Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

A FUSION-BASED CNN-BiLSTM FRAMEWORK FOR ROBUST VIOLENCE DETECTION IN REAL-WORLD VIDEO SURVEILLANCE

MD. SHAFIUL AZAM
Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: msacse@pust.ac.bd

TAHMID RAHMAN
Department of Computer Science and Engineering, Hamdard University, Bangladesh.
Email: tahmid.rahman.cs@gmail.com

ABU SALEH MUSA MIAH
School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima, Japan.
Email: musa@u-aizu.ac.jp

NAKIB AMAN
Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: nakib.cse@pust.ac.bd

MD ABDUR RAHIM*
Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. *Corresponding Author Email: rahim@pust.ac.bd

Abstract

Due to the enormous volume of video material and growing demand for automated surveillance systems, violence detection has become a crucial area of study in computer vision. Law enforcement and security workers might be able to prevent or lessen violent situations by detecting violence in real-time video streams. Deep learning techniques, such as CNNs and LSTMs, have shown promising results in detecting violent activity. However, existing approaches have some limitations, including reduced performance when detecting violence in real-world situations and difficulties differentiating between violent and non-violent activities with similar motion patterns. This paper presents a fully integrated violence detection system that overcomes these limitations by incorporating CNN architectures and BiLSTM with fusion techniques. We analyzed in-depth approaches to violence detection and proposed a novel, effective method. Using a combination of CNNs and a BiLSTM, a reliable framework was built to improve violence detection. This study assesses five CNN designs, including MobileNetV2, ResNet50V2, DenseNet201, Xception, and VGG19, and then integrates them with the BiLSTM network to recognize violent scenes in video data. Furthermore, this paper examines two fusion approaches: intermediate fusion and late fusion. These approaches are tested on two datasets: RLVS and HF. The results reveal that late fusion delivers the highest performance in different metric scores, demonstrating its potential as a superior violence detection approach. We have achieved an accuracy of 98.50% and 97.50% on the RLVS and HF datasets, respectively. This framework might help address the serious issue of violence that affects communities worldwide.

Keywords: Violence Detection, Deep Learning, CNN, BiLSTM, Late Fusion, Video Surveillance, RLVS Dataset.

1. INTRODUCTION

The increasing prevalence of violence has made the identification of violent activities in video feeds essential. Manually analyzing surveillance videos, social media content, and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

media footage is challenging; however, automatic violence detection reduces the amount of data to be analyzed by focusing on key moments. With the rise of video content and surveillance systems, there is an increasing need to analyze vast visual data, particularly to detect violent events crucial to security and public safety. Research on violence detection has grown significantly, especially for real-time applications in public spaces. Advances in human action recognition and computational power have driven the development of intelligent surveillance systems capable of analyzing video footage for applications across sectors such as healthcare, traffic monitoring, and security. Violence detection, which began in 2002, has evolved from handcrafted feature-based methods to more advanced deep learning approaches [1]. Unlike traditional methods that rely on domain knowledge for feature extraction, deep learning models can autonomously detect patterns and features from raw video data. This process enables real-time identification of physical and psychological violence in videos, helping to prevent harm. As violence-detection technology expands, it raises ethical concerns, including privacy issues and the potential for misuse. Effective violence detection systems must accurately distinguish between violent and non-violent behavior, such as sports or social interactions.

Video surveillance systems are increasingly deployed in crowded public areas; however, challenges like poor video quality and inconsistent lighting often hinder accurate detection. Real-time, automated systems are critical to preventing violence escalation, particularly with the rise of live-streamed content on social media. Effective violence detection methods are crucial for creating safer environments and reducing the harm caused by violence. Recent research has applied various machine learning and deep learning techniques to violence detection, including 3D-CNNs [2], dynamic texture identification like Violent Flows (ViF) [3], 3D CNN [4], sparse Gaussian process latent variable model (SGPLVM) [5], multimodal approaches [6], and R-CNNs [7]. However, these methods often fall short due to challenges such as ineffective feature extraction.

To address these limitations, we propose integrating Bi-LSTMs with multi-transfer learning-based CNNs, employing both early and late fusion mechanisms. In our approach, we preprocess the video dataset using YOLOv8 and apply CNNs trained with four transfer learning models: MobileNetV2, DenseNet201, ResNet50V2, Xception, and VGG19, all initialized with ImageNet weights. The input is first fed into the CNN streams, then fused to produce concatenated features, which are passed to the BiLSTM. Alternatively, in the late-fusion approach, we feed inputs into a CNN integrated with a BiLSTM, creating features from two streams that are then concatenated.

The contributions of this paper include:

- A real-time violence detection system for surveillance and media content.
- A systematic comparison of different CNN architectures for violence detection.
- An analysis of CNN+BiLSTM architectures in detecting violence.
- A study on the application of fusion methods to violent video datasets.

This paper is organized as follows. Section 2 presents the literature review on violence detection in real-world video surveillance. Section 3 describes the RLVS and HF datasets.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Section 4 explains the proposed methodology. Section 5 reports the evaluation results. Finally, Section 6 concludes the paper by summarizing the research outcomes.

2. LITERATURE REVIEW

A crowd violence detection model, HD-Net, with good generalizability was presented by Chexia et al. (2022) [2]. HD-Net focuses on human features and dynamic information from neighbouring frames, using 3D-CNN and LSTM for spatial and temporal feature fusion. Hassner et al. (2012) developed Violent Flows (ViF), a dynamic texture-based approach for identifying violence with a linear SVM [3]. Gkountakos et al. introduced a 3D-CNN for analyzing crowd video footage, suitable for standalone desktop applications [4]. Mumtaz et al. (2018) and Naik et al. (2022) used transfer learning to recognize aggressive human behaviours, outperforming traditional models [8-9], while Mugunga et al. (2021) applied ConvLSTM for violence detection in surveillance cameras, improving performance across six benchmark datasets [10]. Moaaz et al. (2020) proposed an end-to-end neural network for detecting violent scenes in surveillance footage [11]. Abdelfatah et al. (2017) used SGPLVM to detect violence in Arabic social media by performing nonlinear dimensionality reduction without labelled data [5]. Several studies have applied deep learning to non-traditional video sources, such as cartoons, video games, activity recognition, and social media platforms [12-13]. Several studies also focused on violence detection using object detection methods such as Faster R-CNN, which Chao et al. (2020) applied to identify terrorist videos on cell phones [7], and Alaquil and Fernandez-Carrobles (2019) used to detect weapons in videos [14-15].

Image processing techniques, including feature extraction and pattern recognition, were used to improve the accuracy of violence detection [16]. Facial recognition techniques such as DeepFace and FaceNet were also employed to identify individuals involved in violent incidents [17], which used YOLO to detect weapons in images with high accuracy. Speech and audio recognition have also been explored, such as Cheng et al.'s (2003) hierarchical approach for identifying gunshots, car brakes, and explosions [18], and Giannakopoulos et al. (2010), who used speech recognition to identify violent behavior through aggressive language analysis [19]. Bakhshi et al. (2023) applied a deep neural network–based voice recognition method for detecting violence in real-world audio signals [20]. Video analysis is a popular method for detecting violence, providing visual data for identifying violent incidents. Nam et al. (1998) pioneered this approach by identifying violent incidents using blood, fire, motion, and distinctive sounds [21]. YOLO has been widely used in detecting hostile gestures, weapons, and other violent behaviors in videos. Sethi et al. (2025) applied YOLO to detect hostile gestures in crowded environments [22]. Motion analysis techniques, such as Optical Flow and Background Subtraction, have also been employed for violence detection, with studies by Garje et al. (2018), Jain et al. (2020), and Clarin et al. (2005) examining motion patterns in videos [23-25]. Bermejo et al. (2011) proposed using the Bag-of-Words framework and MoSIFT (an extension of SIFT) to detect violence through motion [26]. Earlier methods for action recognition and feature extraction often relied on handcrafted descriptors like MoSIFT, which itself builds upon foundational keypoint detection algorithms [27, 28].

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Scene understanding methods, such as Deep Learning-based Semantic Segmentation, have been used to analyze video context to detect violence. Pham et al. (2022), Wu et al. (2017), and Ilyas et al. (2024) applied scene understanding techniques to enhance violence detection [29-31]. Multimodal approaches combining video, audio, and text have been shown to improve detection accuracy by capturing multiple aspects of violent incidents.

The CASSANDRA system (Aktı et al., 2019) analyses motion features and audio cues, such as screams, to detect violence in surveillance footage [32]. Gong et al. (2008) used low-level visual and auditory features, along with high-level audio effects, to detect violence in movies [33]. Peixoto et al. (2020) examined decomposed subconcepts of aggression in both visual and auditory forms, combining results from several neural networks [34]. Giannakopoulos et al. (2010) proposed a k-Nearest Neighbor classifier that combined audio statistics and video motion data to detect violence [35]. Another method condenses entire video sequences into motion-detailed grayscale images for classification via 2D CNN [36]. The use of 3D convolutional networks to directly learn spatiotemporal features from video data was a significant advancement, as demonstrated by Tran et al. [37]. Chunhui et al. (2014) used a 3D ConvNet to learn spatiotemporal properties of video data without prior knowledge, while Zihang et al. (2017) employed ConvNet streams to detect violent movements using temporal and spatial features [38-40]. Swatikiran et al. (2017) introduced a convLSTM architecture that combines CNNs and LSTMs for spatiotemporal analysis of video frames [41]. For feature extraction, the authors employ a variety of CNN architectures, including VGG16 [42] and Xception [43]. A Bi-LSTM is used for the categorization to understand the relationship between historical and prospective data. An additional attention layer then determines the significant input regions.

3. DATASET DESCRIPTION

In this study, we selected relevant datasets from various sources to support our research. Since project-specific datasets were limited, we incorporated several previously used datasets after verifying their compatibility with our proposed system. The datasets used in this work are the Real-Life Violence Situations (RLVS) [44] and Hockey Fights (HF) [26] datasets.

Table 1: Statistical Information of the RLVS and HF dataset.

Dataset	Videos	Violent	Non-Violent	Duration(s)	FPS
RLVS	2000	1000	1000	2-6	30
HF	1000	500	500	1-2	25

3.1 Real Life Violence Situations Dataset (RLVS)

The RLVS dataset contains 1,000 violent and 1,000 non-violent YouTube videos. The violent clips showcase street fights, while the non-violent ones depict everyday activities such as sports, eating, and walking. Each video lasts 2 to 6 seconds, with over 100 frames at 25 frames per second. Figure 1 shows an example of the RLVS dataset.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Figure 1: Example of RLVS Dataset.

3.2 Hockey Fight (HF) Dataset

The Hockey Fight Dataset for violence detection contains 1,000 videos, divided into two categories: 500 fight videos and 500 non-fight videos. All videos are sourced from hockey matches. The 'fight' category includes videos with violent scenes, while the 'non-fight' category consists of non-violent videos. Each video lasts 1 second and contains 41 frames. Figure 2 presents a representative example from the HF dataset, and Figure 3 illustrates the sample counts per class in the RLVS and HF datasets.

Figure 2: Example of HF Dataset

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

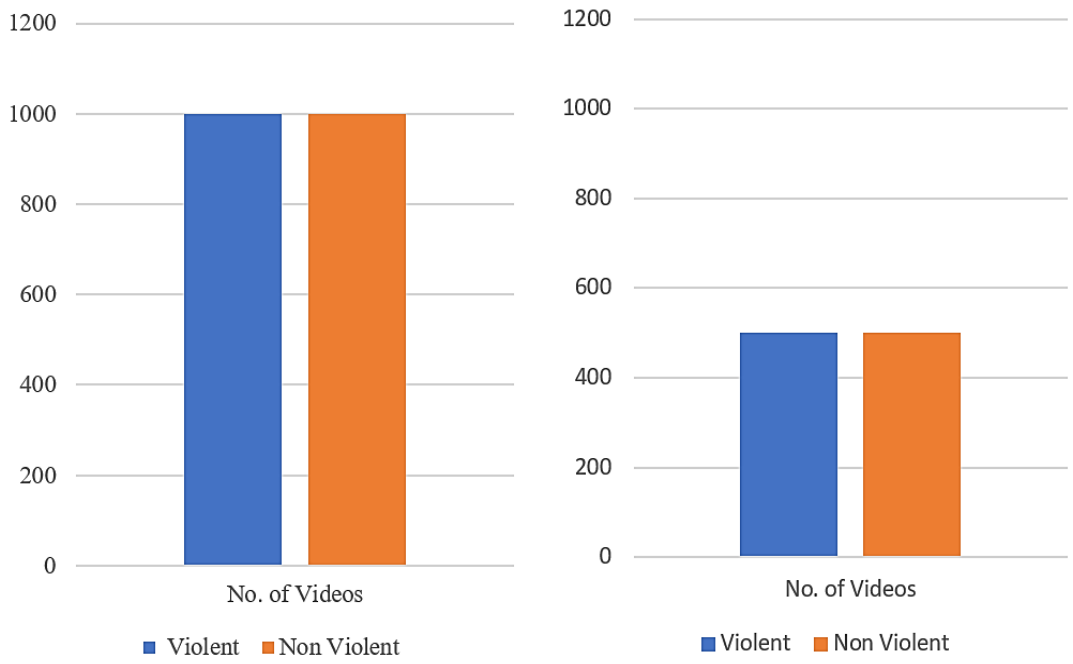


Figure 3: Sample Count Per Class in RLVS and HF Dataset

4. PROPOSED METHODOLOGY

Figure 4 depicts the overall architecture of the increasing demand for automated systems capable of reliably identifying violent behaviour across settings such as schools, public spaces, and public transportation, which has made violence detection a significant research topic in recent years. Despite its importance, detecting violence remains a challenging task due to the complex and dynamic nature of violent behaviour, as well as the difficulties in effectively recording and processing visual data. To address these challenges, we propose a novel approach to violence detection that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, along with several feature fusion strategies. This approach leverages popular pre-trained CNN models, including MobileNetV2, DenseNet, ResNet50V2, Xception, and VGG19, all initialized with ‘imagenet’ weights. These models, trained on large image datasets, are well-known for their ability to extract robust spatial features from visual signals. Meanwhile, BiLSTMs are used to capture temporal dependencies in data sequences, thereby enhancing the detection of motion patterns over time.

This study compares the performance of violence detection using only spatial information (via CNNs) with that of the combination of spatial and temporal information (via CNN+BiLSTM). Additionally, we evaluate the effectiveness of early and late fusion strategies to determine the most effective approach for combining features extracted by the CNN and BiLSTM layers. Our goal is to improve the accuracy of violence detection by integrating the strengths of both CNNs and BiLSTMs while exploring optimal fusion strategies.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

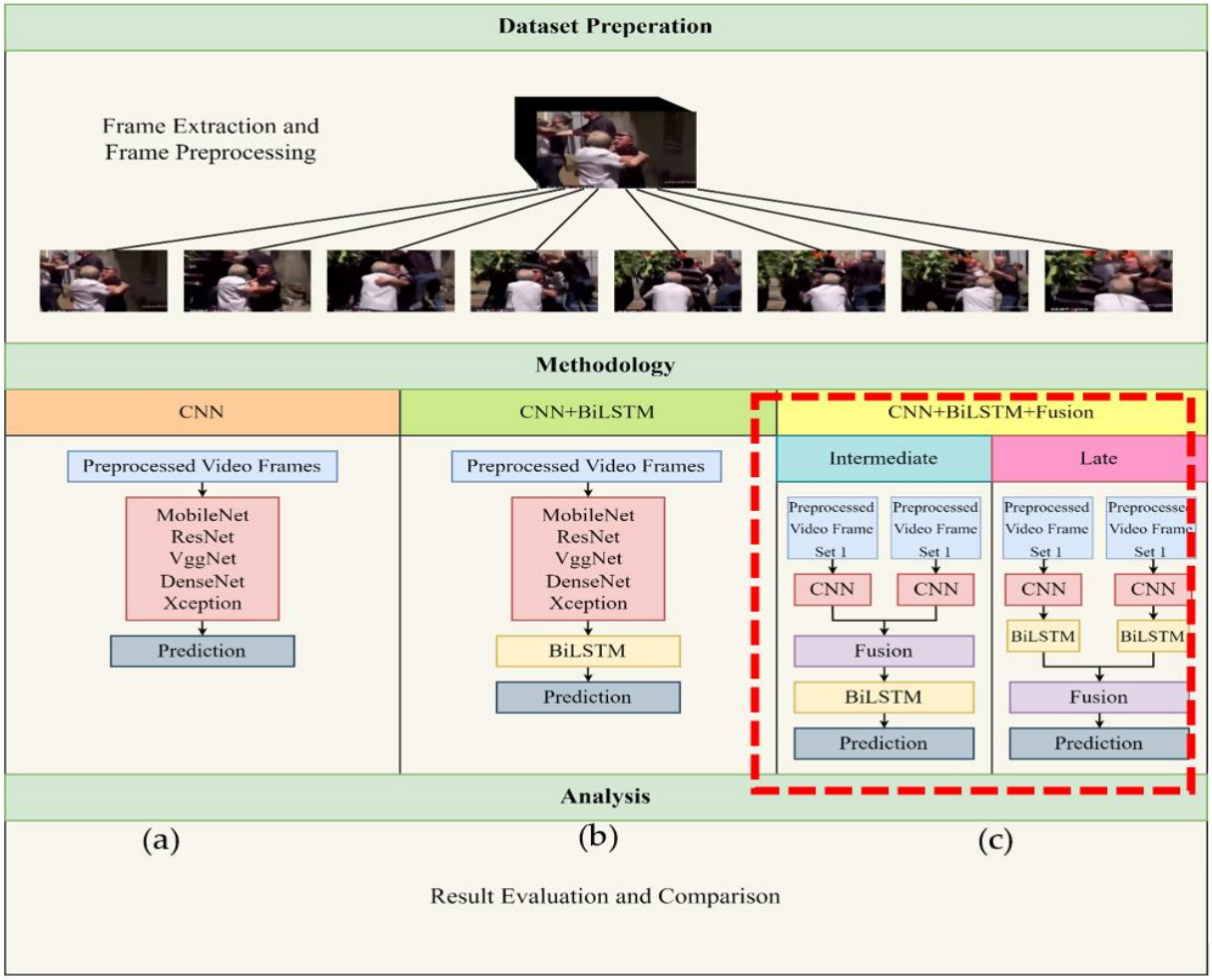


Figure 4: Overall structure of the proposed model.

4.1. Preprocessing

The preprocessing phase begins with extracting 16 frames from each video. The skip window for frame extraction is calculated using Equation (1).

$$skip\_value = \frac{video\_frames\_count}{sequence\_length} \quad (1)$$

where the sequence length is set to 16 frames. The frames are resized to the specified dimensions, and their pixel values are normalized to the range [0, 1]. Finally, the dataset is split into 80% for training and 20% for testing.

4.2. Stage-1: CNN Approach

According to Figure 4(c), Figure 5 shows the proposed CNN approach, which was constructed using transform learning and other CNN Layers. The Convolutional Neural Network (CNN) approach has become a powerful tool for image and video analysis, including violence detection. This section explores CNNs' role in extracting features from video data and examines the various CNN architectures used in previous studies. We will

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

discuss the advantages and limitations of CNNs for violence detection and highlight how they have been optimized for this task. By reviewing relevant studies, this section aims to identify the most promising CNN architectures for effective violence detection.

In this study, 5 CNN architectures were examined. These are MobileNetV2 [45], DenseNet201 [46], ResNet50V2 [47], Xception [47], and VGG19 [48]. Their speciality is feature extraction, particularly spatial features from video frames. Figure 5 illustrates the CNN architecture workflow followed in our study.

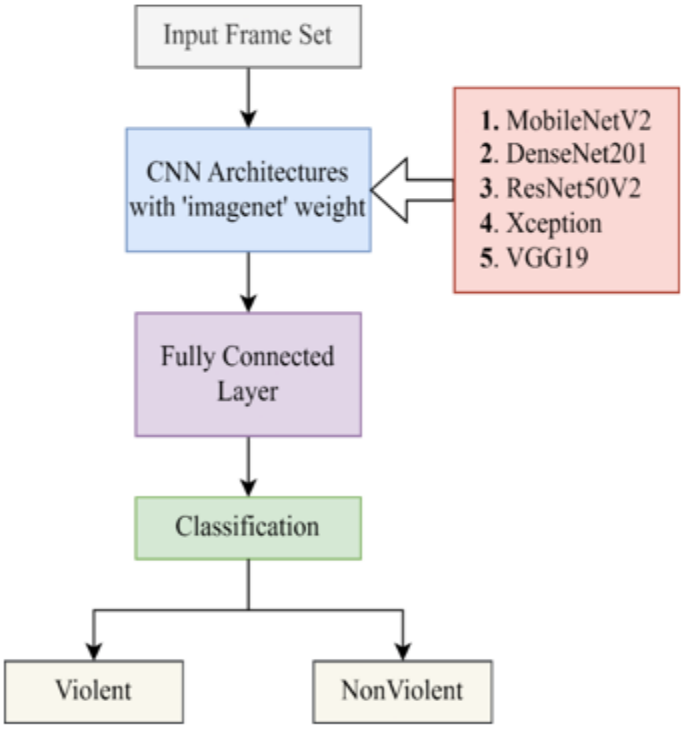


Figure 5: Basic workflow of CNN architecture.

The model architecture, shown in Figure 5, is designed for video analysis, specifically for detecting violent scenes. It takes a set of frames as input, which a time-distributed CNN processes. The CNN can be selected from five options: MobileNet, ResNet, VGGNet, Xception, or DenseNet. The CNN output is passed through a dropout layer to prevent overfitting, followed by a dense layer with ReLU activation and another dropout layer. Finally, the output is processed by a dense layer with softmax activation, providing the predicted class label.

4.2.1 MobileNet

MobileNet, developed by Google in 2017, is a deep CNN architecture designed for efficient processing on mobile and embedded devices, balancing accuracy and computational cost. Its key feature is depth-wise separable convolution, which reduces computation and memory usage without compromising accuracy. In conventional convolutions, each filter processes all input channels, but in depth-wise separable

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

convolution, filters process one channel at a time, followed by a pointwise convolution to combine the results. This makes MobileNet faster and more resource-efficient, ideal for devices with limited capabilities. The MobileNet V2 architecture consists of two types of convolution layers: 1x1 Convolution and 3x3 Depthwise Convolution. Each block includes three layers: 1x1 convolution with ReLU, depthwise convolution, and 1x1 convolution without non-linearity. For our study, we used pre-trained "imagenet" weights, excluding the top layer to tailor the network to our specific task. The last 40 layers are trainable, and the output is passed to a fully connected layer for class prediction [45].

4.2.2 DenseNet

DenseNet, introduced by Gao Huang et al. in 2016, is a deep CNN architecture known for its dense connectivity, in which each layer connects to every other layer. This structure allows for efficient data flow and enables DenseNet to achieve high accuracy with fewer parameters than traditional CNNs. Each layer receives input from all preceding layers and passes its feature maps to all subsequent layers, promoting feature reuse and enhancing information flow. DenseNet also incorporates batch normalization and transition layers. Batch normalization normalizes activations, reducing internal covariate shift and enabling faster convergence. Transition layers perform feature pooling, reducing spatial dimensions and computational cost. For our analysis, we used pre-trained "imagenet" weights, disabling the top layer to tailor the network for the task. The final 40 layers are made trainable, and the output is passed to a fully connected layer for class prediction [46].

4.2.3 ResNet

Introduced by Kaiming He et al. in 2015, ResNet is a deep CNN architecture designed to overcome challenges in training intense neural networks, such as vanishing gradients and performance degradation. Its core innovation is the use of residual connections, which allow data to bypass several layers. Instead of learning new representations at each layer, ResNet layers learn to add a residual signal to the representation from the previous layer, addressing the vanishing gradients problem. ResNet’s simplified architecture requires fewer parameters and less memory than traditional CNNs, making it computationally efficient and low-latency. For this study, we used ‘imagenet’ pre-trained weights, excluding the top classification layer to tailor the model to our task. The last 40 layers are trainable, and a fully connected layer processes the output to predict the class [47].

4.2.4 Xception

Introduced by Google in 2016, Xception is a deep CNN architecture designed to enhance computational efficiency while maintaining accuracy across various computer vision tasks. The key innovation in Xception is the use of depthwise separable convolutions, which reduce computation and memory usage without sacrificing accuracy. Unlike conventional convolutions, which process all input channels, depthwise separable convolutions process each channel individually, followed by a pointwise convolution to combine the results. Xception's simplified architecture requires fewer parameters and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

less memory, making it computationally efficient with low latency and ideal for real-world applications. For our study, we modified the Xception network to train the last 40 layers, initializing the model with pre-trained 'imagenet' weights. The final layer is connected to a fully connected layer for class prediction [47].

4.2.5 VGGNet

In 2014, the Visual Geometry Group at the University of Oxford introduced VGGNet, a deep CNN architecture designed for image classification and object recognition tasks. VGGNet, shown in Figure 3.17, is characterised by its use of many convolutional and pooling layers with a small number of neurons per layer, resulting in a dense, deep network with numerous parameters. Its depth, small filters, and non-linear activation functions allow it to learn a detailed representation of the input data. VGGNet's consistent architecture, with the same number of neurons and activation function in each layer, simplifies training and reduces the risk of overfitting. The architecture has proven effective for various computer vision tasks and is widely used for transfer learning, where pre-trained models are fine-tuned for specific tasks with limited data. For our study, we used the VGG19 model, which has 19 layers. We used all layers for training, initialised them with 'imagenet' pre-trained weights, and modified the final layer to a two-neuron fully connected layer to suit our task [48].

4.3 Stage-2: BiLSTM Integration with CNN Approach

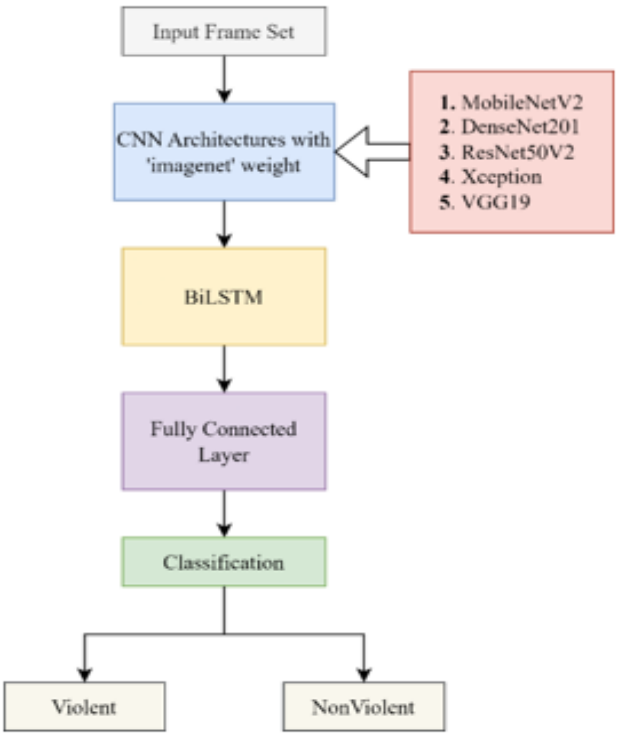


Figure 6: Workflow for BiLSTM integration with Bi-LSTM architecture

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Accurate video classification requires capturing both spatial attributes and their temporal evolution. The Convolutional Neural Network (CNN) extracts spatial features, such as edges, textures, and shapes, but cannot capture temporal dependencies in data sequences.

In contrast, Bidirectional Long-Term Memory (BiLSTM) networks are designed to capture temporal dependencies, making them ideal for processing sequences of video frames. Figure 6 presents the workflow for integrating BiLSTM into the Bi-LSTM architecture.

This section explores how the hybrid CNN+BiLSTM approach, shown in Figure 7, enhances violence detection accuracy by combining CNN spatial feature extraction with BiLSTM temporal sequence modelling. We will discuss how BiLSTMs process temporal information in video data and how combining them with CNNs improves performance in violence detection.

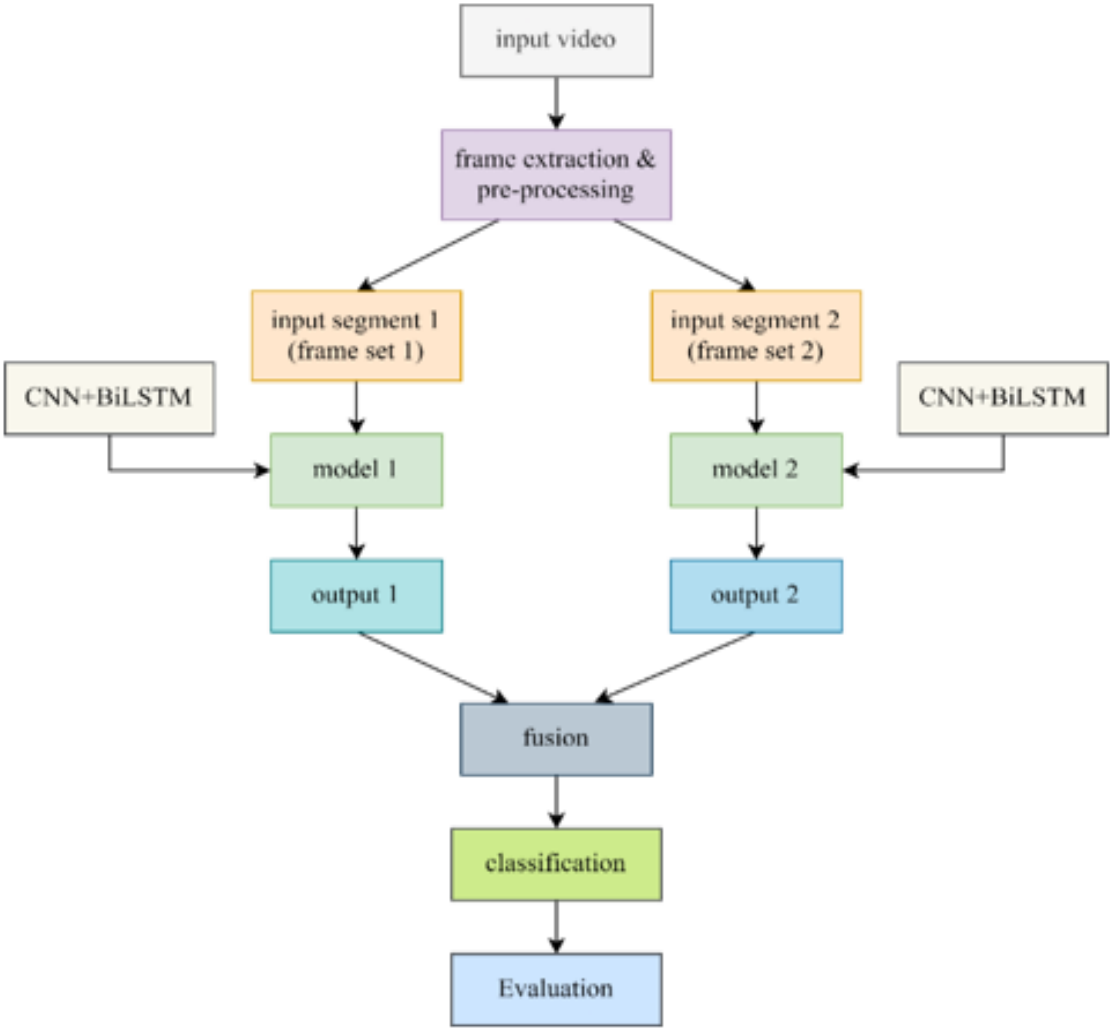


Figure 7: Workflow of the hybrid CNN–BiLSTM model with fusion strategies for violence detection

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

4.4 Fusion

Fusion is the process of combining multiple data sources to provide a more comprehensive representation. Fusion combines various data sources to enhance system performance and provide a more comprehensive representation. It can merge features, methods, or modalities to improve outcomes. The primary fusion approaches are early, intermediate, and late. Early fusion integrates multiple sources into a single representation at the start. Intermediate fusion combines data at an intermediate stage, while late fusion analyzes each source independently before merging the outputs. The choice of fusion method depends on system requirements, such as data types and performance goals. This section outlines the data fusion methodology used in this study, which combines data from various sources to provide a holistic understanding of the research problem. We implemented intermediate fusion and late fusion, using different sets of frames for each model (Figure 7).

In intermediate fusion, data from multiple sources is processed separately and then combined into a single representation, offering a more comprehensive description than individual sources. The model structure, shown in Figure 8 (a), uses two inputs with identical structures but different frame sets. The data is passed through a time-distributed CNN, followed by a dropout layer to prevent overfitting, and then flattened. The outputs are concatenated and processed by a Bidirectional LSTM layer that considers both past and future contexts. The final predictions are made after passing through a dense layer, then another dropout layer, and finally a dense layer. The model's input shape is (? 16, 64, 64, 3), where '?' is the number of sequences, 16 is the number of time steps, and (64, 64, 3) represents the image dimensions.

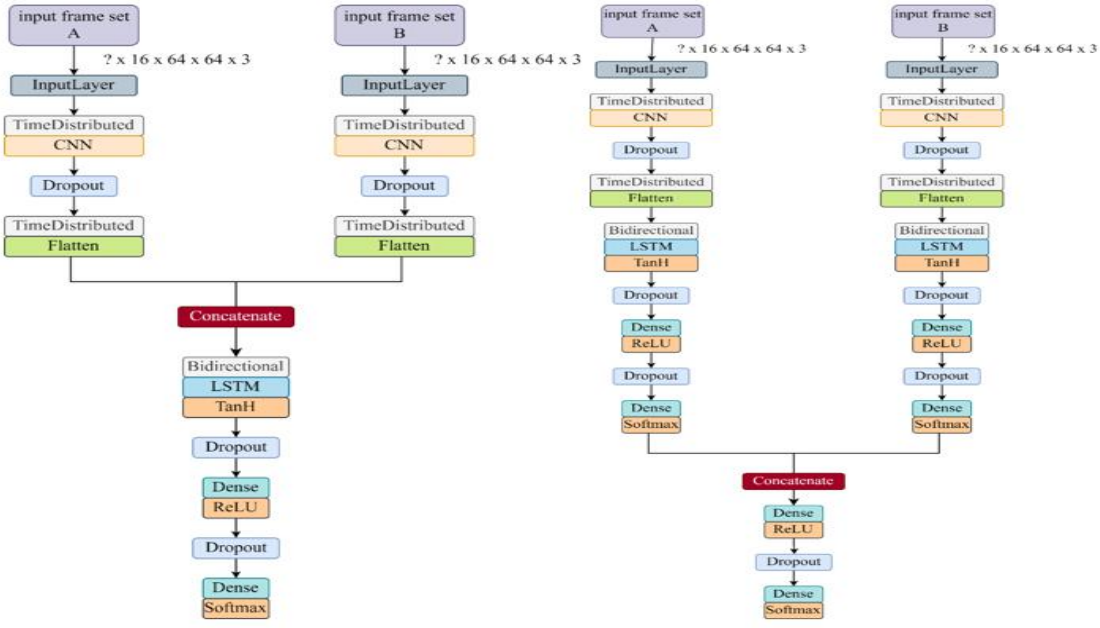


Figure 8: Fusion model structures: (a) intermediate fusion, (b) late fusion

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Late fusion extracts decisions from single-modality architectures and applies a fusion algorithm to compute the final decision. In this approach, information from multiple sources is analyzed independently before combining the results for a final prediction. Figure 8(b) illustrates the architecture of our late fusion implementation.

Initially, we trained separate models on different input frame sets. The predictions from each model were then concatenated and passed through fully connected layers for the final decision. As shown in Figure 8, the model consists of two separate structures, each designed to handle different frame sets.

Both structures follow the same sequence: an input layer, a TimeDistributed CNN for feature extraction, a dropout layer for regularization, TimeDistributed Flatten, a Bidirectional LSTM to model temporal dynamics, a Dense layer with ReLU activation, another Dropout layer, and a Dense layer with softmax activation.

The outputs from both structures are concatenated, passed through a Dense layer with ReLU activation, a Dropout layer, and a final Dense layer with softmax activation to produce the final prediction.

5. EXPERIMENTAL RESULTS

5.1 Contrast and Ablation Experiment

For the CNN + BiLSTM model, BiLSTM is used to capture both past and future context, addressing the CNN's limitation to local information. This allows the model to detect temporal changes in video frames, which are essential for recognizing violence. Tables 2 and 3 show the improved performance of the CNN + BiLSTM architecture in identifying violence across multiple metrics.

Table 2: Performance Comparison on the RLVS Dataset with the CNN+BiLSTM Approach

Methods	Precision	Recall	F1	Accuracy
MobileNet+BiLSTM	0.9347	0.9490	0.9418	0.9425
DenseNet+BiLSTM	0.9196	0.9482	0.9337	0.9350
ResNet+BiLSTM	0.9447	0.9641	0.9543	0.9550
Xception+BiLSTM	0.9347	0.9163	0.9254	0.9250
VGG19+BiLSTM	0.9548	0.9500	0.9524	0.9525

Table 3: Performance Comparison on the HF Dataset with the CNN+BiLSTM Approach

Methods	Precision	Recall	F1	Accuracy
MobileNet+BiLSTM	0.9479	0.9192	0.9333	0.9350
DenseNet+BiLSTM	0.9479	0.9286	0.9381	0.94
ResNet+BiLSTM	0.9375	0.9184	0.9278	0.93
Xception+BiLSTM	0.9271	0.9468	0.9368	0.94
VGG19+BiLSTM	0.9167	0.9072	0.9119	0.9150

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

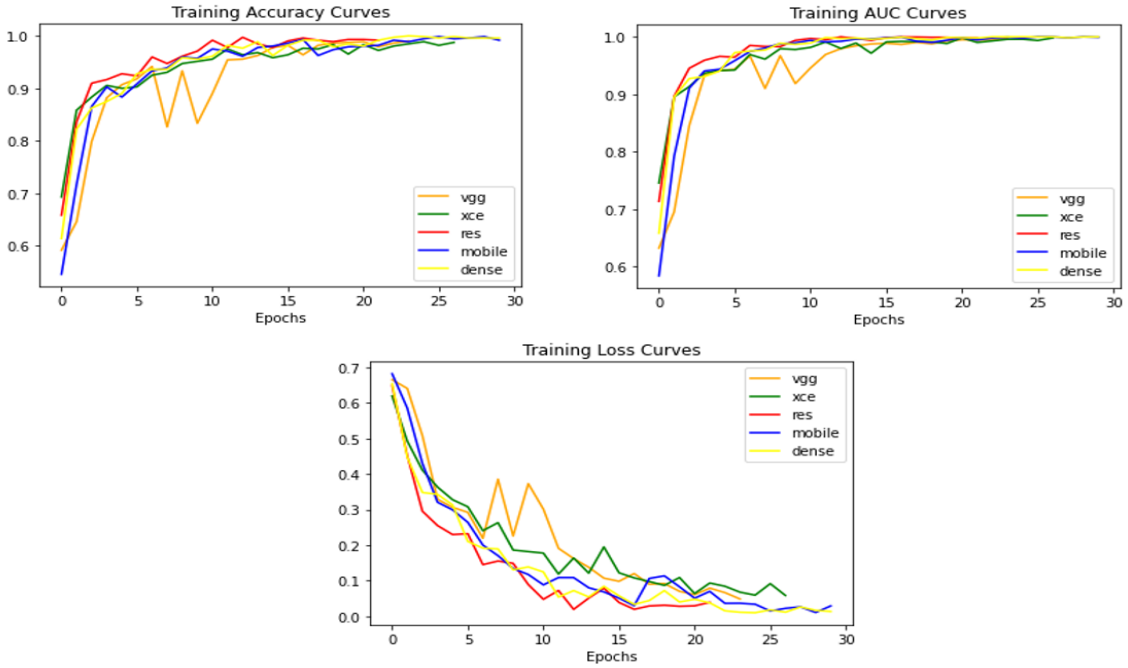


Figure 9: Training Accuracy, AUC, and Loss Curves for RLVS Dataset using CNNs + BiLSTM

The graphs in Figures 9 and 10 illustrate the networks' training history. We can see that ResNet performed better across all metrics than other models.

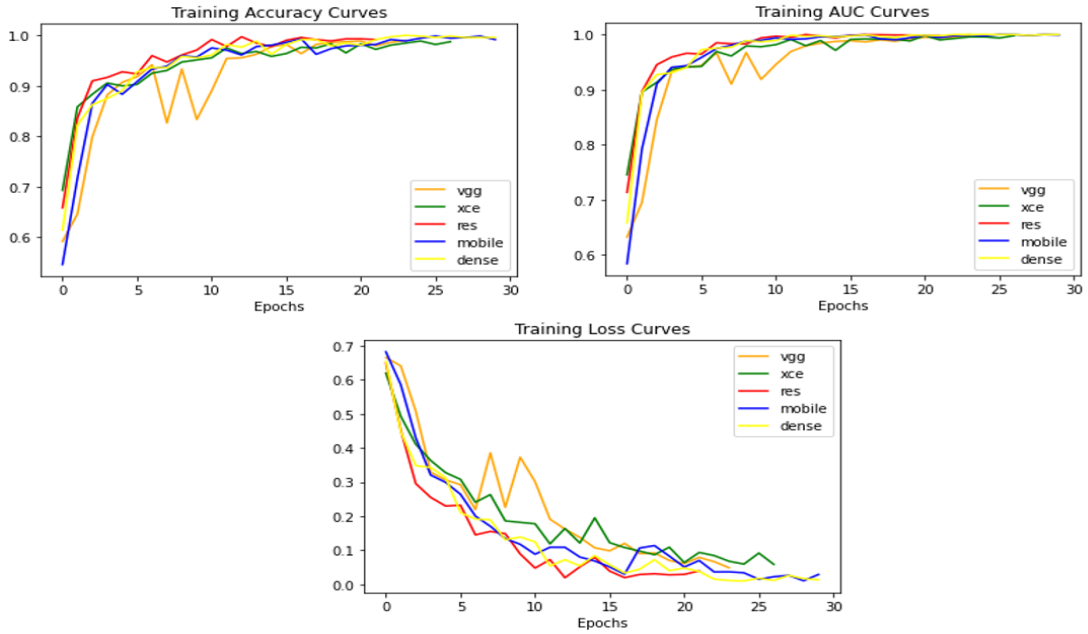


Figure 10: Training Accuracy, AUC, and Loss Curves for HF Dataset using CNNx + BiLSTM

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

5.2 Performance and Real-Life Violence Detection

Table 4 compares the performance of the CNN+BiLSTM architecture across the datasets. ResNet+BiLSTM performed significantly better than the other architectures.

Table 4: Comparative performance analysis of different CNN architectures combined with BiLSTM

Methods	Real Life Violence Dataset	Hockey fight dataset
MobileNet+BiLSTM	0.9425	0.9350
DenseNet+BiLSTM	0.9350	**0.94**
ResNet+BiLSTM	**0.9550**	0.93
Xception+BiLSTM	0.9250	**0.94**
VGG19+BiLSTM	0.9525	0.9150

5.3 Results of Fusion Methods

In the previous sub-section, we observed the improved results from combining CNNs with BiLSTMs. In this section, we explore the performance of fusion techniques. We used MobileNet and ResNet for the fusion implementation, as they performed well in earlier approaches.

The fusion method effectively captures both spatial and temporal information (CNN + BiLSTM) across different input frame sets, making it more efficient and better suited to our objective. The results of the intermediate and late fusion approaches are displayed in Tables 5 and 6, respectively, for precision, recall, F1, and accuracy scores on the RLVS and HF datasets.

In Table 7 and Figure 11, we compared the performance of the intermediate and late fusion approaches on the selected datasets. We can see that late fusion performs considerably better than intermediate fusion.

Table 5: Performance Comparison of Intermediate Fusion on the RLVS and HF Dataset

Dataset	Precision	Recall	F1	Accuracy
RLVS Dataset	0.9688	0.9588	0.9637	0.9650
HF Dataset	0.9688	0.9490	0.9588	0.96

Table 6: Performance Comparison of Late Fusion on the RLVS and HF Dataset

Dataset	Precision	Recall	F1	Accuracy
RLVS Dataset	0.9899	0.9801	0.9850	0.9850
HF Dataset	0.9688	0.9789	0.9738	0.9750

Table 7: Fusion Result Comparison

Methods	Real Life Violence Dataset	Hockey Fight Dataset
Intermediate	0.9650	0.96
Late	0.9850	0.9750

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

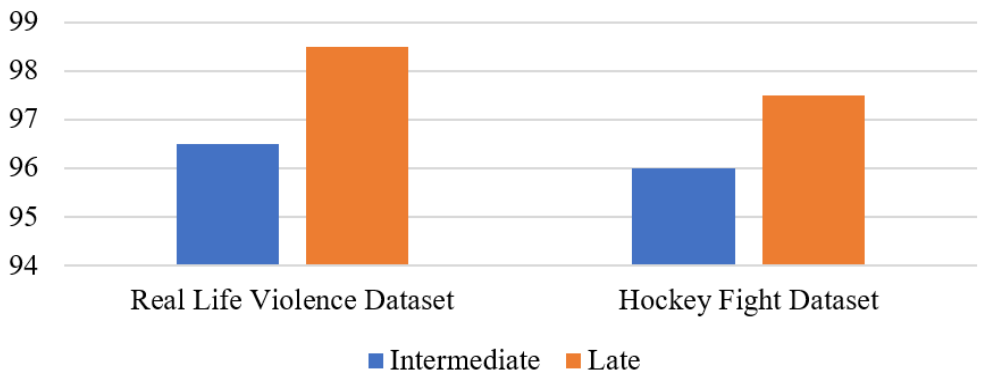


Figure 11: Fusion Method Performance Comparison Chart on the RLVS and HF Datasets

5.4 State of the Art Comparison

Table 8 and Figure 12 summarize the best results from each approach. The chart highlights the performance evolution from CNN to CNN+BiLSTM to Fusion. CNN architectures, which capture only local spatial information, performed worst, with accuracies of 91.50% and 90% on the two datasets. Adding temporal information through Bi-directional LSTM CNN+BiLSTM) improved the model's performance, achieving 95.50% and 94% accuracy. Finally, our proposed Late Fusion approach delivered the best results, achieving state-of-the-art accuracy of 98.50% on the Real-Life Violence Situations Dataset and 97.50% on the Hockey Fights Dataset.

Table 8: Comparison of the Best Performance of Each Approach

METHODS	RLVS Dataset	HF Dataset
CNN	0.9125	0.90
CNN + BiLSTM	0.9550	0.94
Proposed Approach (CNN+BiLSTM with Late Fusion)	0.9850	0.9750

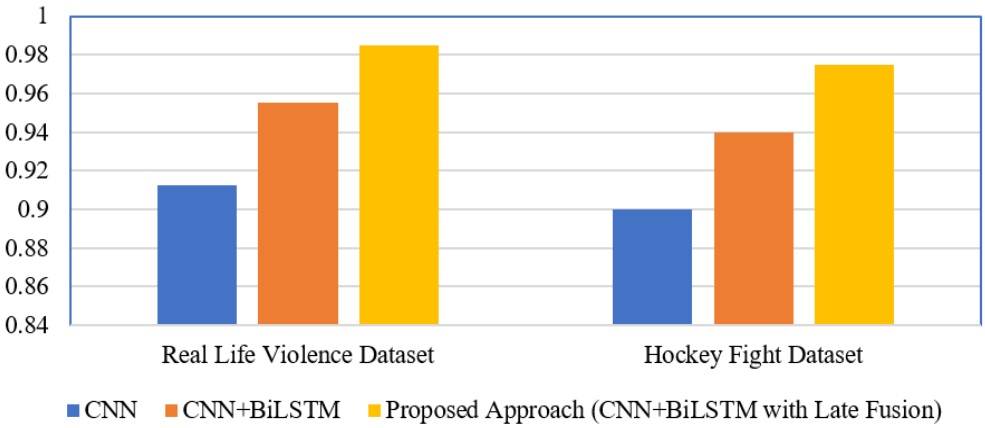


Figure 12: CNN, CNN+BiLSTM and Proposed Approach Performance Comparison Chart

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Table 9 compares our results with previous approaches. The late-fusion approach we proposed has achieved better results than all previous approaches.

Table 9: Comparison with Previous Methods and the Proposed Method: RLVS Dataset

Methods	RLVS Dataset (%)
VGG16+LSTM (Soliman et al.) [48]	88.80
CNN+LSTM+FeedForward (Lima et al.) [48]	91.00
CNN+LSTM (Moaaz et al.) [8]	92.00
DeVTr (Abdali et al.) [49]	96.25
Proposed Approach (CNN+BiLSTM with Late Fusion)	98.50

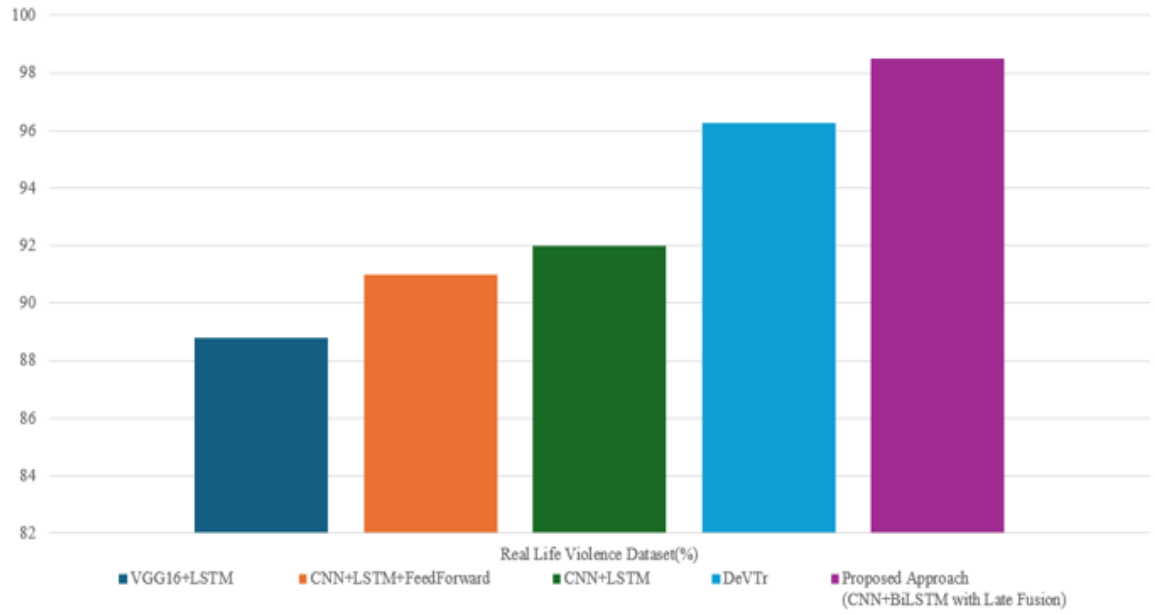


Figure 13: Comparison Chart with Previous Methods and Proposed Approach for RLVS Dataset

Table 10: Comparison with Previous Methods and Proposed Method for Hockey Fights Dataset

METHODS	HF Dataset
Bag-of-Words (Nievas et al.) [28]	90.9
ViF (Hassner et al.) [3]	82.9
MoSIFT+KDE (Long et al.) [50]	94.3
Three streams+LSTM (Zhihong et al.) [51]	93.9
Proposed Approach (CNN+BiLSTM with Late Fusion)	97.50

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

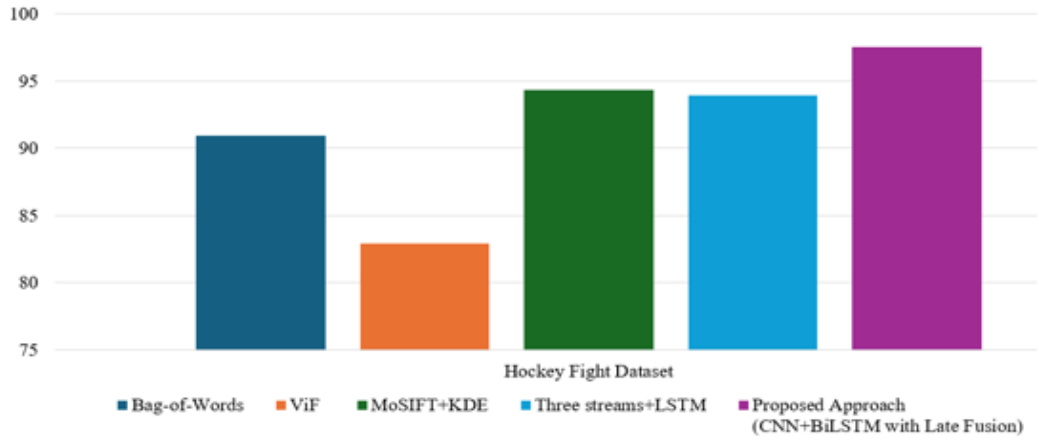
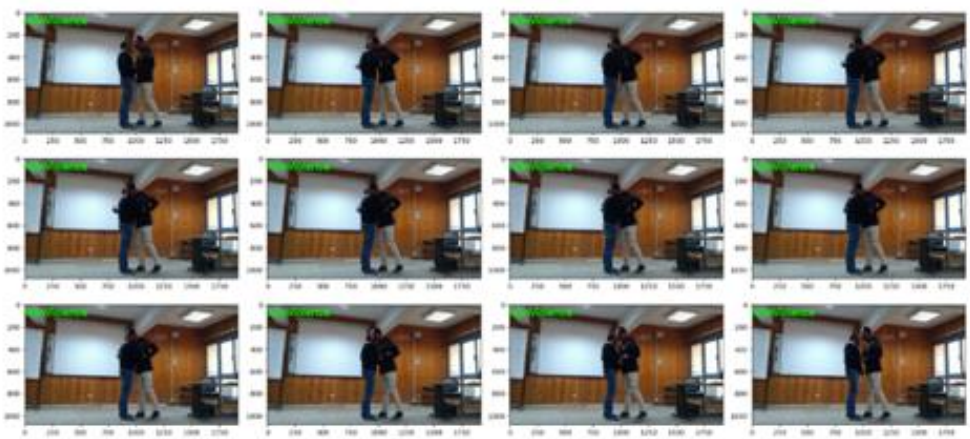


Figure 14: Comparison Chart with Previous Methods and Proposed Approach for HF Dataset

Tables 9 and Figure 13 compare the performance of different methods for violence detection in the RLVS Dataset, with the proposed fusion method achieving the highest accuracy of 98.50%, surpassing all other techniques, including the previous state-of-the-art method, DeVTr. Similarly, Table 10 and Figure 10 show that the fusion method also achieved the highest accuracy of 97.50% on the Hockey Fights Dataset, outperforming all other methods. These results demonstrate the effectiveness of our method for violence detection across various scenarios and datasets. However, it's important to note that evaluation metrics and dataset characteristics can differ across studies, potentially affecting the comparability of results. Therefore, thorough and rigorous evaluations are recommended.

5.5 Testing Random Data

Figure 15 and Figure 16 show visual frames and model outputs from a random street brawl video.



Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

Figure 15: Frame-by-Frame Prediction for (a) a Non-violent Video sample and (b) a Violent Video Sample

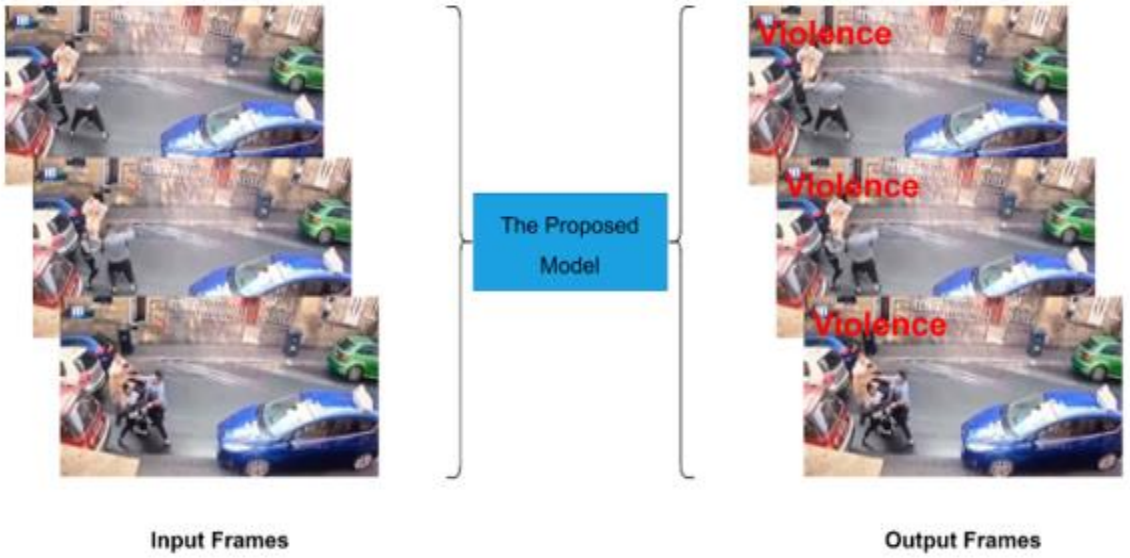


Figure 16: The Proposed Model's Visual Outputs Displayed over Testing Data Obtained from YouTube

The model receives a single video as input, deconstructs it into frames, selects a set number of frames with a fixed interval, processes them, and then makes a prediction about the video's nature. It also provides a prediction of the input data frame by frame.

6. CONCLUSION

This study introduces an advanced violence detection system by integrating CNNs with BiLSTM networks and fusion techniques. The system utilizes five state-of-the-art CNN architectures (MobileNetV2, ResNet50V2, DenseNet201, Xception, and VGG19) alongside BiLSTM to improve the detection of violent activities in video streams. The late

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

fusion method demonstrated the best performance, achieving 98.50% and 97.50% accuracy on the RLVS and HF datasets, respectively, surpassing existing methods. These results show the system's effectiveness in overcoming challenges, such as distinguishing between violent and non-violent actions with similar motion patterns. Looking ahead, several promising areas for future research emerge. While violence detection has seen considerable progress, further improvements in accuracy and handling complex, unpredictable scenarios are needed. Incorporating more diverse datasets and advancing algorithms for real-world challenges will be essential. Ethical considerations should also be addressed by minimizing biases in data and ensuring fairness, transparency, and respect for privacy. Additionally, combining violence detection with technologies like facial recognition and audio analysis could provide a more comprehensive understanding of situations. Furthermore, the focus could shift from post-event detection to real-time intervention, where systems could detect escalating violence and trigger immediate responses. Expanding the technology to include audio analysis and social media monitoring would also enable broader applications, enhancing detection in various environments.

References

1) A. Datta, M. Shah and N. da Vitoria Lobo, "Person-on-person violence detection in video data," 2002 International Conference on Pattern Recognition, Quebec City, QC, Canada, 2002, pp. 433-438 vol.1, doi: 10.1109/ICPR.2002.1044748. Z.

2) Z. Chexia, Z. Tan, D. Wu, J. Ning and B. Zhang, "A Generalized Model for Crowd Violence Detection Focusing on Human Contour and Dynamic Features," 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Taormina, Italy, 2022, pp. 327-335, doi: 10.1109/CCGrid54584.2022.00042.

3) T. Hassner, Y. Itcher and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 2012, pp. 1-6, doi: 10.1109/CVPRW.2012.6239348.

4) K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, “Crowd Violence Detection from Video Footage,” 2021 International Conference on Content-Based Multimedia Indexing (CBMI), Jun. 2021, doi: https://doi.org/10.1109/cbmi50038.2021.9461921.

5) K. E. Abdelfatah, G. Terejanu, and A. A. Alhelbawy, “Unsupervised Detection of Violent Content in Arabic Social Media,” Fourth International Conference on Computer Science and Information Technology, pp. 01–07, Mar. 2017, doi: https://doi.org/10.5121/csit.2017.70401

6) T. Giannakopoulos, A. Pikrakis and S. Theodoridis, "A Multimodal Approach to Violence Detection in Video Sharing Sites," 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 3244-3247, doi: 10.1109/ICPR.2010.793.

7) C. Yi, S. Wu, B. Xi, D. Ming, Y. Zhang, and Z. Zhou, “Terrorist Video Detection System Based on Faster R-CNN and LightGBM,” Proceedings of the 4th International Conference on Computer Science and Application Engineering, pp. 1–8, Oct. 2020, doi: https://doi.org/10.1145/3424978.3425121.

8) A. J. Naik and M. T. Gopalakrishna, “Automated Violence Detection in Video Crowd Using Spider Monkey-Grasshopper Optimization Oriented Optimal Feature Selection and Deep Neural Network,” Journal of Control Automation and Electrical Systems, vol. 33, no. 3, pp. 858–880, Jan. 2022, doi: https://doi.org/10.1007/s40313-021-00868-w.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

9) A. Mumtaz, A. B. Sargano, and Z. Habib, “Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning,” 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), Dec. 2018, doi: https://doi.org/10.1109/eecs.2018.00109.

10) I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa, and H. S. Nawaz, “A Frame-Based Feature Model for Violence Detection from Surveillance Cameras Using ConvLSTM Network,” 2021 6th International Conference on Image, Vision and Computing (ICIVC), Jul. 2021, doi: https://doi.org/10.1109/icivc52351.2021.9526948.

11) M. M. Moaaz and E. H. Mohamed, “Violence detection in surveillance videos using deep learning,” FCI-H Informatics Bulletin, vol. 2, no. 2, pp. 1–6, 2020, doi: 10.21608/fcihib.2020.42233.1003.

12) M. Khan, M. A. Tahir and Z. Ahmed, "Detection of Violent Content in Cartoon Videos Using Multimedia Content Detection Techniques," 2018 IEEE 21st International Multi-Topic Conference (INMIC), Karachi, Pakistan, 2018, pp. 1-5, doi: 10.1109/INMIC.2018.8595563.

13) Akash, H.S., Rahim, M.A., Miah, A.S.M., Lee, H.S., Jang, S.W. and Shin, J., 2024. Two-stream modality-based deep learning approach for enhanced two-person human interaction recognition in videos. Sensors, 24(21), p.7077. DOI: https://doi.org/10.3390/s24217077.

14) R. M. Alaqil, J. A. Alsuhaibani, B. A. Alhumaidi, R. A. Alnasser, R. D. Alotaibi and H. Benhidour, "Automatic Gun Detection from Images Using Faster R-CNN," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 149-154, doi: 10.1109/SMART-TECH49988.2020.00045.

15) M. M. Fernandez-Carrobles, O. Deniz, and F. Maroto, “Gun and Knife Detection Based on Faster R-CNN for Video Surveillance,” Pattern Recognition and Image Analysis, pp. 441–452, 2019, doi: https://doi.org/10.1007/978-3-030-31321-0_38.

16) J. Yu, W. Song, G. Zhou, and J. Hou, “Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation,” Multimedia Tools and Applications, vol. 78, no. 7, pp. 8497–8512, Dec. 2018, doi: https://doi.org/10.1007/s11042-018-6923-3.

17) U. Ihsan, N. Z. Jhanjhi, H. Ashraf, F. Ashfaq, and F. A. Wicaksana, “A real-time intelligent surveillance system for suspicious behavior and facial emotion analysis using YOLOv8 and DeepFace,” Engineering Proceedings, vol. 107, no. 1, Art. no. 59, 2025. doi: 10.3390/engproc2025107059.

18) W.-H. Cheng, W.-T. Chu, and J.-L. Wu, “Semantic context detection based on hierarchical audio models,” in Proc. 5th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval (MIR ’03), Berkeley, CA, USA, 2003, pp. 109–115. doi: 10.1145/973264.973282.

19) T. Giannakopoulos, A. Makris, D. I. Kosmopoulos, S. J. Perantonis, and S. Theodoridis, “Audio-Visual Fusion for Detecting Violent Scenes in Videos,” Lecture Notes in Computer Science, vol. 6040, pp. 91–100, May 2010, doi: https://doi.org/10.1007/978-3-642-12842-4_13.

20) A. Bakhshi, Joaquín García-Gómez, R. Gil-Pita, and S. Chalup, “Violence Detection in Real-Life Audio Signals Using Lightweight Deep Neural Networks,” Procedia computer science, vol. 222, pp. 244–251, Jan. 2023, doi: https://doi.org/10.1016/j.procs.2023.08.162.

21) J. Nam, M. Alghoniemy and A. H. Tewfik, "Audio-visual content-based violent scene characterization," Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), Chicago, IL, USA, 1998, pp. 353-357 vol.1, doi: 10.1109/ICIP.1998.723496.

22) P. M. Sethi, H. Mohapatra, A. K. Dalai, P. B. Landge and S. R. Mishra School, "Deep Learning-Based Violence Detection: A YOLO V7 Approach for Real-World Security Applications," 2025 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2025, pp. 1-8, doi: 10.1109/ASSIC64892.2025.11158209.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

23) P. D. Garje, M. S. Nagmode and K. C. Davakhar, "Optical Flow Based Violence Detection in Video Surveillance," 2018 International Conference on Advances in Communication and Computing Technology (ICACCT), Sangamner, India, 2018, pp. 208-212, doi: 10.1109/ICACCT.2018.8529501.

24) A. Jain and D. K. Vishwakarma, "Deep NeuralNet for Violence Detection Using Motion Features from Dynamic Images," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 826-831, doi: 10.1109/ICSSIT48917.2020.9214153.

25) C. Clarin, J. Dionisio, M. Echavez, and P. Naval, “DOVE: Detection of movie violence using motion intensity analysis on skin and blood,” PCSC, vol. 6, pp. 150–156, 2005.

26) E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, “Violence Detection in Video Using Computer Vision Techniques,” Computer Analysis of Images and Patterns, pp. 332–339, 2011, doi: https://doi.org/10.1007/978-3-642-23678-5_39.

27) M. Y. Chen and A. Hauptmann, “MoSIFT: Recognizing human actions in surveillance videos,” Computer Science Department, no. 929, pp. 1–16, 2009.

28) D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: https://doi.org/10.1023/b:visi.0000029664.99615.94.

29) H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, “Video-based Human Action Recognition using Deep Learning: A Review,” arXiv:2208.03775 [cs], Aug. 2022, Available: https://arxiv.org/abs/2208.03775

30) D. Wu, N. Sharma and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2865-2872, doi: 10.1109/IJCNN.2017.7966210.

31) A. Ilyas and N. Bawany, “Crowd dynamics analysis and behavior recognition in surveillance videos based on deep learning,” Multimedia Tools and Applications, vol. 84, no. 23, pp. 26609–26643, Sep. 2024, doi: 10.1007/s11042-024-20161-7.

32) S. Akti, G. A. Tataroglu, and H. K. Ekenel, “Vision-based Fight Detection from Surveillance Cameras,” 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Nov. 2019, doi: https://doi.org/10.1109/ipta.2019.8936070.

33) Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, “Detecting Violent Scenes in Movies by Auditory and Visual Cues,” Advances in Multimedia Information Processing - PCM 2008, pp. 317–326, 2008, doi: https://doi.org/10.1007/978-3-540-89796-5_33.

34) B. Peixoto, B. Lavi, P. Bestagini, Z. Dias and A. Rocha, "Multimodal Violence Detection in Videos," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 2957-2961, doi: 10.1109/ICASSP40776.2020.9054018.

35) T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-Visual Fusion for Detecting Violent Scenes in Videos," in Artificial Intelligence: Theories, Models and Applications, S. Konstantopoulos, S. Perantonis, V. Karkaletsis, C.D. Spyropoulos, and G. Vouros, Eds. Berlin, Heidelberg: Springer, 2010, pp. 132–141 (Lecture Notes in Computer Science, vol. 6040). doi: 10.1007/978-3-642-12842-4_13.

36) I. Serrano, O. Deniz, J. L. Espinosa-Aranda and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," in IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 4787-4797, Oct. 2018, doi: 10.1109/TIP.2018.2845742.

37) D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4489-4497, doi: 10.1109/ICCV.2015.510.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 12:2025
DOI: 10.5281/zenodo.17896835

38) C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, “Violence Detection in Video by Using 3D Convolutional Neural Networks,” Advances in Visual Computing, pp. 551–558, 2014, doi: https://doi.org/10.1007/978-3-319-14364-4_53.

39) P. Wang, P. Wang, and E. Fan, “Violence detection and face recognition based on deep learning,” Pattern Recognition Letters, vol. 142, pp. 20–24, Feb. 2021, doi: https://doi.org/10.1016/j.patrec.2020.11.018.

40) Z. Meng, J. Yuan, and Z. Li, “Trajectory-Pooled Deep Convolutional Networks for Violence Detection in Videos,” Lecture notes in computer science, pp. 437–447, Jan. 2017, doi: https://doi.org/10.1007/978-3-319-68345-4_39.

41) S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

42) S. Liu and W. Deng, "Very deep convolutional neural network-based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 2015, pp. 730-734, doi: 10.1109/ACPR.2015.7486599.

43) F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

44) M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky and D. Khattab, "Violence recognition from videos using Deep Learning Techniques," 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019.

45) Deshpande, G. D., T. D. Devare, and A. G. Gunjal. "Comparative study: Enhancing crime event classification in video surveillance through ResNet50 and mobileNetV2 analysis." In Intelligent Computing and Communication Techniques, pp. 622-627. CRC Press, 2025. DOI: https://doi.org/10.1201/9781003530176

46) Muhammad, S.S. and Alrikabi, J.M., 2024. Fire detection by using densenet 201 algorithm and surveillance cameras images. Journal of Al-Qadisiyah for Computer Science and Mathematics, 16(1), pp.81-91. DOI: https://doi.org/10.29304/jqcsm.2024.16.11437

47) Imah, E.M. and Puspitasari, R.D.I., 2024. Violent crowd flow detection from surveillance cameras using deep transfer learning–gated recurrent unit. ETRI Journal, 46(4), pp.671-682. DOI: https://doi.org/10.4218/etrij.2023-0222

48) de Oliveira Lima, J.P. and Figueiredo, C.M.S., 2021. A temporal fusion approach for video classification with convolutional and LSTM neural networks applied to violence detection. Inteligencia Artificial, 24(67), pp.40-50. DOI: https://doi.org/10.4114/intartif.vol24iss67pp40-50

49) Abdali, A.R., 2021, July. Data efficient video transformer for violence detection. In 2021 IEEE international conference on communication, networks and satellite (COMNETSAT) (pp. 195-199). IEEE. DOI: 10.1109/COMNETSAT53002.2021.9530829

50) Xu, L., Gong, C., Yang, J., Wu, Q. and Yao, L., 2014, May. Violent video detection based on MoSIFT feature and sparse coding. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 3538-3542). IEEE.

51) Dong, Z., Qin, J. and Wang, Y., 2016, October. Multi-stream deep networks for person-to-person violence detection in videos. In Chinese Conference on Pattern Recognition (pp. 517-531). Singapore: Springer Singapore.