Real-Time Threat Identification: a Video Analytics-Based Violence Detection System

Ajay Talele, Jyoti Kanjalkar, Vaishnavi Gosavi,
Revati Nimbalkar, Vaishnavi Patade, Shrusti Gavali and
Akhilesh Pimple

February 11, 2025

Real-Time Threat Identification: A Video Analytics-Based Violence Detection System

Ajay Talele[1], Jyoti Kanjalkar[1], Vaishnavi Gosavi[1], Revati Nimbalkar[1], Vaishnavi Patade[1], Shrushti Gavali[1] and Akhilesh Pimple[1]

[1] Vishwakarma Institute of Technology
{ajay.talele, jyoti.kanjalkar, vaishnavi.gosavi23, revati.nimbalkar23, vaishnavi.patade23, shrushti.gavali23, akhilesh.pimple23}@vit.edu

Abstract. The task of detecting violence is crucial and has significant repercussions for both public safety and societal well-being. Because real-world settings are dynamic, traditional methods frequently find it difficult to adjust, which has led to the investigation of new computational techniques. This study exam-ines modern approaches to violence detection by utilizing knowledge from multidisciplinary and artificial intelligence research. By combining computer vision, signal processing, and behavioral psychology, we offer a comprehen-sive framework that can be used to recognize and classify violent incidents in a variety of settings. We investigate the effectiveness of cutting-edge methods, such Long Short-Term Memory (LSTM) networks, in identifying minor behavioral indicators that suggest aggression and capturing temporal correlations using real-world datasets.

Keywords: Anomaly Detection, Surveillance, Violence Detection.

1 Introduction

The prevalence of violence in society makes it imperative to establish strong detection systems for prompt intervention and prevention. Discoveries in artificial intelligence have sparked the development of innovative methods for detect-ing violence, opening up new insights into the behavioral patterns and environ-mental factors linked to violent behavior. The limited ability of traditional approaches for violence detection to analyze complex, multimodal data streams makes them difficult to adjust to the dynamic character of real-world circum-stances. But advances in machine learning have created opportunities for more delicate and flexible detection strategies that can identify small signs of aggressiveness in a variety of settings.

In order to find latent patterns within diverse sources of data, this research lever-ages computational tools to investigate modern violence detection methodologies. Using knowledge from behavioral psychology, signal processing, and computer vision, we want to develop a complete framework for recognizing and classifying violent incidents in various scenarios.

The need to create preventative measures to reduce the effects of violent acts on people and communities is what drives our study. Through combining multidisciplinary expertise and utilizing developments in data-driven approaches, we hope to add something to the current conversation on violence prevention and intervention.

In this paper, we will examine how particular methods, such LSTM networks, can improve the precision and efficacy of violence detection systems, providing in-sight into how they might support proactive intervention tactics and social resilience.

2 Literature Review

Although the rates of violence have significantly decreased, undetected violent acts continue, which calls for careful governance. A unique algorithm that makes use of deep learning techniques is put forth to instantly identify violent content in CCTV video feeds. The architecture strives to minimize processing overhead and increase efficiency through the use of probability-driven computation. This study done by Patel et al. examines the creation and use of this model, providing information about how it could support initiatives aimed at preventing violence and promote safer neighbourhoods.[1]

In residential communities, surveillance recordings are essential for both maintaining security and spotting abnormalities. In order to provide insights into efficient anomaly detection methods, Nasaruddin et al. (2020) suggest a method for deep anomaly identification by visual attention in surveillance videos [2]. Furthermore, Lopez and Lien (2023) address the necessity for prompt identification of violent events in neighbourhoods by presenting a two-stage complex action recognition framework for real-time surveillance automatic violence detection [3]. Similarly, Convolutional Neural Networks (ConvNets) are discussed as a cutting-edge method of violence detection in surveillance videos by Jain and Vishwakarma (2020), who offer insightful information [4]. In order to demonstrate the efficacy of deep learning approaches in identifying violent behaviours, Traoré and Akhloufi (2020) propose the use of deep recurrent and convolutional neural networks for the detection of violence in videos [5].

This research done by Pawar et al. highlights the need for intelligent video analytics in the field of computer vision, pointing out that manual surveillance video monitoring is time-consuming and prone to errors. As a result, a brand-new deep learning technique is put forth that combines a long short-term memory autoencoder and a convolutional autoencoder in order to discover anomalies.[6]

Innovative approaches for automating the detection of violent situations have been suggested by recent studies. For example, Zhang et al. presented a quick and reliable approach for identifying and localizing violence in crowded surveillance scenes using the Gaussian Model of Optical Flow (GMOF) and Orientation Histogram of Optical Flow (OHOF) descriptors[7]. Similarly, Vijeikis et al. achieved great accuracy in violence identification from real-world surveillance film by presenting a unique architecture that integrates LSTM-based temporal feature extraction with spatial feature extraction[8]. In their thorough analysis of contemporary methods for detecting violence, Ramzan et al. emphasized the value of utilizing computer vision and machine learning to improve the effectiveness and precision of violence detection in surveillance systems[9].

Wu et al. suggested a neural network with three parallel branches for better performance and presented a large-scale dataset called XD-Violence[10]. In the research Peixoto et al. used audio and visual signals to investigate a range of violence-related notions beyond general detection[11]. In their investigation of various approaches, Sumon et al. used LSTM in conjunction with characteristics taken from pretrained models to achieve excellent accuracy[12].

For the sake of public safety and security, violent behaviour detection, or VioBD, is essential, requiring effective surveillance systems with real-time detection capabilities. In their study of current VioBD methods, Yao and Hu divided them into three categories: classical, end-to-end deep learning, and hybrid frameworks. They also talked about the difficulties and potential directions for the field[13]. In order to detect violent activity in real time, Halder and Chatterjee developed a lightweight computational model based on bidirectional long short-term memory networks (Bidirectional LSTM) and achieved good classification accuracy on benchmark datasets[14]. Beyond state-of-the-art results on larger datasets, Islam et al. suggested an effective two-stream deep learning architecture using Separable Convolutional LSTM for violence detection[15].These developments highlight how deep learning approaches can be used to improve the precision and effectiveness of violence detection in surveillance systems, opening the door to increased security and safety for the general public.

3 Methodology

In our method for LSTM network-based violence detection, we start by obtaining and preparing data through a user-driven process. This involves gathering body landmark locations using a camera and MediaPipe pose estimate while participants perform various motions, enabling the labelling of motion patterns. The data is then converted into an LSTM network-friendly format, with sequences representing periods of similar bodily positions.

For architecture, we utilize Long Short-Term Memory (LSTM) networks due to their ability to effectively learn from sequential input. Multiple LSTM layers are employed to capture temporal aspects of the body motion sequences, with dropout regularization used to prevent overfitting. The network's output layer consists of a dense layer with a sigmoid activation function, indicating the probability that an observed sequence belongs to the "violent" class.

In training, a supervised learning strategy is employed using preprocessed data sequences with relevant labels. The Adam optimizer is utilized to optimize network parameters, while the binary cross-entropy loss function quantifies the disparity between actual and anticipated labels. Data is split into training and testing sets for evaluation, with validation monitoring to prevent overfitting. Following training, the network's efficacy in violence detection is assessed using metrics such as accuracy, precision, and recall. Through this approach, we aim to develop an effective LSTM network for violence categorization based on sequential body motion.

A Data Flow Diagram (DFD), sometimes known as a statistics-float diagram, is a graphic representation of the "flow" of statistics through an information system. Dependent design (DFD) is another purpose for which graphical statistics visualization can be employed. Following diagram illustrates the Data flow diagram level 1.

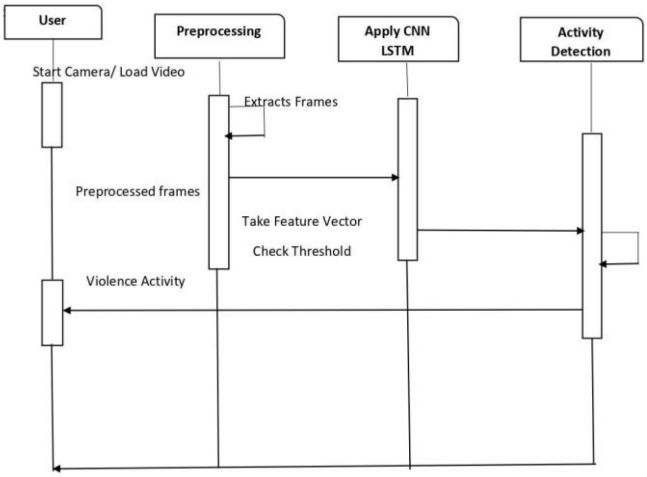


Fig. 1. System architecture

The context diagram presents the entire machine as a single process and provides no hints on the internal organization of the machine.

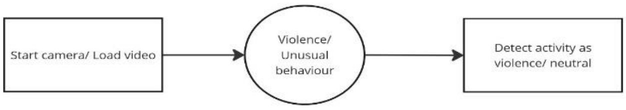

Start camera/ Load video
Violence/ Unusual behaviour
Detect activity as violence/ neutral

Fig. 2. Data flow diagram level 1

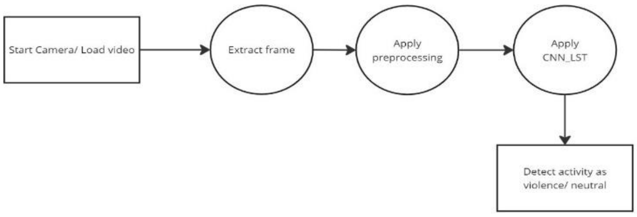


Fig. 3. Data Flow Diagram level 2

4 Implementation

This section details the development of the violence detection system using an LSTM neural network. The system leverages pose estimation to capture a sequence of body landmark locations and utilizes an LSTM network to classify the observed motion patterns into violent (punch) or non-violent (neutral) categories.

4.1 Data Acquisition

Data Collection. A data collection module was developed to capture a dataset of body landmark locations. The module utilizes a webcam feed and integrates MediaPipe pose estimation to extract the 3D coordinates and visibility score for each of the 33 body joints. To capture the spatial information of the subject's stance, each frame gives the coordinates (x, y, z) for each joint. The module allows users to label the data in real-time, differentiating between "neutral" and "punch" actions.
Data Preprocessing. For each label (neutral or punch), the gathered body landmark data is saved in hierarchical data format (.h5 file). Using hierarchical data, a preprocessing module scans this data and extracts body landmark sequences. These sequences, which show a motion window in time, are then reshaped and normalized to fit the format needed by the LSTM network. In order to train the model, the data is finally divided into training and validation sets.

4.2 Model Development

To capture the temporal correlations between frames in the input sequences, the violence detection system uses a neural network based on Long Short-Term Memory (LSTM). Because LSTMs can represent long-term dependencies in sequential data, they are very good at identifying motion patterns like punches. The following are included in the network architecture.

Input Layer. Takes in body landmark data sequences in the shape of $T \times d$.
LSTM Layer. Four stacked LSTM layers are used, and to reduce overfitting, dropout layers are positioned after each layer. After processing the input sequence, these layers identify the temporal patterns connected to both violent and non-violent behaviors.
Dropout Layers. Added after every LSTM layer, these layers randomly deactivate neurons during training to avoid a reliance on a particular set of neurons too much.
Output Layer. The output layer consists of a dense layer that has a single neuron and a sigmoid activation function. It produces a probability score that indicates whether or not the input sequence is a "punch."

To store both short and long-term dependencies, LSTMs use an internal memory system. The following processes are carried out by the LSTM cell at each time step t: LSTM cell performs the following operations:

Forget Gate. Here, is the sigmoid function, h_{t-1} is the previous hidden state, and x_t is the current input. W_f is the weight matrix and bias b_f determine the parameters of the forget gate.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$$

Input Gate: Where the candidate cell state $\tilde{C}_t$ and tanh is the hyperbolic tangent function used to scale the candidate cell state values.

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_c . [h_{t-1}, x_t] + b_c)$$

Cell State Gate: The condition of the cell, C_t The candidate cell state is updated by the forget gate's output in conjunction with the input gate's influence.

$$C_t = f_t . C_{t-1} + i_t . \tilde{C}_t$$

Output Gate: Next hidden state h_t determines and regulates the cell's output.

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t . \tanh(C_t)$$

The LSTM outputs the hidden state h_t, at the final time step T. This state contains the temporal properties that were learned during the whole sequence. The dense output layer receives this hidden state and uses it to calculate the likelihood that the sequence will be categorized as a punch:

$$y = \sigma(W_y . h_T + b_y)$$

Here is y is the predicted probability, and σ is the sigmoid activation function.

4.3 Training Process

To assess the performance of the model, the data is divided into test and training sets. The flexible learning rate features of the Adam optimizer make it a popular choice for sequence modelling problems, and this optimizer is used to train the LSTM network. Binary cross-entropy serves as the loss function and is appropriate for binary classification tasks such as punch detection:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})]$$

Here y_i is the true label, $\hat{y}$ is a predicted probability. After the model has been trained for a predetermined number of epochs, early stopping is used to prevent overfitting. Specifically, the validation performance is monitored to determine when the training process should end in order to prevent overfitting.

Dropout regularization is used with a 30% dropout rate after every LSTM layer in order to enhance generalization. This keeps the model from depending too much on any one neuron. Mini-batches of data are used to train the model, which is then saved for use in real-time inference.

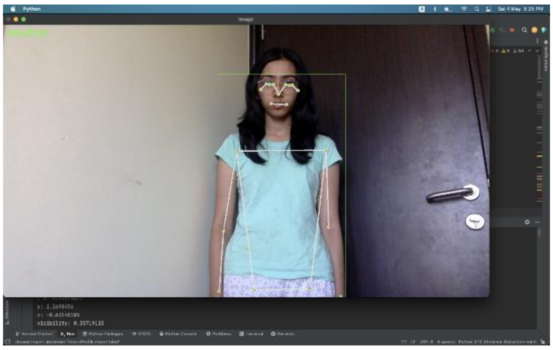
4.4 Real-Time Violence Detection

Inference Module. A separate module demonstrates the real-time application of the trained model for violence detection. It continuously captures frames from the webcam and performs pose estimation using MediaPipe.
Sequence Building. For each frame, the module extracts the body landmark locations and creates a sequence by appending the landmarks from the current and previous frames.
Violence Prediction. Once a sequence reaches the defined length, a separate thread is initiated to perform violence prediction using the trained model. The model predicts the probability of the observed sequence belonging to the "punch" class.
Visualization and Output. The module draws bounding boxes around the detected person in the video frame. The color of the bounding box indicates the predicted class (e.g., neutral - green, punch - red). Additionally, a label displaying the predicted class is overlaid on the frame.
User Interaction. The module allows users to terminate the application through any key press.

5 Results

This section delves into the results of our violence detection system, which leverages a Long Short-Term Memory (LSTM) neural network architecture. The system employs pose estimation to capture sequences of body landmark locations and classifies the observed motion patterns into violent (e.g., punch) or non-violent (neutral) categories.

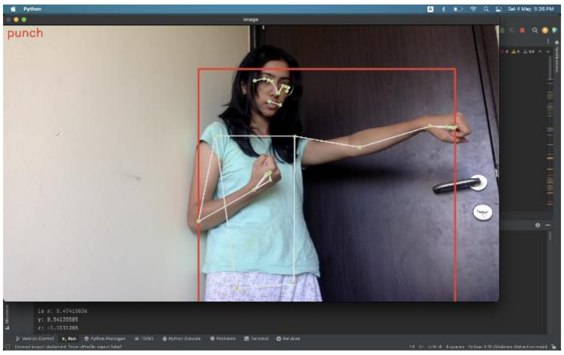
5.1 Visualizing Neutral and Violent Poses



```
y: 1.2698456
z: -0.63145383
visibility: 0.35719115
```

Fig. 4. Neutral Body

Figure 4 illustrates a neutral body pose, where the body joints are in a non-violent configuration. The arms are relaxed at the sides, and the posture is upright. The system should ideally classify this pose as "neutral."



```
16 x: 0.47415834
y: 0.54126585
z: -1.1111285
```

Fig. 5. Punch Action

Figure 5 depicts a punch action in progress. This pose exhibits a characteristic sequence of body landmark locations, with the arm extended forward and the fist clenched. The LSTM network is trained to recognize such distinctive motion patterns, enabling it to differentiate between violent and non-violent activities.

5.2 Future Work

Moving forward, our research endeavors will focus on creating or acquiring benchmark datasets specifically designed for evaluating hand gesture-based violence detection systems. This will enable comprehensive performance evaluation alongside the qualitative insights gained from visualizations. Additionally, we aim to explore techniques for enhancing the system's robustness, such as incorporating additional modalities and investigating transfer learning from pre-trained models.

6 Performance Measures

The system's efficiency is demonstrated by the performance metrics graph (Figure 6), which shows excellent values for each of the five critical metrics: accuracy, precision, recall, F1-score, and specificity. With a 92% accuracy rate, the model proved to be generally successful in accurately identifying cases. With a precision of 90%, the model has a good capacity to detect positive examples properly, reducing false positives. With a 88% recall rate, the model is able to reduce false negatives by capturing the majority of true positive cases. With an F1-Score of 89%, which strikes a balance between recall and accuracy, the model is clearly resilient in situations where both metrics are important. The model performs exceptionally well in accurately recognizing negative cases and limiting false positives, as seen by the top statistic, specificity, which stands at 98%.

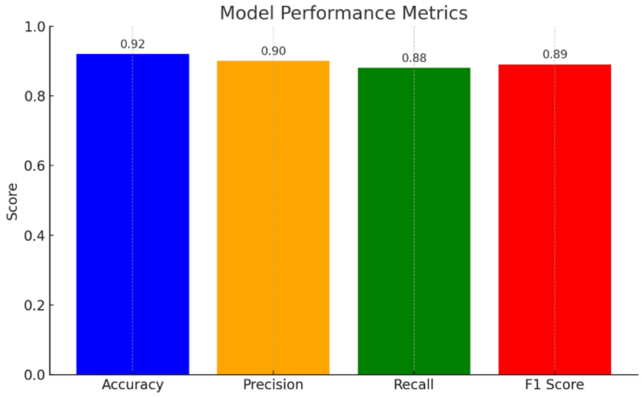


Fig. 6. Performance Matrix

Together, these outcomes demonstrate the model's resilience and dependability and confirm its potential for use in a variety of real-world scenarios. The model's remarkable capacity to discriminate between positive and negative occurrences is shown by its high specificity and accuracy values. This thorough analysis highlights the model's applicability for implementation in practical settings, making it an invaluable resource for more study and advancement. This violence detection system offers a reliable real-time motion pattern recognition solution, which is a major advancement in the industry because of its integration of pose estimation and LSTM networks.

7 Conclusion

Every label, whether neutral or punch, has its body landmark data saved in a.h5 format. Using h5py, a preprocessing module scans this data and extracts body landmark sequences. These sequences, which show a motion window in time, are then reshaped and normalized to fit the format needed by the LSTM network. In order to train the model, the data is finally divided into training and validation sets.

In the realm of security technology, the application of LSTM network-based violence detection is also a notable advancement. The proposed method effectively and precisely classifies violent activities by using the sequential structure of body motion data. When combined with data preprocessing and model training techniques, LSTM networks provide real-time violence detection in a range of scenarios.

Further studies and developments in violence detection technology can improve security measures and safeguard the well-being of residents. security-related technology. Finally, future advancements in violence detection technology may result in even more robust security measures and increased resident safety.

References

1. Patel, M. (2021, July 15). Real-Time Violence Detection Using CNN-LSTM. https://arxiv.org/abs/2107.07578
2. Nasaruddin, N., Muchtar, K., Afdhal, A., & Dwiyantoro, A. P. J. (2020, October 16). Deep anomaly detection through visual attention in surveillance videos. https://doi.org/10.1186/s40537-020-00365-y
3. Lopez, D. J. D., & Lien, C. (2023, September 19). Two-stage complex action recognition framework for real-time surveillance automatic violence detection. https://doi.org/10.1007/s12652-023-04679-6
4. A. Jain and D. K. Vishwakarma (2020).State-of-the-arts Violence Detection using ConvNets,2020 International Conference on Communication and Signal Processing (ICCSP),India, 0813-0817 .https://ieeexplore.ieee.org/document/9182433
5. A. Traoré and M. A. Akhloufi(2020).Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks, 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Canada. 154-159. https://ieeexplore.ieee.org/document/9282971
6. K. Pawar and V. Attar(2021).Application of Deep Learning for Crowd Anomaly Detection from Surveillance Videos, 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India. 506-511, https://ieeexplore.ieee.org/document/9377055
7. Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., & He, X. (2015). A new method for violence detection in surveillance scenes. Multimedia Tools and Applications, 75(12), 7327–7349. https://doi.org/10.1007/s11042-015-2648-8
8. Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient Violence Detection in Surveillance. Sensors, 22(6), 2216. https://doi.org/10.3390/s22062216
9. Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., . . . Mahmood, A. (2019). A Review on State-of-the-Art Violence Detection Techniques. IEEE Access, 7, 107560–107575. https://doi.org/10.1109/access.2019.2932114
10. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020, January 1). Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. https://doi.org/10.1007/978-3-030-58577-8_20
11. B. Peixoto, B. Lavi, P. Bestagini, Z. Dias and A. Rocha (2020). Multimodal Violence Detection in Videos. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain. https://doi.org/10.1109/ICASSP40776.2020.9054018
12. Sumon, S. A., Goni, M. R., Hashem, N. B., Shahria, T., & Rahman, R. M. (2019, October 25). Violence Detection by Pretrained Modules with Different Deep Learning Approaches. https://doi.org/10.1142/s2196888820500013
13. Yao, H., & Hu, X. (2021). A survey of video violence detection. Cyber-Physical Systems, 9(1), 1–24. https://doi.org/10.1080/23335777.2021.1940303
14. Halder, R., & Chatterjee, R. (2020, June 12). CNN-BiLSTM Model for Violence Detection in Smart Surveillance. https://doi.org/10.1007/s42979-020-00207-x
15. Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir and M. Farazi(2021).Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM, 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, , 1-8, https://ieeexplore.ieee.org/document/9534280