International Journal of Research Publication and Reviews

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

Real-Time Violence Detection with Visual Recognition

P.Sai Rishi[1],P.Praveen[2],P.Simhadri[3],P.Gnaneswari[4],S.Venkata Prasad[5]

GMRIT, Rajam, Vizianagaram, 532127, India

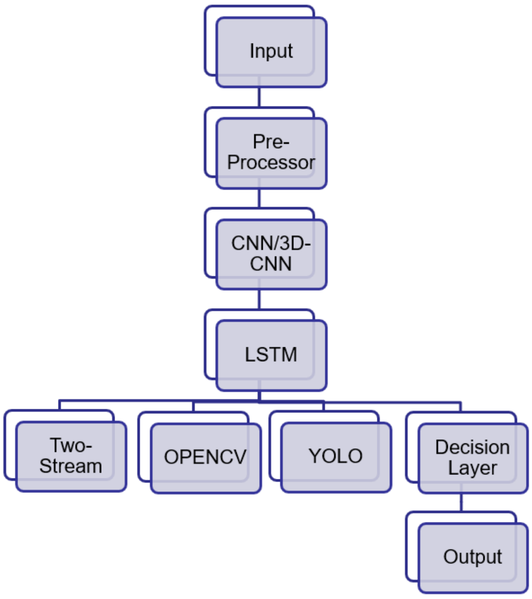
ABSTRACT :

Visual recognition using AI involves enabling computer systems to perceive and analyze the external world. This falls under a subdomain of AI known as computer vision. Visual recognition emerged as a promising field with the convergence of neuroscience, psychology, computer science, and engineering. It has significantly transformed various industries, altering how we interact with technology and the world around us. Real-life applications of visual recognition include self-driving cars, healthcare, security, and social media, among others. Algorithms employed for violence detection through visual recognition include OpenCV, Media Pipe, Support Vector Machines (SVM), and Convolutional Neural Networks (CNN) from deep learning and Rule-based Systems for activating an alarm when the violence is detected. However, challenges associated with violence detection include occlusions, varying lighting conditions, complex backgrounds, and imbalanced datasets. This paper aims to explore these aspects and address the complexities involved in violence detection using visual recognition.

Keywords: Violence Detection, Visual Recognition, Artificial Intelligence, Computer Vision, Media Pipe, Rule-based Systems, CNN ,YOLO.

1. INTRODUCTION :

With the rapid advancements in digital video and surveillance technologies, automated security systems have become essential for maintaining public safety in various environments, including smart cities, airports, and other high-traffic areas. Surveillance cameras are increasingly deployed to monitor activities, deter potential criminal behavior, and provide evidence for post-incident investigations. However, the massive volume of video data generated daily makes it challenging for human operators to manually monitor and analyze events in real time. This has spurred significant research into automatic violence detection through computer vision, aiming to enable quicker responses by automating the detection of suspicious or aggressive behaviors. Automatic violence detection systems leverage the power of deep learning, particularly Convolutional Neural Networks (CNNs) and 3D convolutional architectures, to recognize and analyze violent actions in video feeds. These advanced models can capture both spatial and temporal features, allowing them to detect violence even in complex and dynamic environments such as crowded public spaces or locations with variable lighting. Unlike traditional image processing techniques, deep learning models can learn intricate patterns and movements associated with violence, making them effective at recognizing aggressive behavior in real time. Recent progress in deep learning has made violence detection systems more reliable, especially in challenging scenarios involving partial occlusions or rapid changes in movement. This capability is essential for creating automated video surveillance solutions that can accurately identify violent incidents without relying on constant human oversight. As a result, real-time violence detection systems are increasingly being adopted in public and private sectors to enhance safety, reduce the burden on human operators, and allow for faster, more effective responses to security threats. These systems represent a promising advancement toward creating safer environments through technology, embodying the integration of artificial intelligence with real-world security needs.



2. LITERATURE SURVEY :

2.1. Cheng, M., Cai, K., & Li, M. (2021, January). RWF-2000: an open large scale video database for violence detection.(pp. 4183-4190). IEEE.

The RWF-2000 dataset is a comprehensive video collection specifically designed to improve violence detection in real-world surveillance settings. This dataset comprises 2,000 video clips, each captured by surveillance cameras, thus offering a higher level of realism compared to prior datasets that often lack diversity and scalability, with limited image quality and coverage of violent scenarios. Unlike earlier databases, RWF-2000 aims to overcome these shortcomings by providing a diverse range of violent and non-violent interactions across various environments. To effectively leverage this dataset, the authors developed a novel approach called the Flow Gated Network, which combines the strengths of 3D Convolutional Neural Networks (3D-CNNs) with optical flow analysis. This method captures both spatial and temporal features of violent actions, leading to more accurate detection results. The Flow Gated Network achieved an impressive accuracy of 87.25% on the test set, demonstrating its effectiveness and robustness in practical applications.

2.2. Construction of references Vieira, J. C., Sartori, A., Stefenon, S. F., Perez, F. L., De Jesus, G. S., & Leithardt, V. R. Q. (2022). Low-cost CNN for automatic violence recognition on embedded system. IEEE Access, 10, 25190-25202.

The paper presents a low-cost, real-time system for recognizing violence on embedded systems by employing lightweight Convolutional Neural Network (CNN) architectures such as SqueezeNet, MobileNet-V1, MobileNet-V2, and NASNet-Mobile. Designed for deployment on resource-limited devices like the Raspberry Pi, the system demonstrates that violence detection can be achieved effectively without requiring high power computational resources. The authors created a self-assembled dataset featuring both violent and non-violent scenes to train and test the models. Among the architectures tested, MobileNet-V2 achieved the highest balance between accuracy and computational efficiency, reaching an impressive accuracy of 92.05% on the dataset. This approach, running at 4 frames per second on Raspberry Pi, highlights the system's suitability for real-time monitoring in environments with limited resources. By optimizing CNN models for low computation, the authors provide a practical solution for violence detection in settings where traditional, high-power solutions may be impractical. The paper shows that MobileNet-V2 stands out, offering a strong accuracy-to-parameter ratio, making it an ideal choice for embedded applications in real-time surveillance.

2.3. Kang, M. S., Park, R. H., & Park, H. M. (2021). Efficient spatio-temporal modeling methods for real-time violence recognition. IEEE Access, 9, 76270-76285.

The paper presents an advanced approach for efficiently detecting violence in video surveillance systems in real time. It enhances 2D Convolutional Neural Networks (CNNs) by incorporating spatio-temporal features through a novel frame-grouping technique. This technique averages input frames, improving temporal awareness without imposing significant computational overhead. To further optimize the model, two lightweight modules are introduced: the “Motion Saliency Map (MSM)” and the “Temporal Squeeze-and-Excitation (T-SE)” module. The MSM prioritizes crucial spatial regions by focusing on motion-based attention, while the T-SE module emphasizes the most relevant temporal frames dynamically. Together, these components enable the system to detect violent actions, such as punching or kicking, by highlighting critical spatio-temporal regions efficiently. Demonstrating state-of the-art results on various violence detection datasets, the system minimizes computational costs, making it suitable for real-time, on-device deployment. Its lightweight design ensures practicality in real-world surveillance applications, meeting the demands of real-time performance.

2.4. Kwan-Loo, K. B., Ortíz-Bayliss, J. C., Conant-Pablos, S. E., Terashima-Marín, H., & Rad, P. (2022). Detection of violent behavior using neural networks and pose estimation. IEEE Access, 10, 86339-86352.

The paper presents an AI-driven approach for detecting violent behavior in real-time using video analysis, leveraging pose estimation combined with neural networks. This method classifies actions as violent or non-violent by examining the angles between joints in detected pedestrian poses, allowing for more precise behavior analysis. A key contribution of this work is the introduction of the “Kranok-NV dataset”, which includes annotated footage of both violent and non-violent actions, specifically segmented to enhance model training and improve the reliability of classification. Achieving an impressive accuracy rate of over 98%, this approach demonstrates high effectiveness in distinguishing violent behaviors from normal actions in varied environments. Additionally, the model's architecture is flexible enough to be adapted for other behavior detection tasks beyond violence, highlighting its potential applications in broader surveillance or monitoring contexts where behavior analysis is critical for safety or operational efficiency.

2.5. Huszar, V. D., Adhikarla, V. K., Négyesi, I., & Krasznay, C. (2023). Toward fast and accurate violence detection for automated video surveillance applications. IEEE Access, 11, 18772-18793.

The paper explores the need for efficient and reliable violence detection in automated video surveillance systems, addressing growing safety and security concerns. It focuses on using deep learning techniques, specifically “3D Convolutional Neural Networks (CNNs)”, to capture spatial and temporal details in video data, which are critical for accurate violence recognition. By incorporating 3D CNNs, the method effectively analyzes sequences of frames, allowing it to understand dynamic movements that are often indicative of violent behavior. Additionally, the authors enhance this approach by leveraging pre-trained action recognition models, which serve as a foundation to boost efficiency and accuracy. The models are further refined using smart network architectures that can model dynamic relationships between entities in the video, improving the system’s ability to detect subtle and complex patterns associated with violence. This approach provides a promising solution, combining speed and precision, and is well-suited for real-time surveillance applications that require quick response capabilities.

3. METHODOLOGY :

3.1. RWF-2000 Dataset for Realistic Violence Detection

Methods used in this paper utilizes 3D-CNNs for learning spatio-temporal features from video data, capturing both spatial and temporal information to detect violence. The authors mention using the Optical Flow and RGB frames Used to capture the motion between video frames, which helps in understanding the movement of objects, aiding in the identification of violent actions. The implementation of the algorithms in the paper involves using 3D-CNNs to extract spatiotemporal features from video frames, while optical flow is applied to capture motion patterns, and these are combined in the Flow Gated Network to improve violence detection accuracy.

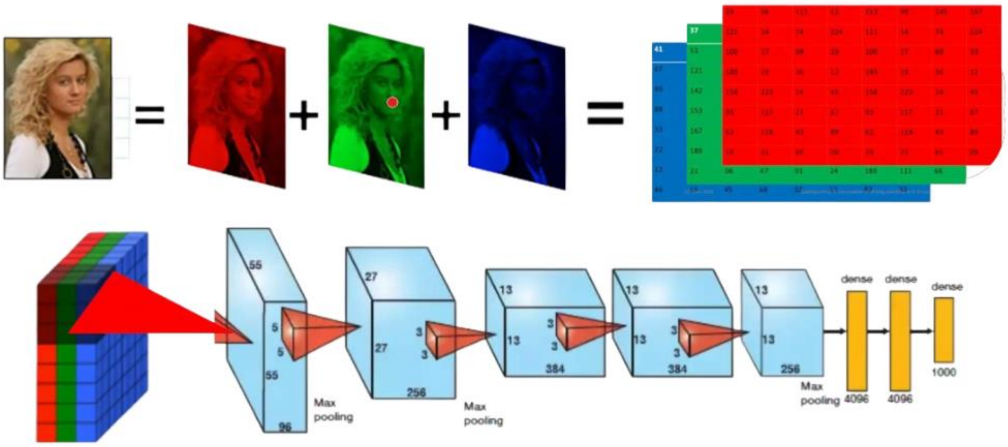


Fig. 1 - Image Classification using RGB Frames

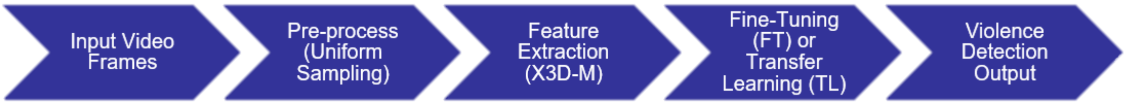

Input Video Frames
Pre-process (Uniform Sampling)
Feature Extraction (X3D-M)
Fine-Tuning (FT) or Transfer Learning (TL)
Violence Detection Output

Fig. 2 - Flow Chart of Flow Gated Network implementing on RWF-2000

3.2. Low-Cost Violence Recognition on Embedded Systems Using Mobile CNN Architectures:

The paper discusses the use of various methods and techniques for violence recognition, including a lightweight CNN method that has shown higher precision rates compared to traditional methods. The paper also mentions the violence recognition on embedded systems like Raspberry Pi.The paper also discusses the use of the SqueezeNet, MobileNet-V1, MobileNet-V2, and NASNet-Mobile architectures for violence detection. SqueezeNet is a lightweight CNN architecture used for violence detection on embedded systems with limited resources.MobileNet-V1 is a lightweight CNN used for efficient violence detection in resource-constrained environments. MobileNet-V2, an optimized version of MobileNet, achieves the best balance between accuracy and computational efficiency for real-time violence detection. NASNet-Mobile is a CNN architecture optimized through neural architecture search, used for violence detection with a focus on computational efficiency.

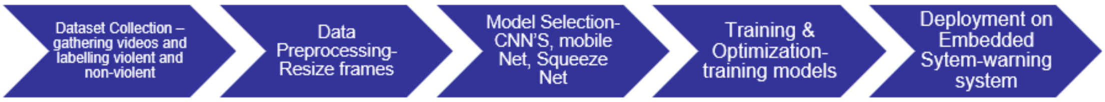

Dataset Collection – gathering videos and labelling violent and non-violent
Data Preprocessing-Resize frames
Model Selection-CNN'S, mobile Net, Squeeze Net
Training & Optimization-training models
Deployment on Embedded Sytem-warning system

Fig. 3 - Flow Chart of CNN Architectures implementing on Dataset resulting warning system

3.3. Efficient Spatio-Temporal Modeling for Real-Time Violence Detection:

The paper incorporates Motion Saliency Map (MSM) to identify moving regions in a video by calculating the differences between consecutive frames. This method highlights areas with significant motion, isolating active regions such as people involved in violent actions while minimizing background distractions.

By computing the absolute difference between pixels of consecutive frames, MSM efficiently captures dynamic changes and suppresses static elements, making it highly effective in detecting violent behaviors. The Temporal Squeeze-and-Excitation (T-SE) Block adjusts temporal attention by emphasizing critical frames related to violent actions. By using global average pooling (squeeze) followed by excitation through fully connected layers, this mechanism assigns higher importance to the most relevant moments in the video.

This improves the detection accuracy of violent actions while maintaining minimal computational overhead, allowing the system to focus on key timeframes for more efficient action recognition.

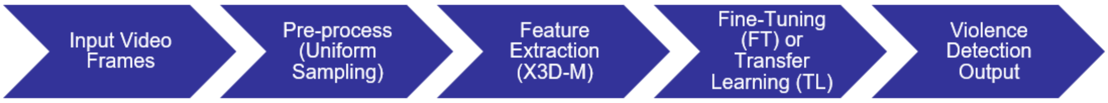

Input Video Frames
Pre-process (Uniform Sampling)
Feature Extraction (X3D-M)
Fine-Tuning (FT) or Transfer Learning (TL)
Violence Detection Output

Fig. 3 - Flow Chart of MSM and T-SE implementing on Dataset

3.4. Pose Estimation and Neural Networks for Behavior Classification:

The paper utilizes a combination of advanced methods to efficiently detect violence in surveillance video using YOLOv3, it is employed for pedestrian detection, accurately isolating humans in video frames and focusing on potential violent subjects. OpenPose is used for pose estimation, identifying key body joints like elbows and knees, and calculating joint angles to capture movement dynamics. This allows the system to detect the subtle changes in body posture associated with violent behavior.A simple neural network classifies the pedestrian's actions by analyzing joint angles across multiple frames, distinguishing between violent and non-violent movements.

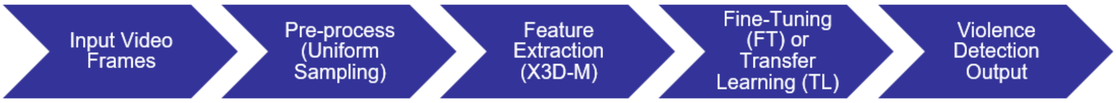

Input Video Frames
Pre-process (Uniform Sampling)
Feature Extraction (X3D-M)
Fine-Tuning (FT) or Transfer Learning (TL)
Violence Detection Output

Fig. 4 - Flow Chart of MSM and T-SE implementing on Dataset

3.5. 3D CNNs for Fast and Accurate Violence Detection in Surveillance:

The paper employs the X3D-M model (Expandable-3D), a 3D convolutional neural network (CNN) designed to capture both spatial and temporal features from video frames. This dual focus makes it ideal for action recognition and violence detection. Initially pre-trained on the Kinetics-400 dataset, is fine-tuned to specifically detect violent behavior.

The methodology integrates two configurations:

- FT model, which fine-tunes the X3D-M model for violence detection by updating its parameters.
- TL model, which uses pre-trained X3D-M features and adds fully connected layers to emphasize violence-specific cues.

The models are evaluated across multiple datasets to test their generalizability and are subjected to experiments simulating real-world conditions, such as video compression and standalone system use, ensuring their robustness and applicability in practical settings.

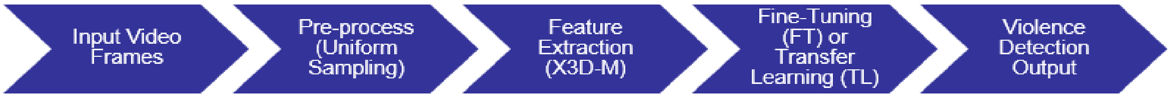

Input Video Frames
Pre-process (Uniform Sampling)
Feature Extraction (X3D-M)
Fine-Tuning (FT) or Transfer Learning (TL)
Violence Detection Output

Fig 4- Flow Chart of X3D-M and FL &TL implementing on Dataset

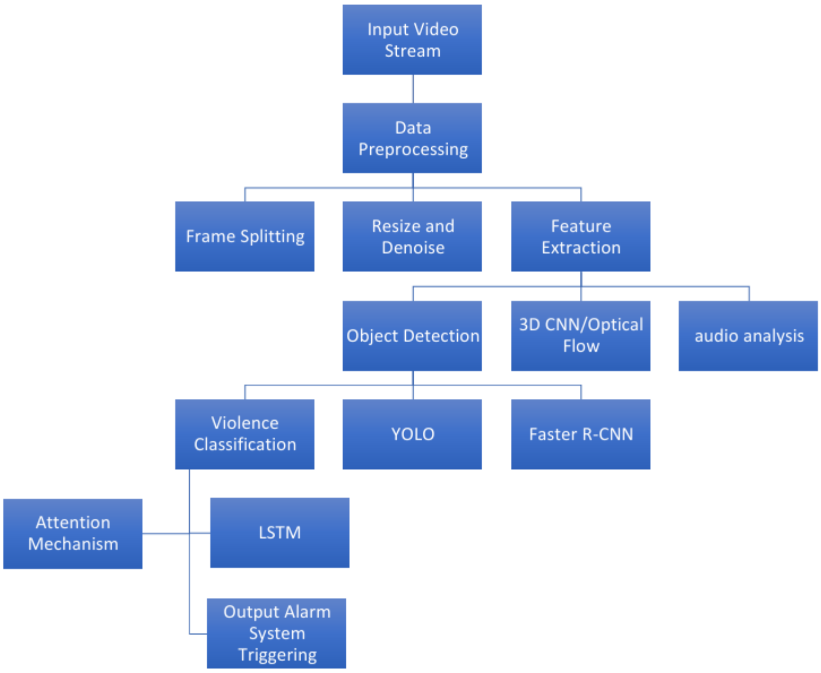


Fig 5 - Flow Chart of violence Detection with alarm System

4. RESULTS and DISCUSSION :

Paper	Methodology	Results & Accuracy	Dataset	Advantages	Disadvantages
Cheng, M., Cai, K., & Li, M. (2021)	Combines RGB frames and optical flow using Flow Gated Network (FGN). Trained on RWF-2000.	Accuracy: 87.25%	RWF-2000 (2,000 video clips)	- Large-scale dataset - Efficient model design	- Manual data cleaning - Limited to 5-second clips
Vieira, J. C., Sartori, A., et al. (2022)	Mobile CNN architectures (MobileNet-V2) optimized for embedded systems like Raspberry Pi.	Accuracy: 92.05%	Self-assembled dataset + UCF-101	- Low-cost - Efficient models for constrained environments	- Limited hardware performance - Struggles in crowded scenarios
Kang, M. S., Park, R. H., & Park, H. M. (2021)	Spatiotemporal modeling using Motion Saliency Map (MSM) and T-SE blocks.	State-of-the-art on multiple datasets	RWF-200	- Real-time efficiency - Lightweight modules	- Dependence on specific camera setups - Limited long-term temporal understanding

Kwan-Loo, K. B., et al. (2022)	Combines YOLOv3 for pedestrian detection, OpenPose for pose estimation, and cnn for classification.	Accuracy: 98%	Kranok-NV	- High accuracy - Adaptable for detecting behaviors beyond violence	- Dataset limitations - Challenges in uncontrolled environments
Huszar, V. D., et al. (2023)	Leverages X3D-M 3D-CNNs with fine-tuning and transfer learning for spatial-temporal features.	High accuracy across datasets	Multiple datasets, including Kinetics-400	- High accuracy - Robustness in various conditions	- Heavy reliance on data - Struggles in crowded scenarios

Based on the comparison table provided, **"Real-Time Violence Detection"** is a crucial area of research that focuses on designing systems capable of promptly identifying violent actions in videos or live scenarios. Several studies have been conducted to address this challenge, utilizing different techniques and datasets. For instance, Peixoto et al. (2019) developed a method that combines subjective violence detection approaches with advanced signal processing and machine learning techniques. These systems aim to improve accuracy and reduce detection time by analyzing key features such as motion dynamics, object interactions, and environmental context in video data. The table highlights various methodologies, datasets, and evaluation metrics employed in violence detection studies, providing a foundation for understanding the strengths and limitations of existing approaches. Real-time violence detection systems are integral to enhancing public safety, offering immediate responses to potential threats, and paving the way for advancements in automated surveillance systems.

4.Conclusion :

The research addresses the complex challenge of real-time violence detection through visual recognition. By utilizing advanced algorithms and deep learning models, particularly Convolutional Neural Networks (CNNs) and techniques like pose estimation, this study aims to significantly enhance the accuracy and efficiency of detecting violent behavior in surveillance footage. The model is trained on a variety of datasets, including RWF-2000 and Kinetics-400, ensuring robustness and adaptability across different settings and environments. In addition, the study proposes several key enhancements to address existing limitations, such as latency and the occurrence of false positives. These include multi-modal detection, optimized real-time processing, and the implementation of a dynamic alarm system to alert authorities promptly and reliably. Furthermore, context-aware detection and a continuous learning mechanism are incorporated, enabling the system to adapt to evolving environments and recognize new patterns of violent behavior over time. These advancements highlight the potential of AI-driven surveillance in supporting public safety by providing timely and accurate detection in complex and high-risk scenarios.

REFERENCES :

1. Cheng, M., Cai, K., & Li, M. (2021, January). RWF-2000: an open large scale video database for violence detection. (pp. 4183-4190). IEEE.
2. Vieira, J. C., Sartori, A., Stefenon, S. F., Perez, F. L., De Jesus, G. S., & Leithardt, V. R. Q. (2022). Low-cost CNN for automatic violence recognition on embedded system. *IEEE Access*, *10*, 25190-25202.
3. Kang, M. S., Park, R. H., & Park, H. M. (2021). Efficient spatio-temporal modeling methods for real-time violence recognition. *IEEE Access*, *9*, 76270-76285.
4. Kwan-Loo, K. B., Ortíz-Bayliss, J. C., Conant-Pablos, S. E., Terashima-Marín, H., & Rad, P. (2022). Detection of violent behavior using neural networks and pose estimation. *IEEE Access*, *10*, 86339-86352.
5. Huszar, V. D., Adhikarla, V. K., Négyesi, I., & Krasznay, C. (2023). Toward fast and accurate violence detection for automated video surveillance applications. *IEEE Access*, *11*, 18772-18793.
6. Ullah, F. U. M., Ullah, A., Muhammad, K., Haq, I. U., & Baik, S. W. (2019). Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors*, *19*(11), 2472.
7. Wu, P., Liu, X., & Liu, J. (2022). Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, *25*, 1674-1685.
8. Khan, M., El Saddik, A., Gueaieb, W., De Masi, G., & Karray, F. (2024). VD-Net: An Edge Vision-Based Surveillance System for Violence Detection. *IEEE Access*, *12*, 43796-43808.
9. Serrano, I., Deniz, O., Espinosa-Aranda, J. L., & Bueno, G. (2018). Fight recognition in video using hough forests and 2D convolutional neural network. *IEEE Transactions on Image Processing*, *27*(10), 4787-4797.
10. Bhatti, M. T., Khan, M. G., Aslam, M., & Fiaz, M. J. (2021). Weapon detection in real-time cctv videos using deep learning. *Ieee Access*, *9*, 34366-34382.

11. Roman, D. G. C., & Chávez, G. C. (2020, November). Violence detection and localization in surveillance video. In *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)* (pp. 248-255). IEEE.
12. Mumtaz, A., Sargano, A. B., & Habib, Z. (2018, December). Violence detection in surveillance videos with deep network using transfer learning. In *2018 2nd European conference on electrical engineering and computer science (EECS)* (pp. 558-563). IEEE.
13. Pang, W., Xie, W., He, Q., Li, Y., & Yang, J. (2022). Audiovisual dependency attention for violence detection in videos. *IEEE Transactions on Multimedia*, *25*, 4922-4932.
14. Choqueluque-Roman, D., & Camara-Chavez, G. (2022). Weakly supervised violence detection in surveillance video. *Sensors*, *22*(12), 4502.
15. Peixoto, B., Lavi, B., Martin, J. P. P., Avila, S., Dias, Z., & Rocha, A. (2019, May). Toward subjective violence detection in videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8276-8280). IEEE.