Violence Detection in Videos Using Discriminant Frame Extraction and Convolutional Long Short-Term Memory

Venkatesh Akula * and Ilaiah Kavati

Department of Computer Science and Engineering, National Institute of Technology, Warangal, Telangana, India
Email: avenkatesh@student.nitw.ac.in (V.A.); ilaiahkavati@nitw.ac.in (I.K.)
*Corresponding author

***Abstract*—Violent activity detection in videos is a challenging task in computer vision due to the complex and diverse motion patterns present among human subjects in real-world environments. Automated surveillance systems are necessary in sensitive monitoring areas such as smart cities, educational institutions, sports grounds, hospitals, public gatherings, and other critical surveillance domains. This paper aims to investigate the effective combination of hand-crafted and deep learning features within a hierarchical framework to enhance the accuracy and efficiency of violent action detection in videos. In the first stage, the You Only Look Once (YOLO) object detection model is employed for person detection in video frames, thereby discarding frames without humans. A precise cropping operation is then performed on the resultant frames around the maximum overlapped human intersection area (i.e., the Region of Interest (ROI)). In the second stage, interest points from the cropped region are generated using optical flow. Histogram of Oriented Gradients (HOG) is then applied to extract features from the image data. These feature maps are forwarded to a customized Convolutional Long Short-Term Memory (ConvLSTM) deep neural network for efficient training. The model is evaluated on three benchmark datasets: Action Movies, hockey fights, and violent flows. The experimental results demonstrate that our method performs better than existing state-of-the-art approaches.**

***Keywords*—violence detection, key frames, You Only Look Once (YOLO), Histogram of Oriented Gradients (HOG), Convolutional Long Short-Term Memory (ConvLSTM)**

I. Introduction

In recent years, there has been a rapid increase in the generation and use of video content for surveillance purposes. The widespread deployment of Closed-Circuit Television (CCTV) cameras in sensitive areas has significantly contributed to this growth. As a result, there has been a substantial increase in the demand for video data in security-related applications. Surveillance footage is carefully analyzed to monitor human activities, making video data analysis a critical area of focus. As a result, it has attracted significant attention from researchers in the computer vision field. In surveillance domains, Human Activity Recognition (HAR) [1, 2], human object tracking, and human behavior analysis are key functionalities. Detecting violent actions is crucial for enabling immediate responses to aggressive behavior. However, due to the vast amount of video content generated and the infrequent occurrence of violent incidents, it is challenging for human observers to monitor surveillance footage continuously. Manual monitoring also requires substantial human resources, making it impractical. Therefore, there is a strong need to develop efficient automated surveillance systems.

Violence Activity Detection (VAD) has received comparatively less attention than normal action recognition in HAR, primarily due to the following challenges [3]:

(1) The presence of intricate and unpredictable motion patterns among human subjects.
(2) Variations in camera angles further increases the complexity, as the same action may appear different from multiple perspectives.
(3) The difficulty in distinguishing between normal and violent behaviors, as normal actions such as hugging or handshaking may visually resemble aggressive interactions.

Furthermore, processing all video frames is challenging, as successive frames often contain redundant information. Therefore, it is essential to identify significant frames during the preprocessing stage for further analysis. Since motion is key parameter for violence detection, focusing on motion regions in keyframes helps extract meaningful features for classification. A hybrid method combining handcrafted spatial and deep learning temporal features improves spatio-temporal extraction [4].

To address these challenges, we propose a hierarchical violence detection approach that integrates a robust framework combining hand-crafted features with deep learning techniques. As illustrated in Fig. 1, we first employ the You Only Look Once (YOLO) v10 object detection model to identify humans in video frames [5]. Since violent actions typically involve close human

Manuscript received April 9, 2025; revised May 14, 2025; accepted July 16, 2025; published February 5, 2026.

interactions, frames are filtered by retaining those with the highest overlap among detected individuals to reduce processing overhead. Subsequently, optical flow is computed between successive frames to highlight high-motion regions. Histogram of Oriented Gradients (HOG) features are then extracted from these key motion areas, capturing essential motion patterns. Finally, the extracted HOG features are fed into a deep neural network for effective video classification. The datasets used in the experimental analysis are hockey fight, action movies, and violent flows [6, 7].

The key contributions of this paper are:

- We propose to utilize YOLOv10 model to detect human objects in video frames and significantly reduce processing overhead by discarding frames that do not contain any humans.
- We introduce a strategy to discard irrelevant frames that lack significant interaction between human objects, thereby focusing only on frames with potential conflicts among detected individuals to improve detection accuracy.
- We leverage motion information obtained from optical flow between selected frames to identify regions of interest and extract features using HOG.
- We employ a customized Convolutional Long Short-Term Memory (ConvLSTM) model trained on the extracted significant features, thereby enhancing the overall model performance.

The paper is structured as follows: Section II reviews related work on violence detection using both hand-crafted and deep learning features. Section III describes the proposed architecture in detail. Section IV presents a performance comparison with existing approaches, experimental results, and a summary of the datasets used. Finally, Section V concludes the paper and outlines directions for future work.

II. Literature Review

In recent years, numerous methods have been proposed to address the challenges associated with detecting violent activities in videos. This section provides a brief review of these approaches, covering both hand-crafted feature-based and deep learning-based methods.

A. Handcrafted Features

In handcrafted feature-based methods, key regions within frames are extracted using motion information. An extreme acceleration pattern-based method is proposed for video classification utilizing motion features [8]. Hassner *et al*. [7] introduced a descriptor called Violent Flow (ViF), which is based on the magnitude of optical flow. Gao *et al*. [9] improved ViF by incorporating orientation information along with the magnitude of optical flow, resulting in a descriptor called Oriented Violent Flows (OViF). Mahmoodi and Salajeghe [10] proposed the Histogram of Optical flow Magnitude and Orientation (HOMO) descriptor, which leverages both magnitude and orientation components of optical flow. This descriptor performs well on the hockey fight dataset. Chen and Hauptmann [11] proposed the motion SIFT descriptor, which combines HOG and optical flow information to represent spatiotemporal features in videos. Febin *et al*. [12] presented a motion boundary SIFT feature-based violence detection method, where motion boundary information is used to detect violent actions. A novel descriptor called Distribution of Magnitude and Orientation of Local Interest Frame (DiMOLIF), generates the orientation and magnitude of optical flow for video frames [13].

B. Deep Features

The advancement of deep learning methods, which inherently learn discriminative features, has led to the development of more effective violence detection models that outperform traditional handcrafted approaches. Serrano *et al*. [14] proposed a method combining handcrafted features with a 2D CNN to enhance detection accuracy. Keçeli and Kaya [15] introduced a transfer learning-based approach that constructs 2D motion templates using magnitude and orientation information, which are then processed by a pre-trained network. Roman and Chavez [16] proposed a rank pooling-based technique to generate a single representative image summarizing the video sequence for classification. A keyframe-based strategy was introduced along with a modified 3D CNN [17]. This approach converts RGB frames into grayscale and uses the gray centroid to evaluate the similarity between frames. Keyframes with a relative distance above a set threshold are selected and passed as input to the 3D CNN for classification.

Cobo and SanMiguel [18] presented a violence detection method that leverages dynamic temporal changes in human interactions and human posture estimation. Akula and Kavati [4] introduced a technique that utilizes key frames from a video and employs a densenet architecture for efficient feature learning. Rendon-Segador *et al*. [19] proposed a method that incorporates a multi-head self-attention layer combined with Bi-Directional Long Short-Term Memory (BD-LSTM) for violence detection. Imran *et al*. [20] introduced a method for violent activity recognition, which involves computing Approximate Dynamic Images (ADI) during the preprocessing stage. Tan *et al*. [21] proposed a fusion network for violence detection, capturing both spatial and temporal features using a 3D CNN and frame interactions. Kavati *et al*. [22] developed a human activity recognition framework based on Space and Channel-Based Squeeze and Excitation blocks (SCBSE) [2]. Asad *et al*. [23] presented a violence detection method that employs a feature fusion technique to combine multi-level features extracted from the top and bottom layers of a convolutional neural network across two consecutive frames. Naik *et al*. [24] proposed a method, which integrates YOLOv4 with deep feature-based real-time tracking to address identity switching in sports videos, achieving high accuracy and real-time performance. Singh *et al*. [25] proposed an end-to-end ViViT-based deep learning framework for detecting violent actions in surveillance videos, integrating data augmentation to enhance performance on small datasets. Li *et al*. [26] proposed ACTION-VST framework that

employs a video swin transformer with keyframe selection and multi-path excitation for effective spatiotemporal violence detection, achieving superior accuracy on RLVS and RWF-2000 datasets.

However, we identified several critical gaps in the existing literature on violence detection:

(1) Processing all frames in a video results in significant computational overhead.
(2) Many existing works rely on C3D models, which are limited in capturing long-range temporal dependencies.
(3) There is a lack of efficient region-based filtering methods to focus on meaningful human interactions.

Moreover, early approaches attempted to detect violence by identifying explicit cues such as flames, blood, gunshots, and explosions. To address the identified gaps in the literature, our model employs two key strategies for effective violence detection:

(1) It emphasizes motion-based cues by analyzing movement patterns between human objects, as violent actions are typically characterized by rapid and aggressive motion;
(2) It reduces computational overhead by selectively processing only the most informative frames, recognizing that violence is rare in surveillance video and processing every frame in a video is redundant and inefficient.

III. Proposed Method

The proposed method is detailed in the following subsections: (i) Keyframe extraction and identification of significant spatiotemporal features from high-motion regions in the selected keyframes; (ii) Feature extraction using the Histogram of Oriented Gradients (HOG); (iii) Video action classification using a customized deep neural network. Fig. 1 shows the detailed architecture of the proposed model.

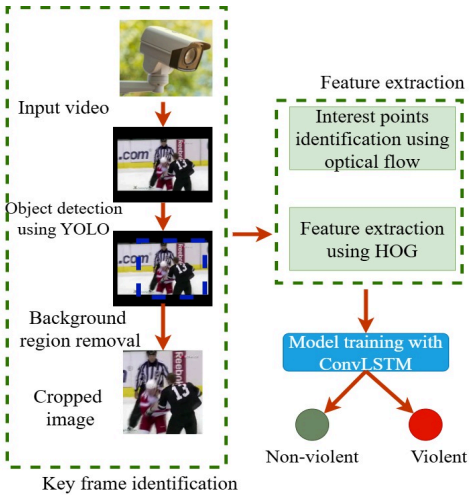

Fig. 1. Proposed model block diagram.

A. Key Frame Identification

Successive frames in a video often contain significant content redundancy, increasing the processing overhead in video analysis models. It is unnecessary to analyze all the frames of a video. Hence, selecting the key frames of the video in the preprocessing stage can significantly improve the model performance. To effectively capture motion patterns, we focus on processing only the keyframes that contain human objects. Human detection within video frames is carried out using the YOLOv10 object detection model [5]. This step enhances model accuracy by eliminating irrelevant frames those without human presence and retaining only the keyframes for further analysis. YOLOv10 precisely draws bounding boxes around detected objects, enabling the isolation of human objects from the background. Within each keyframe, Regions of Interest (ROIs) are identified and assigned confidence scores based on detection reliability. Fig. 2 illustrates sample keyframes extracted from various datasets.

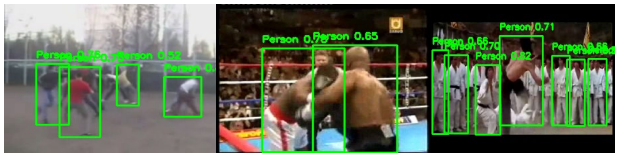

Fig. 2. Keyframes containing detected humans extracted from different datasets.

1) Identifying high-motion regions

The extracted keyframes are further analysed based on the intersection area between human objects in every bounding box detected. Violent actions often involve close interactions such as punching, kicking, and hitting with an object, making regions with minimal human overlap less relevant for processing. In each keyframe, the bounding box with the highest intersection above the preset threshold (th) is retained, while others are discarded to simplify processing and reduce model complexity. The red bounding boxes in Fig. 3 indicate the keyframes containing the maximum intersection regions. This selective approach reduces computational complexity by eliminating irrelevant background regions and focusing only on ROIs. Consequently, this strategy improves the model's efficiency and performance by focusing on key motion regions, which are then used for spatio-temporal feature extraction in the motion analysis stage.

Fig. 3. Maximum intersection regions extracted from keyframes across different datasets.

2) Interest point detection

Detecting interest points with significant optical flow is essential for capturing dynamic motion, particularly in violent scenes with frequent rapid movements [27]. To identify these interest points, optical flow along X and Y directions for all pixels are computed and then separated

into high and low optical flow clusters using K-Means clustering. Pixels belonging to the high optical flow cluster are designated as motion interest points, as they represent localized areas of significant movement. For each group formed by intersecting human bounding boxes, the average motion magnitude of the motion interest points within the group is computed. The intersection group with the highest average motion is then selected for subsequent feature extraction. From the experimental analysis, we observed that the regions in Fig. 4 (highlighted in red) within the keyframes are high-motion areas and are considered as the interest points within a keyframe.

Fig. 4. High-motion regions (shown in red) identified within keyframes of the videos from different datasets.

B. Feature Extraction Using Histogram of Oriented Gradients (HOG)

The high-motion regions identified within the keyframes from the previous stage are utilized for feature extraction. The HOG feature descriptor is employed, which quantifies the distribution of gradient orientations within these localized, high-motion areas (Fig. 5). This approach can effectively capture the appearance and shape of objects within that region. The HOG descriptor is computed by dividing the image into grids, and calculating the magnitude and orientation of each pixel, and then aggregating these orientations into histograms within each cell.

Fig. 5. Extracted HOG features from high-motion frames obtained.

$$g_x = f(x+1, y) - f(x-1, y) \tag{1}$$

$$g_y = f(x, y+1) - f(x, y-1) \tag{2}$$

In above Eqs. (1) and (2), $f(x, y)$ represents the pixel intensity at location (x, y), g_x represents the change in intensity values in the X-direction and g_y represents the change in intensity values in the Y-direction. Eq. (3) represents the gradient magnitude at pixel (x, y), and Eq. (4) represents the gradient orientation.

$$g(x, y) = sqrt(g_x^2 + g_y^2) \tag{3}$$

$$\theta(x, y) = \arctan\left(\frac{g_y}{g_x}\right) \tag{4}$$

C. Network Architecture

The Convolutional Long-Short-Term Memory (ConvLSTM) network is employed in the proposed model to further extract the significant spatio-temporal features. The network is optimized to effectively capture motion patterns and relationships between extracted features. The customized architecture ensures that the model is not only precise, but can also process large volumes of video data, which is critical for real-time applications in surveillance domains. Convolutional layers capture spatial patterns, while Long-Short-Term Memory (LSTM) cells model temporal dynamics among human subjects.

The proposed network architecture (Fig. 6) begins with a ConvLSTM2D layer that captures spatio-temporal patterns using 32 filters with a 3×3 kernel size and ReLU as an activation function. A dropout rate of 0.2 is applied after the convolutional layer to improve generalization. Hence, the overfitting problem reduces. Then, a MaxPooling2D layer is placed to down-sample the spatial dimensions of the generated feature maps. The kernel size utilized in the MaxPooling2D layer is 2×2.

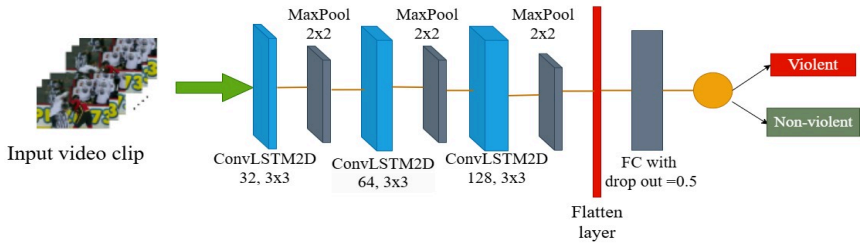

Fig. 6. Internal architecture of the proposed neural network.

The high-level features are extracted in the initial layers. To further capture the effective features from the extracted feature maps in previous layers, the number of filters in the second and third ConvLSTM2D layers are set to 64 and 128, respectively. Then, a MaxPooling2D layer is placed after each ConvLSTM2D layer to reduce the spatial dimensions. The resultant features after the third ConvLSTM2D layer are flattened. Then, a Fully Connected (FC) is placed with the number of neurons set to 128 and a dropout rate of 0.5. Then, the sigmoid function is placed in the output layer to generate the probabilities for each class, i.e., violent and non-violent. The proposed model network architecture is detailed in Table I.

TABLE I. PROPOSED MODEL NETWORK ARCHITECTURE

Layer Type	Architecture	Feature Map Size
Input	[16, 3, 90, 180]	[16, 3, 90, 180]
ConvLSTM2D	32 filters, 3×3 kernel	[16, 32, 90, 180]
Dropout	Rate: 0.2	[16, 32, 90, 180]
MaxPooling2D	2×2	[16, 32, 45, 90]
ConvLSTM2D	64 filters, 3×3 kernel	[16, 64, 45, 90]
MaxPooling2D	2×2	[16, 64, 22, 45]
ConvLSTM2D	128 filters, 3×3 kernel	[16, 128, 22, 45]
MaxPooling2D	2×2	[16, 128, 11, 22]
Flatten	-	[16, 30976]
Dense (FC)	128 units	[128]
Dropout	Rate: 0.5	[128]
Dense	Sigmoid	[1] (violent or not)

IV. EXPERIMENTAL RESULTS

In this section, we analyze the performance of the proposed method through a series of experiments and compare the results with existing approaches.

A. Datasets

In the experimental analysis, the following datasets were used: action movies, hockey fight, and violent flows [6, 7]. Table II outlines the specifications of each dataset.

TABLE II. DETAILS OF THE DIFFERENT DATASETS

Dataset	Samples	Resolution
Action Movies [6]	200	360×250
Hockey Fight [6]	1000	360×288
Violent Flows [7]	246	320×240

Fig. 7. Sample video frames from different datasets.

Each dataset contains an equal number of violent and non-violent videos. Sample frames depicting both types of scenes from different datasets are shown on Fig. 7. The hockey fight dataset is specifically curated for violence detection, with videos captured from hockey games. The action movies dataset comprises videos collected from movies and public gatherings. The violent flows dataset is comparatively more challenging due to its complex, noisy backgrounds and densely crowded scenes.

B. Experimental Setup

All the experiments were conducted using the Google Colab environment, equipped with an L4 GPU and TensorFlow 2.0, to build the model architecture. In experimental analysis, video samples from each data set are divided into three sets: 70% of the data is assigned to the training set, 10% to the validation set, and 20% to the test set. The model was trained for 300 epochs using mini-batch Stochastic Gradient Descent (SGD). A learning rate of 0.0005 was used to ensure stable convergence, while a batch size of 4 was chosen to balance learning efficiency and model performance.

C. Performance Evaluation

Several experiments are conducted during model training for the performance evaluation of the proposed method.

1) Intersection over Union (IoU) between human objects vs. model performance

The proposed model is evaluated by varying the size of the overlapping regions in the detected human bounding boxes during the preprocessing stage (Fig. 8). The performance of the model is assessed across different IoU threshold (th) ranges, specifically 0.85–0.90, 0.90–0.95, and 0.95–0.98, respectively. From the experimental analysis, we can observe that violent action scenes exhibit maximum overlapping regions among human objects. As the IoU area increases, the accuracy of the model (*i.e.,* detecting the violent actions) increases (Fig. 8). Across all datasets, the proposed model achieves the highest accuracy within the IoU threshold range of 0.95–0.98. For the hockey fight dataset, the model accuracy increases significantly as the range increases. As a sports dataset, it naturally exhibits frequent instances of player overlapping. Consequently, the model more accurately detects violent actions when the IoU threshold is high. In contrast, the other two datasets exhibit fewer overlapping regions among humans.

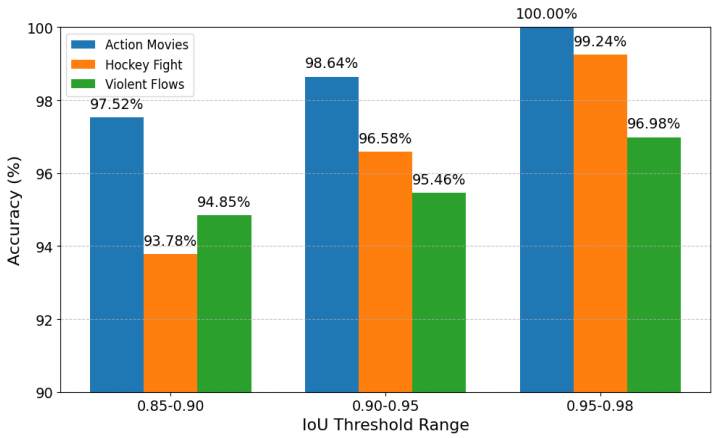

Fig. 8. Model performance vs IoU threshold range.

However, the model demonstrates improved performance when interactions between human objects are high within the frames.

To assess the impact of Intersection over Union (IoU)-based filtering on model performance, additionally we conducted an ablation study by removing IoU component from the processing pipeline. In the modified setup, all person-detected frames were forwarded without assessing spatial interactions with IoU. The exclusion of IoU-based filtering introduced temporal noise and reduced the model's effectiveness in capturing salient motion features, thereby increasing the complexity of spatio-temporal representation learning. From the results shown in Table III, we can observe the critical role of IoU-based filtering in improving violence detection by focusing on regions with significant human interaction. Furthermore, Fig. 8 demonstrates that lower IoU threshold ranges result in reduced model performance.

TABLE III. MODEL PERFORMANCE VS IOU

Dataset	With IoU (%)	Without IoU (%)
Action Movies	100	98.69
Hockey Fight	99.24	97.31
Violent Flows	96.98	95.86

2) Ablation study

In this section, we perform ablation study to investigate the effectiveness of each component in the proposed framework.

a) Effect of keyframe extraction (You Only Look Once (YOLO) detection) on the model performance

Violent activity in a video typically occurs within a specific duration rather than throughout the entire sequence. Moreover, consecutive video frames often exhibit high similarity, resulting in significant temporal redundancy. Therefore, processing all the video frames is both computationally inefficient and unnecessary. To address this, keyframes are selected while redundant frames are discarded, reducing overall processing time and improving model efficiency. In the proposed approach, keyframe selection is performed using YOLO, which identifies frames based on the presence of human objects. Since violent actions primarily occur among humans, focusing on such frames is both meaningful and effective.

TABLE IV. MODEL ACCURACY (%) VS KEYFRAME SELECTION USING YOLO

Dataset	With keyframes identified using YOLO	Without keyframes
Action Movies	100	99.26
Hockey Fight	99.24	97.58
Violent Flows	96.98	94.73

To evaluate this effect, an experiment was conducted to analyze the impact of keyframe selection on the performance of the proposed model. The detailed experimental results are presented in Table IV. A slight decrease in accuracy was observed when the model was evaluated without keyframe selection. The reduced performance may be caused by redundant information in consecutive frames, which introduces noise and reduces the learning of significant spatio-temporal features.

b) Effect of Histogram of Oriented Gradients (HOG) and optical flow features on model performance

To evaluate the impact of temporal motion cues on model performance, we conducted an ablation study to compare HOG-only features with HOG plus optical flow. The results indicate that while HOG captures essential appearance-based features, integrating optical flow significantly enhances the model's ability to capture complex motion dynamics. Table V presents the experimental results comparing the proposed model using HOG-only features and the combination of HOG with optical flow.

TABLE V. MODEL ACCURACY (%) WITH HOG AND OPTICAL FLOW

Dataset	With only HOG features	With HOG and Optical flow features
Action Movies	99.64	100
Hockey Fight	98.97	99.24
Violent Flows	95.26	96.98

c) Temporal modelling: ConvLSTM vs. 3D-CNN

An ablation study was conducted to compare the impact of different temporal modeling architectures. From the experimental analysis, we observed that the ConvLSTM-based architecture demonstrates superior performance over the 3D-CNN model by effectively modeling both spatial and temporal features in video data. The architecture of the proposed ConvLSTM-based model is detailed in Table I. In comparison, the CNN-based variant comprises two convolutional layers with 5 and 10 filters of size 3×3×3, respectively. Each convolutional layer is followed by batch normalization and a max-pooling layer with a 1×2×2 kernel. The model uses same padding and a stride of 1 throughout, with sigmoid as the activation function. Unlike 3D-CNNs, which operate on fixed-length frame sequences and may fail to capture long-range dependencies, ConvLSTM utilizes a recurrent structure that maintains temporal continuity across frame sequences. This allows the network to capture more distinctive motion patterns associated with violent actions, leading to enhanced detection accuracy and more effective temporal feature representation. The experimental results are shown in Table VI.

TABLE VI. MODEL ACCURACY (%) WITH CONVLSTM AND 3D-CNN

Dataset	With ConvLSTM	With 3D CNN
Action Movies	100	99.58
Hockey Fight	99.24	98.73
Violent Flows	96.98	95.85

3) Learning rate and epochs

In experimental analysis, the proposed method is evaluated by varying the hyperparameters, specifically, the learning rate (α) and the number of training epochs. The optimal values considered for α are 0.002 and 0.0005. The proposed model is trained for up to a maximum of 300 epochs.

TABLE VII. ACCURACY OF THE MODEL WITH DIFFERENT LEARNING RATES AND EPOCHS

Dataset	Learning rate	Iterations	Loss	Accuracy (%)
Violent Flows	0.002 and 0.0005	100	0.1814	94.16
		200	0.1825	93.72
		300	0.1731	93.43
		100	0.1229	94.51
		200	0.1262	95.89
		300	0.1103	96.98
Action Movies	0.002 and 0.0005	100	0.0098	99.76
		200	0.0369	99.85
		300	0.0185	100
		100	0.0461	98.97
		200	0.0512	98.91
		300	0.0355	99.03
Hockey Fight	0.002 and 0.0005	100	0.1326	98.62
		200	0.1536	98.22
		300	0.1181	98.47
		100	0.0519	98.76
		200	0.0550	98.94
		300	0.0486	99.24

Table VII presents the model performance for different values of learning rates and epochs. For each configuration, we have recorded the model performance. It is observed that reducing the learning rate from 0.002 to 0.0005 improves the model's accuracy from 94.16% to 96.98% on the violent flows dataset. Furthermore, the loss value decreases from 0.1814 to 0.1103. A similar pattern is observed for the hockey fight dataset, where the accuracy is improved from 98.62% to 99.24%, and the loss value decreases from 0.1326 to 0.0486. For the action movies dataset, the model yields better accuracy results with a higher learning rate. An accuracy of 100% is achieved with a learning rate of 0.002, while 99.03% is obtained when the learning rate is set to 0.0005.

4) Comparative analysis

Our model achieved an accuracy of 100% on the action movies dataset (Fig. 9), an accuracy of 99.24% on the hockey fight dataset (Fig. 10), 96.98% on the violent flows dataset (Fig. 11) respectively. Despite the challenging conditions across different datasets, our model performs better in detecting violent actions. We observed that the model performs better on the action movies and hockey fight datasets for violence detection. However, some videos in the violent flows dataset exhibit heavy occlusion, leading to a slight reduction in detection accuracy, which can impact keyframe selection.

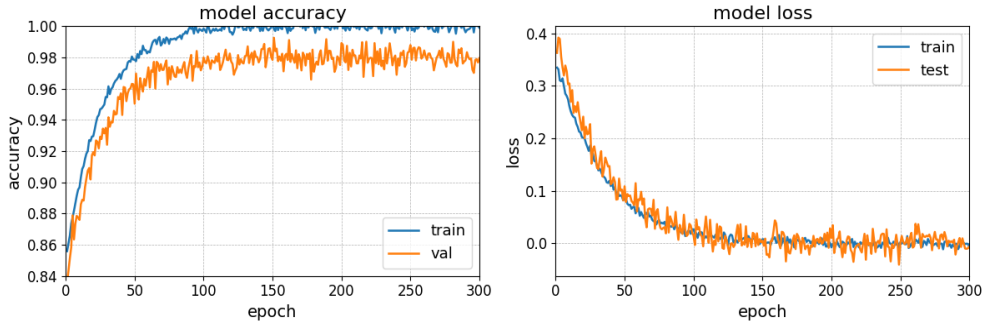

Fig. 9. Model performance on action movies dataset.

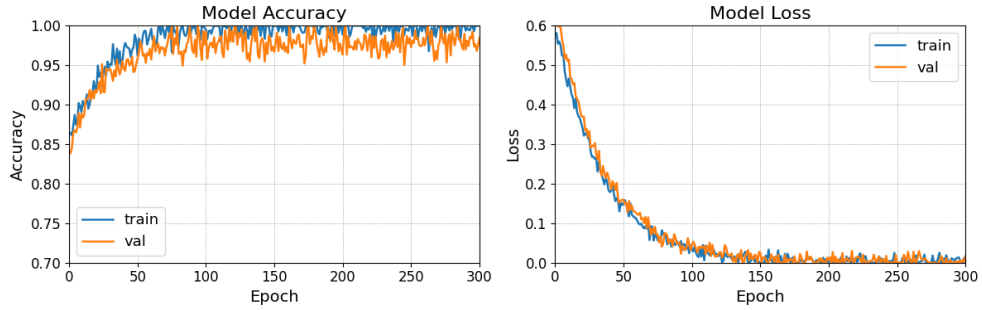

Fig. 10. Model performance on hockey fight dataset.

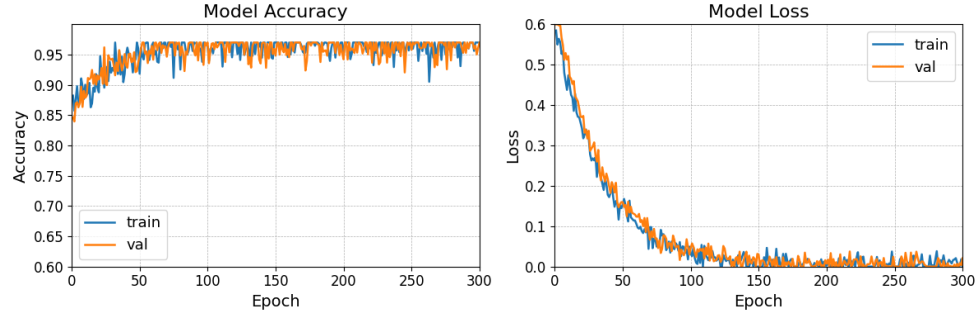

Fig. 11. Model performance on violent flows dataset.

To empirically validate the classification performance and assess no potential overfitting, we have provided confusion matrices for three datasets. The confusion matrices for different datasets (Figs. 12–14), clearly demonstrate the model's robustness and its ability to generalization ability across datasets while maintaining a low false positive and false negative rate.

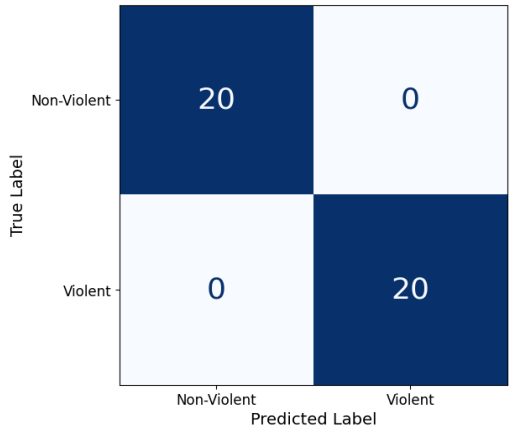

Fig. 12. Confusion matrix for action movies dataset.

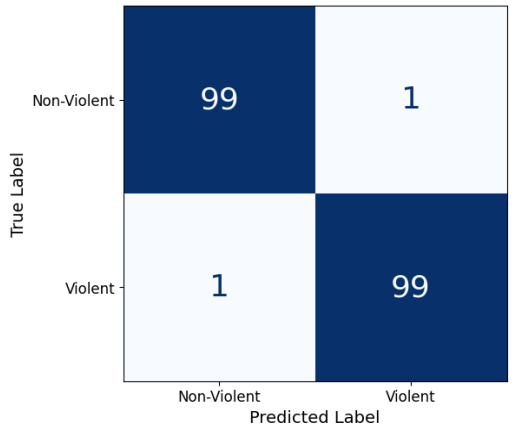

Fig. 13. Confusion matrix for hockey fight dataset.

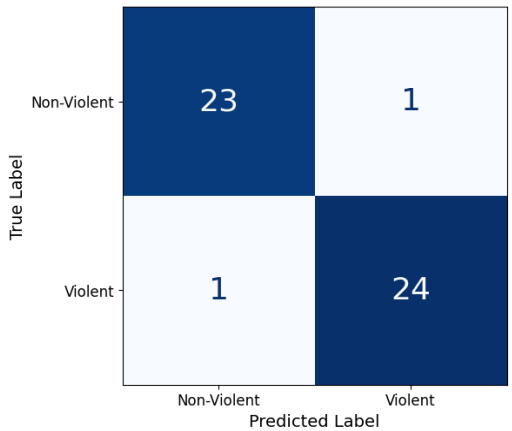

Fig. 14. Confusion matrix for violent flows dataset.

Table IX presents the performance of the proposed model, showing that it consistently performs better than existing state-of-the-art violence detection approaches. The proposed framework optimizes computational efficiency through selective frame processing with keyframe extraction and IoU-based spatial filtering, significantly reducing the number of video frames.

TABLE IX. MODEL ACCURACY (%) COMPARISON ACROSS DIFFERENT DATASETS

Method	Hockey Fight	Violent Flows	Action Movies
DiMOLIF [13]	88.6	85.83	-
ViF + OViF [9]	87.5 ± 1.7	88 ± 2.4	-
HOMO [10]	89.3 ± 0.9	76.8 ±1.7	-
ConvLSTM [28]	97.1 ± 0.5	94.5 ±2.3	100
FightNet [29]	97.0	-	100
HF + 2DCNN [14]	94.6 ± 0.6	-	99.24
3D CNN [17]	98.96	-	99.97
TSM [30]	98.05	96.93	-
ViolenceNet [19]	99.2 ± 0.6	96.9 ±0.5	100
FusionNet [31]	92.89	-	-
U-Net + LSTM [32]	96.10	-	99.50
ConvLSTM [18]	94.5	94.3	98.5
Proposed method	99.24	96.98	100

Additionally, with lightweight YOLOv10 and handcrafted features like optical flow and HOG reduces computational overhead, enabling real-time performance on limited-resource systems. Furthermore, we evaluate the overall processing time for a sample video consisting of 43 frames from the hockey fight dataset. The total processing time is approximately 0.18 s, corresponding to approximate processing rate of 227 Frames per Second (FPS). This demonstrates the high efficiency of the proposed model.

5) Computational complexity

To assess the practicality of our proposed framework for real-time violence recognition, we conducted a thorough analysis of its space and time complexity. The model comprises approximately 4.06 million trainable parameters, occupying just 16.24 MB of disk space. As reported in Table VIII, our method achieves better processing speed compared to existing state-of-the-art approaches.

TABLE VIII. PROCESSING TIME OF A VIDEO FOR VIOLENCE DETECTION

Method	Processing time (s)
HF + 2D CNN [14]	0.46
FTCF Net [21]	0.35
MobileNet + GRU [20]	0.29
Proposed method	0.18

V. CONCLUSION

In this work, we proposed a hierarchical framework to detect violent actions in videos. The YOLO model was employed to identify relevant frames containing humans. Frames with maximum human intersection regions were subsequently selected to discard unnecessary ones. An efficient keyframe extraction strategy was applied during the preprocessing stage to reduce the model's processing overhead. Furthermore, the selected keyframes were refined by identifying high-motion regions using optical flow information between successive frames, which further reduced the processing complexity. For feature

extraction, orientation and optical flow gradient information were used to derive HOG features. A ConvLSTM deep neural network was employed to learn significant spatio-temporal features, thereby enhancing the model's performance. Experimental results on multiple datasets demonstrated that our method outperformed existing approaches. In future work, we aim to improve the preprocessing stage by developing a framework to filter out non-violent videos in the initial phase and optimize the proposed framework for deployment of edge devices to ensure real-time performance in resource-constrained environments.

Ethics Statement

The datasets used in this work (hockey fight, violent flows, and action movies) contain scenes of simulated or real-world violence and are publicly available for research purposes. No personally identifiable information is present. All data usage complies with the respective dataset licenses, and the content is used solely for academic and non-commercial research.

Data Availability

The action movies and hockey fight datasets are collected from [3], and following URL enables to access the violent flows dataset: https://talhassner.github.io/home/projects/violentflows/index.html

The code for the proposed method is publicly available at https://github.com/venkateshakula19/Human-violence-detection-in-videos-using-discriminant-frame-extraction-and-ConvLSTM/tree/main

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Venkatesh Akula: Conceptualization, Methodology, Formal analysis, Investigation, Writing—original draft, Visualization. Ilaiah Kavati: Validation, Resources, Writing—review and editing, Supervision, Project administration. All authors had approved the final version.

References

[1] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, "Dynamic image networks for action recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034–3042. https://doi.org/10.1109/CVPR.2016.331

[2] B. Zhang, H. Xu, H. Xiong *et al*., "A spatiotemporal multi-feature extraction framework with space and channel based squeeze-and-excitation blocks for human activity recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 7983–7995, 2021. https://doi.org/10.1007/s12652-020-02526-6

[3] J. Donahue, L. A. Hendricks, S. Guadarrama *et al*., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878

[4] V. Akula and I. Kavati, "Human violence detection in videos using key frame identification and 3D CNN with convolutional block attention module," *Circuits, Systems, and Signal Processing*, vol. 43, no. 12, pp. 7924–7950, 2024. https://doi.org/10.1007/s00034-024-02824-w

[5] A. Wang, H. Chen, L. Liu *et al*., "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, 107984–108011, 2024. https://doi.org/10.52202/079017-3429

[6] E. B. Nievas, O. D. Suarez, G. B. Garcıa, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. Computer Analysis of Images and Patterns: 14th International Conference*, 2011, pp. 332–339.

[7] T. Hassner, Y. Itcher, and O. Kliper-Gross. "Violent flows: Real-time detection of violent crowd behavior," in *Proc. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–6.

[8] O. Deniz, I. Serrano, G. Bueno, and T. K. Kim, "Fast violence detection in video," in *Proc. 2014 International Conference on Computer Vision Theory and Applications (VISAPP),* 2014, pp. 478–485.

[9] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and Vision Computing*, vol. 48, pp. 37–41, 2016.

[10] A. Mahmoodi and A. Salajeghe, "Classification method based on optical flow for violence detection," *Expert Systems with Applications,* vol. 127, pp. 121–127, 2019.

[11] M. Y. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," *Computer Science Department*, vol. 929, pp. 1–16, 2009.

[12] I. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using mobsift and movement filtering algorithm," *Pattern Analysis and Applications*, vol. 23, no. 2, pp. 611–623, 2020.

[13] A. B. Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognition Letters*, vol. 92, pp. 62–67, 2017.

[14] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using hough forests and 2D convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787–4797, 2018.

[15] A. Keceli and A. Kaya, "Violent activity detection with transfer learning method," *Electronics Letters*, vol. 53, no. 15, pp. 1047–1048, 2017.

[16] D. G. C. Roman and G. C. Ch´avez, "Violence detection and localization in surveillance video," in *Proc. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI),* 2020, pp. 248–255.

[17] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019.

[18] G. Cobo and J. C. SanMiguel, "Human skeletons and change detection for efficient violence detection in surveillance videos," *Computer Vision and Image Understanding*, vol. 233, 103739, 2023.

[19] F. J. Rendon-Segador, J. A. Alvarez-Garcıa, F. Enrıquez, and O. Deniz, "ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence," *Electronics,* vol. 10, no. 13, 1601, 2021.

[20] J. Imran, B. Raman, and A. S. Rajput, "Robust, efficient and privacy-preserving violent activity recognition in videos," in *Proc. 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 2081–2088.

[21] T. Zhenhua, X. Zhenche, W. Pengfei, D. Chang, and Z. Weichao, "FTCF: Full temporal cross fusion network for violence detection in videos," *Applied Intelligence*, vol. 53, no. 4, pp. 4218–4230, 2023. https://doi.org/10.1007/s10489-022-03708-9

[22] I. Kavati, V. Akula, E. S. Babu, R. Cheruku, and G. K. Kumar, "Complex human activity recognition with deep inception learning and squeeze-excitation framework," *Journal of Information Assurance & Security*, vol. 17, no. 2, 2022.

[23] M. Asad, J. Yang, J. He, P. Shamsolmoali, and X. He, "Multi-frame feature-fusion-based model for violence detection," *The Visual Computer*, vol. 37, no. 6, pp. 1415–1431, 2021.

[24] B. T. Naik, M. F. Hashmi, Z. W. Geem and N. D. Bokde, "DeepPlayer-Track: Player and referee tracking with jersey color recognition in soccer," *IEEE Access*, vol. 10, pp. 32494–32509, 2022. doi: 10.1109/ACCESS.2022.3161441

[25] S. Singh, S. Dewangan, G. S. Krishna, V. Tyagi, S. Reddy, and P. R. Medi, "Video vision transformers for violence detection," arXiv preprint, arXiv:2209.03561, 2022.
[26] C. Li, X. Yang, and G. Liang, "Keyframe-guided video swin transformer with multi-path excitation for violence detection," *The Computer Journal*, vol. 67, no. 5, pp. 1826–1837, 2024.
[27] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. IJCAI'81: 7th International Joint Conference on Artificial Intelligence*, 1981, vol, 2, pp. 674–679.
[28] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *Proc. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.
[29] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," *Journal of Physics: Conference Series*, vol. 844, 012044, 2017.
[30] Q. Liang, Y. Li, B. Chen, and K. Yang, "Violence behavior recognition of two-cascade temporal shift module with attention mechanism," *Journal of Electronic Imaging*, vol. 30, no. 4, 043009–043009, 2021.
[31] S. A. Jebur, K. A. Hussein, H. K. Hoomod, and L. Alzubaidi, "Novel deep feature fusion framework for multi-scenario violence detection," *Computers*, vol. 12, no. 9, 175, 2023.
[32] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, 2216, 2022.