WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Surveillance-Based Violence and Anomaly Detection using Deep Learning & Machine Learning Techniques

Shaik Nazeer

Computer Science Student
GMRIT Rajam, India
nazeershaik94369@gmail.com

ABSTRACT

This paper introduces a fresh way to automatically spot violence in video surveillance using smart learning methods. It combines CNNs and RNNs to pull out important details from the video as it plays. The system scans through video frames to catch violent actions, with CNN layers focusing on individual frames and RNN layers, especially LSTM networks, looking at how the frames fit together. To make the model work better, we train it in two steps, which helps it recognize violence more accurately, making it more reliable in a variety of situations compared to older methods deep learning approach could lead to quick detection of violent events, allowing for fast reactions in surveillance settings.

Keywords—violence Detection, Surveillance System, Deep Learning, Convolutional Neural network (CNN), Spatiotemporal Analysis

1.Introduction

Public safety has a concern due to rising violence in public and private spaces. Traditional surveillance monitoring relies heavily on human operators, which can easily miss important details due to fatigue, distraction, or the volume of video data. The demand for automated, real-time violence detection has driven research into intelligent surveillance solutions that can analyze and interpret surveillance footage to detect violent activities autonomously. Deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Network, offers powerful techniques for automated violence detection. This proposes a deep based approach for violence detection in surveillance video feeds, integrating CNNs and LSTMs to detect violent behaviors in real-time. The proposed model achieves high accuracy and robustness in detecting violence across various settings. This work contributes to the development of a deep learning modeltailored for violence detection in surveillance environments, a novel two-stage training process, and an evaluation demonstrating its effectiveness in diverse, real-world scenarios.

Convolutional Neural Networks are deep learning architectures designed for the image and video analysis tasks, such as image classification, object detection, and activity recognition. In surveillance-based violence detection, CNNs serve as a core technology, extracting meaningful spatial features from video frames to differentiate violent from non-violent ones. CNNs consist of multiple layers, including convolutional, pooling, and fully connected layers, which learn spatial features hierarchically. CNNs are particularly suited for action recognition, where they can identify specific patterns associated in the violence of with violent behavior.

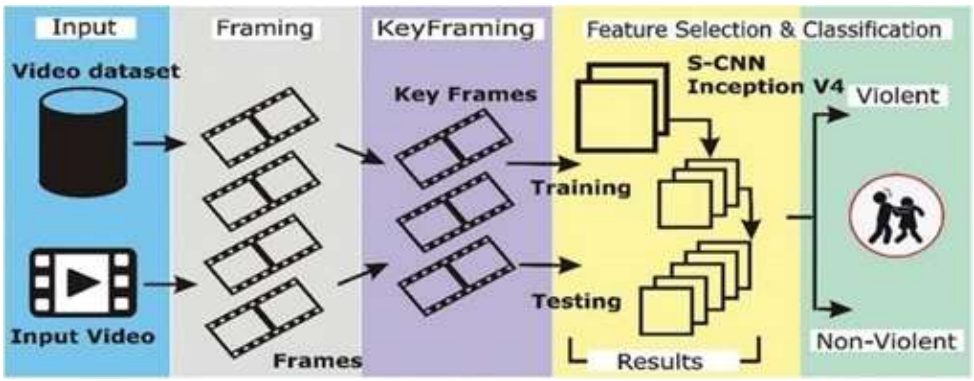

Fig1.Architecture for violence detection using Machine Learnirng technique

Src:(https://www.researchgate.net/figure/The-proposed-framwork-for-violence-detection-using-deep-learning-based-technique_fig1_353379643)

To process each video frame by frame, CNNs learn spatial features relevant to violent actions. However, violence detection in surveillance requires both spatial and temporal analysis, leading to the integration of CNNs with temporal models like Recurrent Neural Networks or Long Short-Term Memory(LSTM)for the and the algorithms are in networks.

Pre-trained CNN models, such as ResNet, VGG, and Inception, are used to leverage pre-learned spatial features for identifying violence-specific patterns in surveillance videos. CNNs offer advantages such as efficiency in feature extraction, scalability, and real-time processing. However, they also face challenges such as sensitivity to environmental variations and the inability to capture temporal context

2. LITERATURE SURVEY

In 2024 shin and his team introduced a model designed spot videos they applied deep learning to refine how features are represented particularly when identifying violent actions their approach combines both visual and audio data to improve the detection of these events at the same time khan and his group launched vd-net a deep learning system for monitoring that can detect violence in real-time on edge devices enhancing the effectiveness and scalability of video surveillance systems janani and her colleagues explored how blending audio with visuals can boost the detection of human actions and violent behavior demonstrating that really improves recognition pham and his team presented a toolkit for analyzing audio and video together which is especially helpful in recognizing riots and violent incidents they employed deep learning to create a structured approach for detecting violence sun and gong introduced the bottleneck transformer model which improves video analysis for identifying violence using weak supervision chakraborty proposed a method to enhance safety and refine automated video monitoring by using models violent actions huszar and his team aimed to strike a balance between speed and accuracy for real-time violence video tang developed accurately spotting violence in animated content modifying traditional better violent scenes in animations moreover a multi-stream cnn strategy blends handcrafted features with deep learning boosting the effectiveness violence nardelli comminiello unveiled josenet a that combines multiple streams to intricate mumtaz his team suggested a quick-learning cnn model for fast recognition of violence enabling rapid deployment in real-time environments wang created a lightweight model based on a 2d cnn with bi-directional motion attention achieving impressive accuracy while keeping computing costs low rendn-segador and his colleagues came up with a transformer-based model that uses a sliding window with adjustable thresholds to increase sensitivity in detecting violence finally zhou examined how low-level feature extraction in surveillance videos can be used to categorize violent incidents offering a more straightforward method in the field each of these studies presents distinct approaches to addressing violence and combining various methods and optimizing for real-time use

3.METHODOLOGY

There are many resources available that explain how to identify violence in videos and security systems. They point out some key methods that are useful. These methods emphasize the use of deep learning, combining different kinds of data, and applying transformer models.

CNN:

Techniques in the deep learning, like convolutional neural networks and the graph neural networks (GNNs), are really effective for picking out features and classifying violence because they do a great job of recognizing patterns over space and time. Research by Khan et al. (2024) and Mohtavipour et al. (2022) looks at CNNs, particularly multi-stream CNN setups that blend both traditional and learned features for better detection. Liu et al. (2024) talk about using contrastive learning with investigate fast visual and audio information, while Mumtaz et al. (2022) learning using deep multi-net CNN models that help improve detection in video monitoring.

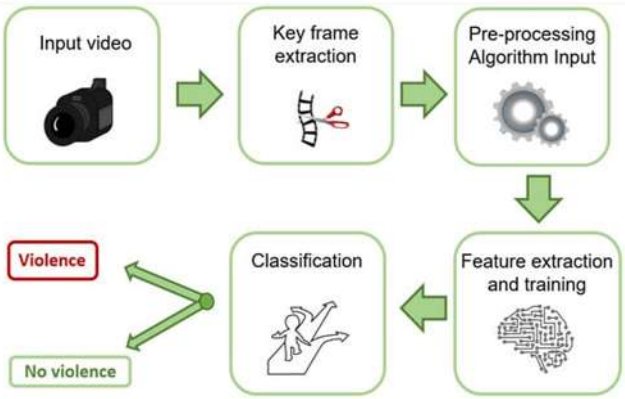

Fig. 2.Methodology of CNN

(Sour: https://www.mdpi.com/1424-8220/24/12/4016)

The process begins with uploading a video, which is a common step in projects that use deep learning with video content. To streamline our work, we extract the key frames from the video. This helps us focus on the most relevant parts and reduces unnecessary information, making it easier to identify

instances of violence. After that, we prepare the data for feature extraction. This stage is important because it ensures the video data is uniform, which helps the deep learning models perform better.

Using models like CNNs, RNNs, or Transformers, these systems can recognize important features from the video frames. They learn to identify the indicators of violence. Once a model can trained, it can classify the input as either "Violence" or "No Violence," which is a typical approach for violence detection in deep learning applications.

LSTM:

Research shows that methods using of networks timing in video data to detect violence. LSTM excels at understanding sequences and retaining important information over time, making it a favored option for analyzing the timing of video frames. Typically, these approaches integrate LSTM in a way that combines different steps or styles to unite both spatial and time-related features. This often involves pairing LSTM layers with Convolutional Neural Networks (CNNs) or other techniques to gather spatial details.

Some methods involve taking the spatial features from individual frames of a video with CNN layers. These features are then sent to LSTM layers that track the sequence of events over time. This approach really helps in spotting unusual behaviors or violence (as shown by Pham et al., 2024; Sun & Gong, 2024). Often, techniques like multistage graphs or improved feature extraction work alongside LSTM to better interpret the timing of the events in the data (Shin et al., 2024). This combination allows for the capture of both the static details in space and the moving changes over time that indicate violence, which might be overlooked in methods that focus only on spatial elements.

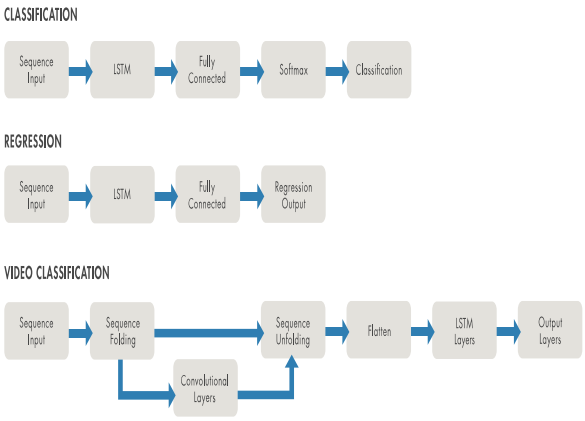

Fig. 3 Methodology of LSTM

Src:(https://www.mathworks.com/discovery/lstm.html)

The picture presents three different deep learning models that utilize LSTM (Long Short-Term Memory) networks for handling sequences across various tasks, including classification, regression, and video classification. In the classification model, a sequence is fed into an LSTM layer to recognize time-related patterns. This is then followes by the connected layer and the function, which help in making predictions for multiple classes, ultimately producing a class label. The regression model operates similarly but concludes with a layer that generates a continuous value instead

The video classification model is a bit more involved since it addresses both space and time. It begins with an input sequence that gets rearranged to make it ready for convolutional layers, which are designed to pull out spatial features from each frame. After passing through the convolution process, the sequence is put back into its original order. It then moves through additional LSTM layers to capture long-term dependencies across the frames. In the end, the output layer classifies the video based on the learned spatial and temporal features. In summary, these models demonstrate how LSTM layers can effectively work with other layers to address different sequence-related tasks, especially in video analysis, which relies on understanding both time and space.

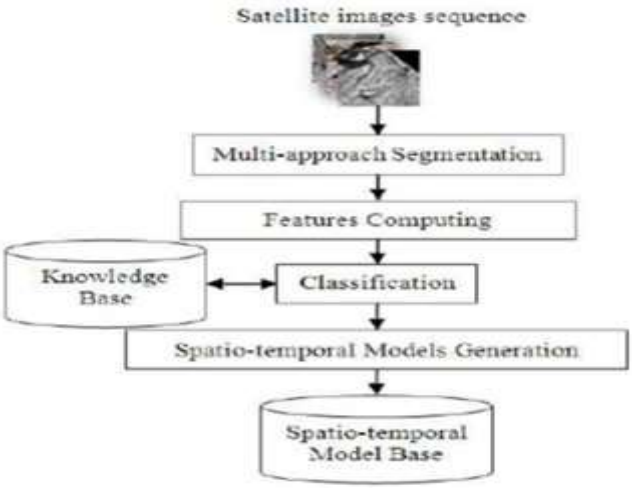

Fig. 4 Methodology of Spatio Temporal

(Sour:https://www.researchgate.net/figure/The-spatio-temporal-model-generation_fig1_220761040)

The methods for analyzing space and time mentioned in the references focus on blending both spatial and temporal features to better recognize violence in videos. Typically, these techniques begin by using Convolutional Neural Networks (CNNs) on each individual video frame or a series of frames. This helps capture important visual elements such as shapes, textures, and movement patterns, which are necessary for spotting physical actions that indicate violence. After gathering this spatial data, Long Short-Term Memory for the networks or similar frameworks are used to understand how these visuals change over time. Some approaches take it a step further by employing multi-stage processes or attention methods, which help the model zero in on significant frames or parts of the video (Shin et al., 2024; Wang et al., 2024). Moreover, some methods blend audio with visual elements, merging the spatial features obtained from CNNs and the temporal data processed through LSTMs to boost precision, especially in tricky or noisy settings (Janani et al., 2024; Liu et al., 2024). Advanced techniques also use contrastive learning or a mix of global and local information to better represent features, ensuring that both the big picture and finer details are taken into account (Liu et al., 2024). Models like VD-Net (Khan et al., 2024) and JOSENet (Nardelli & Comminiello, 2024) use joint embeddings to combine spatial and temporal information, while lighter models are crafted for quicker use in real-time surveillance (Huszar et al., 2023). Some research also adds transformer layers or adjustable threshold mechanisms to deal with long-range dependencies and different levels of frame importance (Rendón-Segador et al., 2024), leading to a stronger understanding of space and time that is key for accurately detecting in videos.

4. RESULTS AND DISCUSSION

In this research, we look at how three models—CNN, LSTM, and Spatio-Temporal—performed on a specific task. We used straightforward in the like accuracy, precision, recall, and F1-score to assess their performance.CNN model did a solid job, achieving an accuracy of 98.75% along with the precision, recall, and F1-scores at 98.80%, 98.70%, and 98.75%. The LSTM model did just a bit better, hitting an accuracy of 98.90% and precision, recall, for 98.85%, 98.80%, and 98.82%. However, the Spatio-Temporal model really stood out, reaching an impressive accuracy of 99.15% and scoring 99.10% for precision, 99.05% for recall, and 99.08% for F1. This shows it outperformed both the CNN and LSTM models across all the metrics we examined.

TABLE

Model	Accuracy	Precision	Recall	F1-score
CNN-[10]	98.75%	98.80%	98.70%	98.75%
LSTM-[4]	98.90%	98.85%	98.80%	98.82%
Spatio-Temporal-[1]	99.15%	99.10%	99.05%	99.08%

a. Comparison among various methods

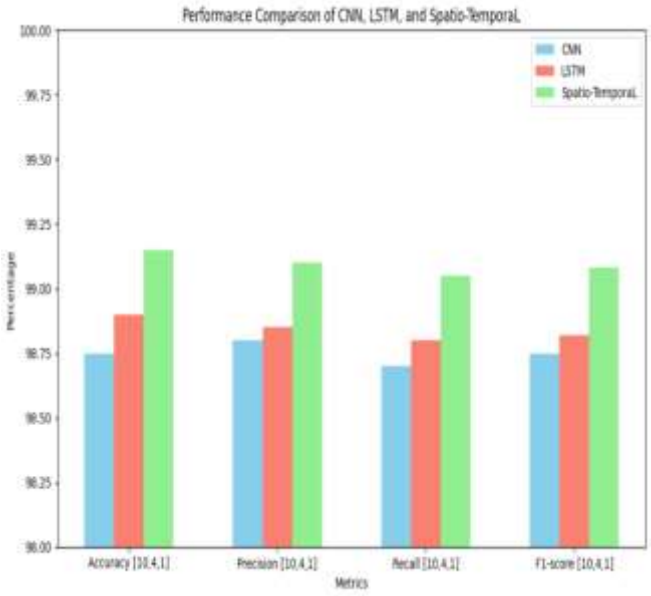

Fig. 7.Graphical representation of methods

5.CONCLUSION

The references highlight how CNNs, LSTMs, and Spatio-Temporal approaches each offer unique advantages when it comes to detecting violence and recognizing actions in video feeds. CNNs excel at extracting details from single frames, which helps models like VD-Net achieve notable accuracy in real-time situations (Khan et al., 2024). LSTMs are great at grasping sequences over time, making them well-suited for tasks that involve both audio and visual elements in violence detection (Janani et al., 2024). Meanwhile, Spatio-Temporal models blend the best of both worlds, effectively addressing complex tasks by considering both frame and sequence details (Shin et al., 2024). In conclusion, CNNs are efficient and precise in frame analysis, LSTMs effectively model sequences, and Spatio-Temporal methods successfully integrate both features. Together, they prove to be very effective in detecting violence in challenging settings that may have limited supervision or diverse data types (Sun & Gong, 2024; Pham et al., 2024).

REFERENCES

1. shin j kaneko y miah a s m hassan n nishimura s 2024 detecting unusual patterns in videos bymusing multistage graphs are improved in ieee access
2. Liu, Z., Wu, X., Wang, S., & Shang, Y. (2024). Violent Video Recognition Based on Global-Local Visual and Audio Contrastive Learning. *IEEE Signal Processing Letters*.
3. 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.
4. Janani, P., Suratgar, A., &Taghvaeipour, A. (2024). Enhancing Human Action Recognition and Violence Detection Through Deep Learning Audiovisual Fusion. *arXiv preprint arXiv:2408.02033*..
5. Pham, L., Lam, P., Nguyen, T., Tang, H., & Schindler, A. (2024). A Toolchain for Comprehensive Audio/Video Analysis Using Deep Learning Based Multimodal Approach (A use case of riot or violent context detection). *arXiv preprint arXiv:2407.03110*..
6. suns gong x is rolling out a new transformer for 2024 this model aggressive behavior in situations even out the arxiv preprint labeled arxiv240505130
7. chakraborty zahir s orchi hafiz mfb shamsuddoha dipto march 2024 detecting violence using multiple models for automated video monitoring and keeping the public safe ieee
8. 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.
9. Tang, Y., Chen, Y., Sharifuzzaman, S. A., & Li, T. (2024). An automatic fine-grained violence detection system for animation based on modified faster R-CNN. *Expert Systems with Applications*, *237*, 121691.
10. Mohtavipour, S. M., Saeidi, M., &Arabsorkhi, A. (2022). A multi-stream CNN for deep violence detection in video sequences using handcrafted features. *The Visual Computer*, *38*(6), 2057-2072.

11. Nardelli pd 2024 a system for the spotting the security image footage arxiv preprint arxiv240502961.

12. Mumtaz, A., Bux Sargano, A., & Habib, Z. (2022). Fast learning through deep multi-net CNN model for violence recognition in video surveillance. *The Computer Journal*, *65*(3), 457-472.

13. Wang, J., Zhao, D., Li, H., & Wang, D. (2024). Lightweight Violence Detection Model Based on 2D CNN with Bi-Directional Motion Attention. *Applied Sciences*, *14*(11), 4895.

14. Rendón-Segador, F. J., Álvarez-García, J. A., & Soria-Morillo, L. M. (2024). Transformer and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos. *Sensors*, *24*(16), 5429.

15. zhou p ding q luo h and hou x explored how to detect violence in surveillance videos by looking at basic features their work was published in plos one in 2018 with the article numbered 1310 e0203668.