.

[19] H. Qian and J. Huang, "Two-stream sparse feature non-local spatiotemporal residual convolutional neural network for human action recognition," in *Proc. Int. Symp. Robot. Intell. Manuf. Technol. (ISRIMT)*, 2024, pp. 300–303.

[20] C.-J. Huang, M. Gochoo, and T.-H. Tan, "Two-stream architecture using RGB-based ConvNet and pose-based LSTM for video action recognition," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, 2023, pp. 127–131.

[21] A. Mihanpour, M. J. Rashti, and S. E. Alavi, "Human action recognition in video using DB-LSTM and ResNet," in *Proc. Int. Conf. Web Res. (ICWR)*, 2020, pp. 133–138.

[22] D. Feng and F. Ren, "Dynamic facial expression recognition based on two-stream CNN with LBP-TOP," in *Proc. IEEE Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, 2018, pp. 355–359.

[23] V.-M. Khong and T.-H. Tran, "Improving human action recognition with two-stream 3D convolutional neural network," in *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, 2018, pp. 1–6.