WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real-Time Violence Detection Using ViT+BiLSTM

Prof. Vidyadhar Hanji [a], Mr. Amit Rangarava Ingale [b], Mr. Karan Chandrakant Patil [c], Mr. Kelvin Paulu Furtado [d], Mr. Mohammad Sameer Mulla [e]

[a] Assistant Professor, [b] UG Student, [c] UG Student, [d] UG Student, [e] UG Student

[a,b,c,d,e] Department of Computer Science And Engineering, Angadi Institute of Technology And Management, Belagavi, Karnataka, India

[a] vidyadhar.hanji@aitmbgm.ac.in, [b] amitingaleimop@gmail.com, [c] karancpatil28@gmail.com, [d] kelvinpf00@gmail.com, [e] imsam8x@gmail.com

ABSTRACT:

Real-time violence detection is a crucial research area in computer vision and artificial intelligence, aimed at identifying violent or harmful human behaviors from video streams for the purpose of enhancing public safety. This project focuses on detecting and classifying violent activities such as fighting or the presence of weapons by implementing advanced deep learning techniques. The proposed system utilizes a hybrid architecture combining Vision Transformer (ViT) for spatial feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) networks for temporal sequence modeling. By processing video frames in real-time, the system effectively captures both visual context and motion dynamics to distinguish violent from normal behavior. The use of large, annotated datasets and pretrained models enables high accuracy and fast inference, making it suitable for live surveillance scenarios. Despite challenges such as privacy concerns and real-time data processing constraints, this approach enhances automated monitoring systems and provides a practical framework for deploying intelligent video surveillance in public spaces, events, and sensitive locations.

Keywords: Real-time violence detection, Vision Transformer, LSTM, BiLSTM, Deep learning, Video surveillance, Human activity recognition.

Introduction:

Violence detection in real-time video streams has emerged as a crucial research area due to the increasing need for automated surveillance in public safety and security applications. Traditional surveillance systems rely heavily on human operators to continuously monitor multiple camera feeds, which is not only inefficient but also prone to human error, delayed response, and fatigue-related inaccuracies. With the advancement of deep learning and computer vision, automated behavior recognition has become a feasible solution for identifying violent gestures, aggressive activities, and the presence of weapons in videos.

Recent studies highlight the effectiveness of hybrid spatial–temporal deep learning architectures for activity recognition. Spatial information captures appearance-based features such as body posture and objects, while temporal information captures movement patterns across consecutive frames. However, most existing systems face challenges such as low accuracy in complex scenes, poor generalization under varying lighting conditions, and high computational requirements that limit real-time deployment.

To address these challenges, this research proposes a hybrid Vision Transformer (ViT) and Bidirectional Long Short-Term Memory (BiLSTM) model designed specifically for real-time violence detection. The Vision Transformer extracts high-level spatial features from each frame using self-attention, while the BiLSTM models temporal dependencies across the video sequence. In addition, auxiliary modules such as MediaPipe-based pose estimation and YOLO-based weapon detection enhance the system's capability to detect fighting poses and weaponized behavior. The system includes a lightweight Flask-based web interface, enabling real-time monitoring, alert generation, and evidence logging.

The objective of this work is to develop a robust, scalable, real-time violence detection framework capable of deployment in CCTV networks, educational institutions, public spaces, and drone-based surveillance systems.

Methodology:

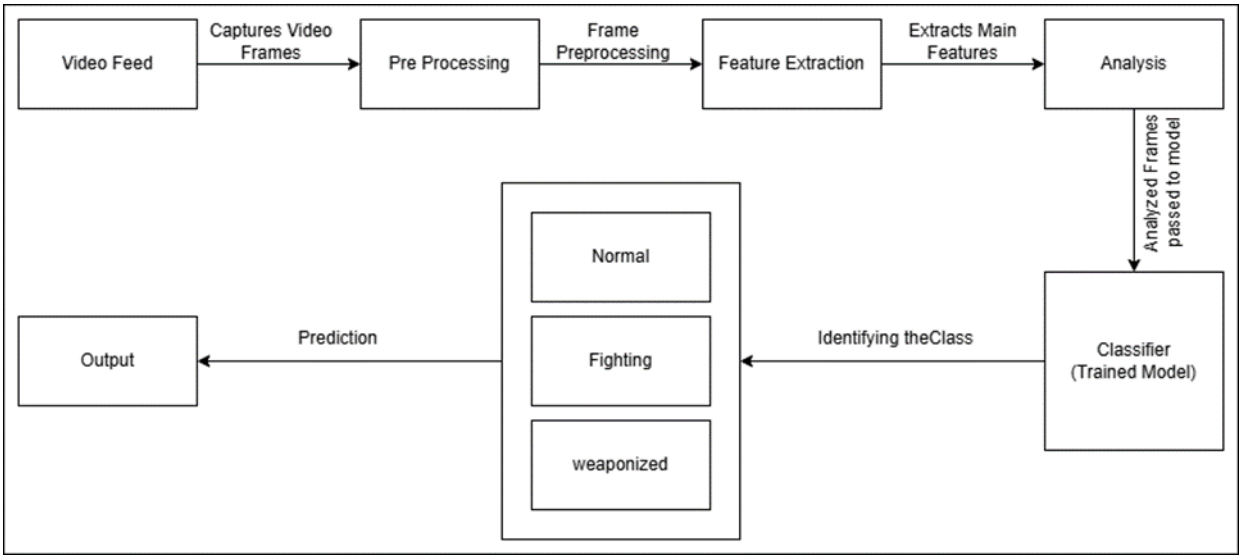

Figure 1: System Architecture

The proposed methodology integrates a sequence of carefully designed stages, each playing a crucial role in achieving high accuracy and real-time performance.

2.1 Data Collection

A diverse dataset of violence and non-violence videos is compiled from public datasets, surveillance footage, and online sources. The dataset includes three primary categories: *Normal*, *Fighting*, and *Weaponized*. Variation in lighting, camera angle, crowd density, and background complexity ensures high generalizability.

2.2 Frame Preprocessing

Each video is decomposed into frames at fixed intervals. Frames undergo resizing (224×224 or 384×384 pixels), normalization, noise reduction, and tensor conversion. Unnecessary frames are discarded using sampling techniques to minimize redundancy. This ensures consistent input quality and reduces computational load.

2.3 Feature Extraction using Vision Transformer (ViT)

ViT divides the image into patches and processes them using multi-head self-attention to extract deep spatial features. Unlike CNNs, ViT does not rely on convolutional filters, enabling stronger global contextual understanding. The extracted embeddings preserve posture details, object shape, and scene semantics.

2.4 Temporal Analysis using BiLSTM

Since violence is inherently temporal, the BiLSTM processes sequential features to capture movement intensity, direction, and rapid posture transitions. The bidirectional architecture analyzes both forward and backward dependencies, enabling better understanding of subtle actions like punching, kicking, or aggressive gestures.

2.5 Auxiliary Modules

- **Pose Detection:** MediaPipe identifies body landmarks to analyze fighting stances, arm movements, and aggressive postures.
- **Weapon Detection:** YOLO (Ultralytics) detects weapons such as knives or firearms, adding another layer of threat assessment.

2.6 Classification

A fusion module combines outputs from ViT, BiLSTM, pose estimation, and YOLO-based detections. A decision logic unit assigns the final class label (Normal, Fighting, or Weaponized) based on weighted confidence scores.

2.7 Deployment and Alert System

The trained model is deployed using Flask. The system streams real-time video, overlays detection results, triggers audio alerts via Pygame, and sends automated email notifications. This ensures instant response and improves situational awareness.

Figure 2 shows the Use Case Diagram of the system, which explains how the user interacts with the application. The primary actor in the system is the Operator, who can perform actions such as viewing live surveillance, configuring system thresholds, and reviewing notifications. In the background, the system automatically performs tasks including capturing live video streams, detecting violent or suspicious activities, classifying frames, and generating alerts. The diagram ensures that both manual and automated responsibilities are clearly defined. This representation helps establish functional requirements, ensuring the system meets real-world operational demands such as automation, monitoring, and real-time alerting.

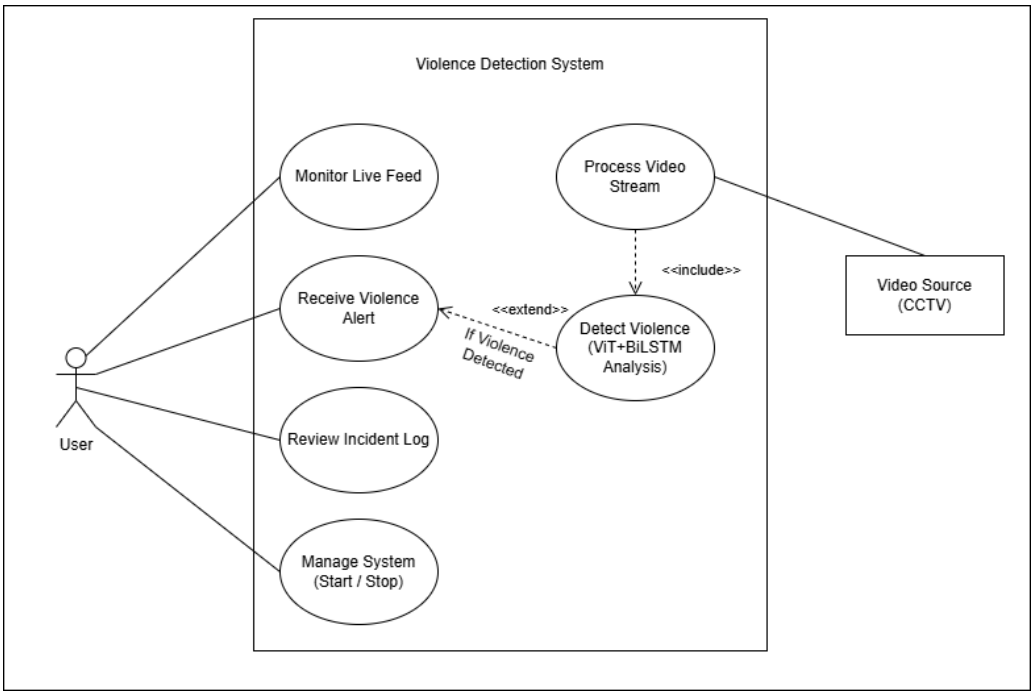

Figure 1: Use Case Diagram

Figure 3 shows the Activity Diagram of the proposed system, representing the step-by-step workflow from input to output. The process begins with the system initializing and capturing frames from the video feed. These frames undergo preprocessing and feature extraction before being analyzed by the trained model. A decision point evaluates whether the detected activity falls under violent or weaponized behavior. If detected, alerts such as messages, logs, or sound notifications are triggered; otherwise, the system continues passive monitoring. This diagram effectively demonstrates how the system operates continuously in a loop while responding dynamically to detected activities.

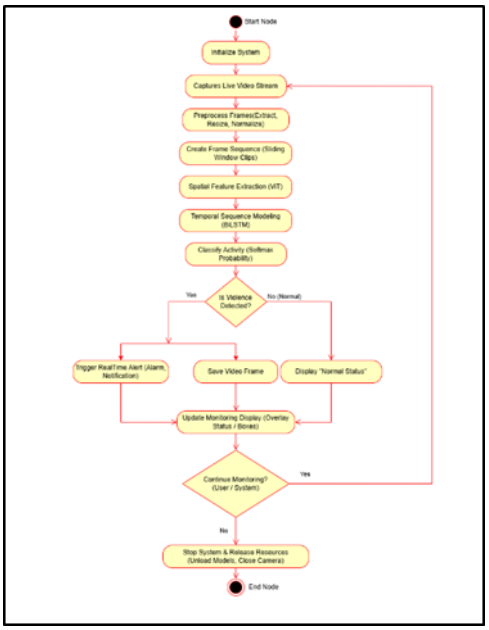

Figure 3: Activity Diagram

3. Equations:

3.1 Frame Representation:

$$F_t \in \mathbb{R}^{H \times W \times 3}$$

3.2 Patch Embedding (ViT):

Divide Image into n patches

$$N = \frac{HW}{P^2}$$

Project each Image into embedding

$$E_i = W_p \cdot x_i + b_p$$

3.3 Positional Encoding:

$$Z_i = E_i + PE_i$$

3.4 Multi-Head Self Attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

For h heads:

$$MHSA(Z) = Concat(head_1, \ldots, head_h)W_o$$

3.5 BiLSTM Temporal Modeling:

Forward Pass:

$$\vec{h_t} = LSTM(S_t, \vec{h_{t-1}})$$

Backward Pass:

$$\overleftarrow{h_t} = LSTM(S_t, \overleftarrow{h_{t+1}})$$

Combined temporal vector:

$$H_t = \left[\vec{h_t} \parallel \overleftarrow{h_t}\right]$$

3.6 Final Classification:

Logits:

$$z = WH_t + b$$

Softmax probability:

$$P(c) = \frac{e^{z_c}}{\sum_{i=1}^{C} e^{z_i}}$$

3.7 Loss Function (Cross Entropy):

$$L = -\sum_{c=1}^{C} y_c log P(c)$$

Algorithm for the Proposed System

Algorithm: Hybrid ViT + BiLSTM

Input: Live video stream V(t)

Output: Activity class {Normal, Fighting, Weaponized}

Step1: Video Frame Accquisation

1. Start capturing live video stream.
2. Extract frames F_t at fixed interval Δt

Step2: Preprocessing

1. Resize each frame to target resolution (224 × 224).
2. Normalize pixel values.
3. Convert frames to tensor representation.
4. Select relevant frames using frame sampling.

Step3: Feature Extraction using Vision Transformer (ViT)

1. Divide frame into patches.
2. Convert patches into patch embeddings.
3. Add positional encoding.
4. Pass tokens through multi-head self-attention encoder.
5. Extract spatial feature vector S_t

Step4: Temporal Analysis using BiLSTM

1. Collect sequence of spatial features $S_1, S_2, \ldots..,S_n$
2. Feed sequence into BiLSTM
3. Compute forward hidden states $\vec{h_t}$
4. Compute backward hidden states $\overleftarrow{h_t}$
5. Concatenate both to obtain temporal feature vector.

$$\boldsymbol{H_t} = \left[\vec{\boldsymbol{h_t}} \parallel \overleftarrow{\boldsymbol{h_t}}\right]$$

Step5: Fusion and Classification

1. Combine ViT features S_t and BiLSTM temporal features H_t
2. Apply fully connected layers.
3. Compute softmax probability for classes.

$$\boldsymbol{P(c)} = \frac{\boldsymbol{e^{z_c}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{C}} \boldsymbol{e^{z_i}}}$$

Step6: Auxiliary Detection Modules

1. Detect human pose landmarks using MediaPipe.
2. Detect weapons using YOLO.

Step7: Decision Logic

1. Fuse Scores:

$$\textbf{Score}_{final} = \alpha P_{ViT} + \beta P_{BiLSTM} + \gamma P_{YOLO}$$

2. Assign class with highest final score.

Step8: Alert and Output

1. Display detection overlay on dashboard.
2. Trigger audio alert if violence detected.
3. Send notification (email/log update).

4. Results:

The proposed hybrid ViT + BiLSTM system was evaluated on a dataset containing diverse violence scenarios. Key performance metrics include:

4.1 Accuracy and Performance

- Overall classification accuracy: 99.3%

- Macro F1-Score: 99.2%
- Matthews Correlation: 0.987

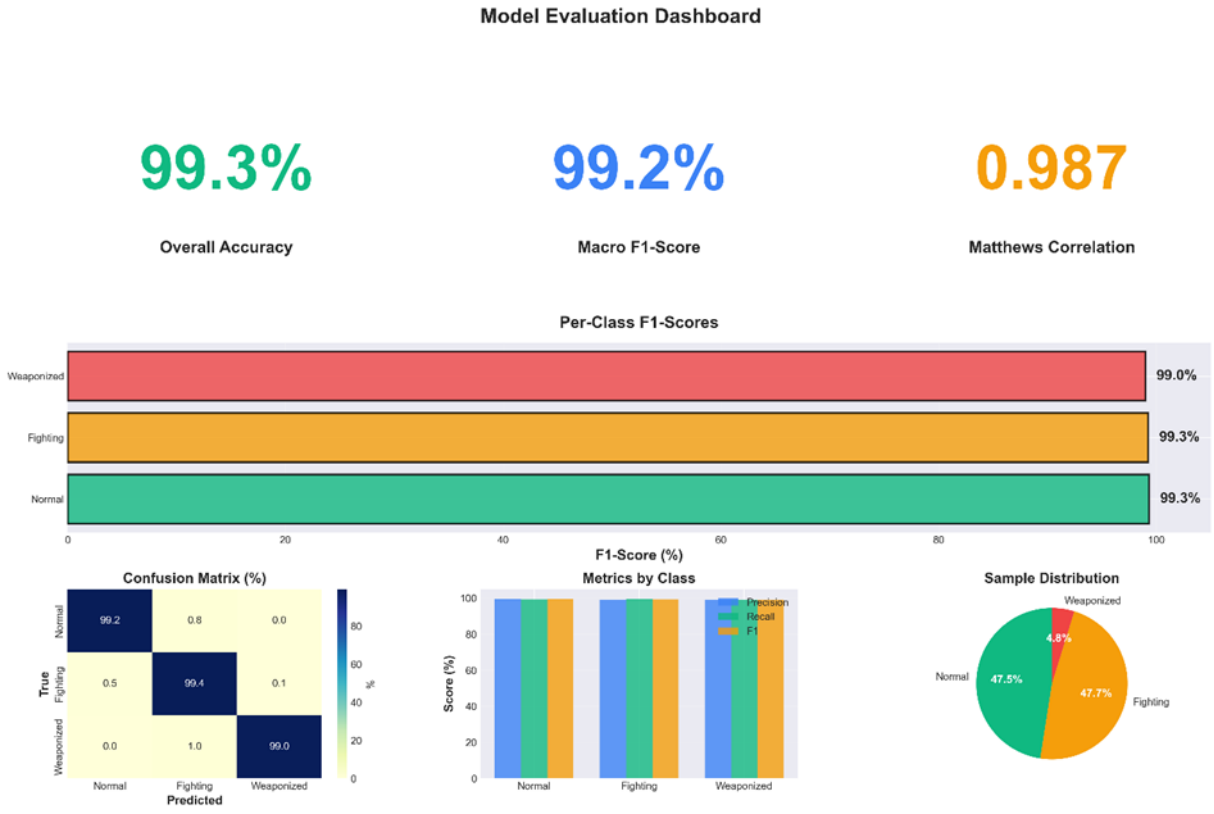

Figure 4: Model Evaluation Dashboard

4.2 Real-Time Performance

- Average inference speed: 18–25 FPS on GPU (NVIDIA 3050)
- End-to-end latency: < 150 ms
- Supports continuous streaming without frame drops due to threaded processing.

4.3 Visual Analysis

- The system correctly identifies:
- Punching, kicking, pushing
- Rapid hand movements indicating aggression
- Weapons such as knives or guns
- Multiple persons involved in chaotic scenes
- Detection overlays and confidence values are displayed in real time on the web dashboard.

4.4 Comparison with Existing Models

- Compared to CNN-LSTM and 3D CNN-based models, the proposed ViT+BiLSTM hybrid shows:
- Higher temporal sensitivity
- Better generalization
- Lower false positives
- Improved embedding quality due to self-attention

5. Conclusion

This research presents a robust and scalable real-time violence detection system combining the strengths of the Vision Transformer and BiLSTM architecture. By integrating auxiliary modules like MediaPipe pose detection and YOLO-based weapon identification, the system demonstrates high accuracy across complex scenarios. The hybrid spatial–temporal learning approach enhances the detection of both violent actions and weaponized behavior.

The real-time web interface, alert system, and evidence logging features make the system practical for deployment in public surveillance, educational institutions, law enforcement, and drone-based monitoring applications. The performance evaluation indicates strong accuracy, reliability, and real-time capability, proving its effectiveness for real-world deployment.

Future extensions may include multi-camera integration, edge-AI deployment on embedded devices, crowd violence prediction, and reinforcement learning-based adaptive thresholds. This work contributes to the field of intelligent surveillance by offering a practical, efficient, and accurate solution for automated violence detection.

References

[1] Monji Mohamed Zaidi, Gabriel Avelino Sampedro, Ahmad Almadhor, Shtwai Alsubai, Abdullah Al Hejaili, Michal Gregus, Sidra Abbas (2024). “Suspicious Human Activity Recognition from Surveillance Videos Using Deep Learning”. IEEE

[2] K Nithesh, Nikhath Tabassum, D. D. Geetha, R D Anitha Kumari (2023). “Anomaly Detection in Video Surveillance Using Deep Learning”. IEEE

[3] Rajesh Kumar Yadav, Rajiv Kumar (2022). “A Survey on Anomaly Detection Techniques in Video Surveillance”. IEEE

[4] S. A. Quadri, Komal (2022). “Suspicious Activity Detection Using 3D Convolutional Neural Networks”. Research Gate

[5] Nahum Kiryati, Tammy Riklin Raviv, Yan Ivanchenko, Shay Rochel (2009). “Real-time Abnormal Human Activity Detection in Surveillance Videos”. IEEE

[6] M. Perez, A. C. Kot and A. Rocha, “Detection of Real-world Fights in Surveillance Videos”, ICASSP 2019 - 2019 IEEE

[7] C. V. Amrutha, C. Jyotsna and J. Amudha, "Deep Learning Approach for Suspicious Activity Detection from Surveillance Video," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020

[8] W. Sultani, C. Chen and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018

[9] J. Wei, J. Zhao, Y. Zhao and Z. Zhao, "Unsupervised Anomaly Detection for Traffic Surveillance Based on Background Modeling," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018

[10] S. Ma, L. Sigal and S. Sclaroff, "Learning Activity Progression in LSTMs for Activity Detection and Early Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016

[11] G. Varol, I. Laptev and C. Schmid, "Long-Term Temporal Convolutions for Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition

[13] Jitendra Musale, Akshata Gavhane, Liyakat Shaikh, Pournima Hagwane, Snehalata Tadge, “Suspicious Movement Detection and Tracking of Human Behavior and Object with Fire Detection using A Closed-Circuit TV (CCTV) camera”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue XII December 2017.

[14] Tian Wanga, Meina Qia, Yingjun Deng, Qi Lyua, Hichem Snoussie, “Abnormal eventdetection based on analysis of movement information of video sequence”, Article-Optik, vol152, January-2018.

[15] Elizabeth Scaria, Aby Abahai T and Elizabeth Isaac, “Suspicious Activity Detection in Surveillance Video using Discriminative Deep Belief Netwok”, International Journal of Control Theory and Applications Volume 10, Number 29 -2017.

[16] Dinesh Jackson Samuel R, Fenil E, Gunasekaran Manogaran, Vivekananda G.N, Thanjaivadivel T, Jeeva S, Ahilan A, “Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM”, The International Journal of Computer and Telecommunications N e tworking,2019.

[17] Kwang-Eun Ko, Kwee-Bo Sim “Deep convolutional framework for abnormal behavior detection in a smart surveillance system." Engineering Applications of Artificial intelligence ,67 (2018). [6] Yuke Li “A Deep Spatiotemporal Perspective for Understanding Crowd Behavior”, IEEE Transactions on multimedia, Vol. 20, NO. 12, December 2018. [7] P. Bhagya Divya, Shalini, R. Deepa, Baddeli Sravya Reddy, “Inspection

of suspicious human activity in the crowdsourced areas captured in surveillance cameras", International Research Journal of Engineering and Technology (IRJET), December 2017.

[18] K. Kranthi Kumar, B. Hema Kumari, T. Saikumar, U. Sridhar, G. Srinivas, G. Sai Karan Reddy (2022). “Suspicious Activity Detection from Video Surveillance”. IJRPR