*Computers, Materials & Continua*
DOI:10.32604/cmc.2021.018103
***Article***

Tech Science Press

# Real-Time Violent Action Recognition Using Key Frames Extraction and Deep Learning

**Muzamil Ahmed[1,2], Muhammad Ramzan[3,4], Hikmat Ullah Khan[2], Saqib Iqbal[5], Muhammad Attique Khan[6], Jung-In Choi[7], Yunyoung Nam[8,*] and Seifedine Kadry[9]**

[1]Deparment of Computer Science and Information Technology, The University of Lahore, Sargodha Campus, Sargodha, 40100, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt, 47040, Pakistan
[3]School of System and Technology, University of Management and Technology, Lahore, 54782, Pakistan
[4]Department of Computer Science and Information Technology, University of Sargodha, Sargodha, 40100, Pakistan
[5]College of Engineering, Al Ain University, Al Ain, United Arab Emirates
[6]Department of Computer Science, HITEC University Taxila, Taxila, Pakistan
[7]Applied Artificial Intelligence, Ajou University, Suwon, Korea
[8]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Korea
[9]Department of Mathematics and Computer Science, Faculty of Science, Beirut Arab University, Lebanon
[*]Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr
Received: 24 February 2021; Accepted: 24 April 2021

**Abstract:** Violence recognition is crucial because of its applications in activities related to security and law enforcement. Existing semi-automated systems have issues such as tedious manual surveillances, which causes human errors and makes these systems less effective. Several approaches have been proposed using trajectory-based, non-object-centric, and deep-learning-based methods. Previous studies have shown that deep learning techniques attain higher accuracy and lower error rates than those of other methods. However, the their performance must be improved. This study explores the state-of-the-art deep learning architecture of convolutional neural networks (CNNs) and inception V4 to detect and recognize violence using video data. In the proposed framework, the keyframe extraction technique eliminates duplicate consecutive frames. This keyframing phase reduces the training data size and hence decreases the computational cost by avoiding duplicate frames. For feature selection and classification tasks, the applied sequential CNN uses one kernel size, whereas the inception v4 CNN uses multiple kernels for different layers of the architecture. For empirical analysis, four widely used standard datasets are used with diverse activities. The results confirm that the proposed approach attains 98% accuracy, reduces the computational cost, and outperforms the existing techniques of violence detection and recognition.

**Keywords:** Violence detection; violence recognition; deep learning; convolutional neural network; inception v4; keyframe extraction

## 1 Introduction

The recognition of human activities from surveillance videos has become an active and progressive research area in computer vision and machine learning [1,2]. The classification of media content in the form of videos is based on human action, which depicts general human behavior. Human behavior and actions are understood based on different video features that classify actions as normal or abnormal [3]. All activities of everyday lives, including walking, running on the ground, eating food, sitting down and rising from a chair, lying in bed, picking an item from a table or floor, and descending stairs, are called normal activities [4,5]. Abnormal activities, also called suspicious activities, deviate from normal human activities. The actions are abnormal for one scenario but may be considered normal for another scenario [6]. For example, running on a playground is normal, but running in a bank or a marketplace is considered abnormal [7]. The most crucial and significant abnormal activities are violent activities that physically depict actions to cause harm or damage with aggressive behaviors. Fighting, killing, and beating someone are the most common examples of violence in public places [8].

In a semi-automated system, violent activities are monitored manually through the monitor screen of a surveillance camera. This is not beneficial because continuous monitoring is required, but watching screens continuously to recognize violent activities is difficult. There is no scope for carelessness while monitoring such activities because these can occur at any time [9,10]. There is a need to transform such semi-automated systems into fully automated intelligent systems that can detect and recognize violent activities without human supervision [11]. Fully automated systems can detect human activity through computer vision and machine learning and are more effective and efficient in detecting object movements and recognizing human activity as compared to semi-automated systems [12,13]. Human activity recognition is a difficult task because of many factors such as real-time classification, low video quality of surveillance cameras, and inconsistent light intensity during monitoring [14].

This study proposes a fully automated system for violent action recognition. The main contributions of this study are as follows:

- Keyframe extraction to eliminate duplicate frames from a video
- Application of two convolutional neural network (CNN) architectures, Sequential and Inception v4 CNN, for feature selection and classification
- Preparation of violent activities dataset for the training of classification models
- Comparison of the proposed framework with state-of-the-art models, violent flow (Vif), CNN Hough Forest, BoW (MOSIFT), and Conv LSTM, on three benchmark datasets

The rest of the paper is organized as follows. Section 2 reviews the existing research on violent activity detection using machine learning models. Section 3 presents the research methodology and architecture of the proposed technique. Section 4 discusses the preparation of the video dataset. Section 5 describes the experimental setup and the results. Section 6 presents the conclusions and outlook for future research.

## 2 Related Work

This section discusses previous studies on video action detection [15]. Fully automatic violence detection methods can be grouped as trajectory-based, non-object-centric, and deep-learning-based techniques [16,17].

### *2.1 Trajectory-Based Methods for Violence Detection*

Trajectory features are widely used for detecting human activities. These features contain information related to the object movements in the foreground. Trajectory-based methods involve two phases. The first phase involves the motion estimation of objects using a statistical model and extracts the trajectory features of the video. In the second phase, the activity is recognized based on the extracted features [18]. The fight action recognition framework proposed in [19] used a bag of words for feature extraction and K-nearest neighbors for classification. The model achieved an accuracy of 86% using the k-th video dataset. The main disadvantage of a bag of words is that it assumes that all words are independent. Another trajectory-based approach was proposed [20] to detect violent activity from videos, wherein the Gaussian mixture method was used to extract three trajectory features: object direction, speed, and centroid. An accuracy of 90% was achieved using a rule-based classifier. However, complex timing rules were required for massive video data. The violent activity detection framework proposed in [21] used region vector motion for feature extraction and an SVM support vector machine for classification. The authors achieved 96% accuracy on the movies video dataset. Another trajectory method [22] used a transfer-learning technique. The authors used animal fight data by extracting trajectory features using local motion features, LMF, and SVM for classification, and they achieved an accuracy of 85%. The violence detection framework was proposed using motion boundary histograms for video feature extraction and SVM for classification. An accuracy of 89% was achieved using the hockey fight dataset. Tab. 1 lists the different trajectory-based approaches to violence detection and their key aspects.

**Table 1:** Trajectory based methods

| Reference | Feature extraction | Classification | Accuracy % |
|---|---|---|---|
| [19] | Begs of words (BoW) | k-NN | 86 |
| [20] | Gaussian mixture method (GMM) | Rule-based | 90 |
| [21] | Region Vector Motion (RVM) | SVM | 96 |
| [18] | Motion boundary histograms (MBH) | SVM | 89 |
| [22] | Local motion feature (LMF) | SVM | 85 |

### *2.2 Non-object Centric Based Methods for Violence Detection*

In non-object-centric methods, video features are extracted based on object behavior rather than object motion. These methods are more complex than trajectory-based methods because of the low-level representation of video features. Non-object-centric methods deal with spatial-temporal (space and time) contexts while extracting the features [23,24]. These methods involve descriptors for low-level representations of video features and use 2D cells or 3D cubes of each frame interest point. After feature description, classifiers are used for classification [25,26].

A framework proposed in [27] for violence action detection used an optical histogram flow (HOG) invariant for feature extraction and description and employed rotation-invariant motion coherence (RIMOC) for classification. A 93% accuracy was achieved using a violent flow dataset. Another non-object-centric violence-detection framework [28] used the Gaussian model for optical flow to extract low-level features. SVM was used as a classifier, and an 89% accuracy was achieved

using a crowded violence video dataset. The framework proposed in [29] detected abnormal activities using OMEGA equations for features and descriptions and used SVM as a classifier. The authors achieved an accuracy of up to 90%. Another abnormal activity detection framework proposed in [30] used Gaussian and fuzzy K-mean approaches for video feature extraction and description. The authors used K nearest neighbor for classification and achieved a 95% accuracy. A motion blob was used for feature description, and SVM was used as a classifier. A 92% accuracy was achieved using the BEHAVE and CAVIAR datasets. Tab. 2 lists the trajectory-based approaches used to detect violence.

**Table 2:** Non-Object-centric based methods

| Reference | Feature extraction | Classification | Accuracy % |
|---|---|---|---|
| [27] | histogram optical flow | Rotation-Invariant Motion Coherence (RIMOC) | 93% |
| [28] | Gaussian Model for optical flow (GMOF) | SVM | 89% |
| [29] | OMEGA equations | SVM | 90% |
| [30] | Gaussian and fuzzy K mean | K-NN | 95% |
| [25] | Motion blob | SVM | 92% |

### *2.3 Deep Learning Based Methods for Violence Detection*

Deep learning methods attain high accuracy over trajectory-based and non-object-centric methods for video activity detection. Deep learning models treat feature selection and classification as a single module [31,32]; there is no need to use feature extractors or descriptors separately. Deep learning techniques have gained more attention and popularity than other techniques to resolve the challenges stated in [33]. Unsupervised learning techniques, including deep belief networks, recurrent neural networks, CNNs, and long short-term memory (LTSM), are used as deep learning methods for activity recognition [34].

The deep learning framework for recognizing abnormal human actions in [35] used a recurrent neural network and achieved an accuracy of 91.43%. Another deep learning framework [36] used LSTM to recognize violent activity. Three benchmark datasets, namely the movie dataset, hockey fight dataset, and violent flow dataset were used, and an accuracy of 94% was achieved. A simple deep neural network framework was proposed to detect violent activities using the Weber local descriptor to extract optical flow [37]. A 90% accuracy was achieved using a crowded violence dataset. Another deep learning framework [38] used CNN Bi-LSTM to detect violent activities from videos. They attained 94% accuracy using three widely used datasets of hockey fights, movies, and violent flow. Tab. 3 lists the different trajectory-based approaches for violence detection.

**Table 3:** Deep learning-based methods

| Reference | Classification and feature selection | Conv Layers /FC layers | Accuracy |
|---|---|---|---|
| [39] | Recurrent neural network (RNN) | 3/2 | 91% |
| [36] | Long short-term memory (LSTM) | 5/3 | 94% |
| [37] | Simple deep neural network | 4/1 | 90% |
| [38] | CNN Bi LSTM | 3/2 | 94% |

## 3 Research Methodology

This section discusses the proposed framework, deep learning models, and their architectures. The proposed framework for violence detection is shown in Fig. 1. First, the framework takes the video sequence as input and generates frames (5 frames per second (fps)). Subsequently, the keyframe extraction technique is used to eliminate consecutive duplicate frames. These extracted frames are used for training the deep learning models. The sequential CNN and inception v4 deep learning architectures were used for feature selection and classification.

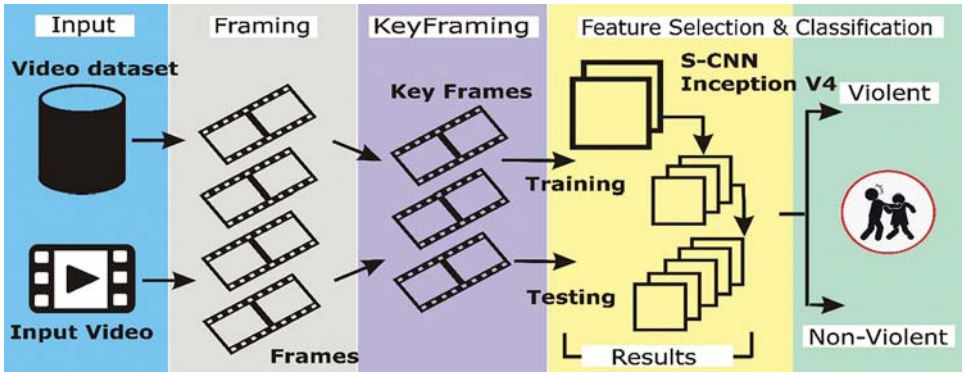


**Figure 1: :** The proposed framwork for violence detection using deep learning based technique

### *3.1 Key Frame Extraction*

In existing violent detection approaches, all frames extracted from the video are used for training purposes. In a normal video sequence, many consecutive frames are duplicated. These consecutive duplicate frames increase the complexity and computational cost of the model. In this study, the keyframe extraction technique eliminates identical successive frames. Thus, keyframe extraction reduces the number of training frames and the computational cost of processing duplicate frames [40].

**Algorithm 1:** Keyframe extraction

Input: List of all frames
Output: Keyframes
1. List of keyframes ( )
2. The previous frame = Read new frame

(Continued)

3. Append the first frame in the List of Keyframes
4. WHILE the end of frames
5. Current frame = Read new frame
6. Difference = absolute difference (current frame, previous frame)
7. Count non zero from difference
8. if non zero count >= frame threshold
9. than append frame in List of Keyframes
10. Previous frame = current frame
11. WHILE END
12. Return List of extracted frames

Algorithm 1 shows the pseudocode for keyframe extraction. All frames extracted from the video data are inputted in the algorithm, and the function returns the list of keyframes. The first frame of the video is considered a keyframe and added to the list of keyframes. The next frame is compared with the previous frame, and the similarity between two consecutive frames is computed. This similarity is based on the absolute difference between two frames, which is determined as a non-zero value using a simple matrix subtraction method. The non-zero value is compared with a threshold value. The threshold, also called the binary decision threshold, has two regions: above the threshold and below the threshold. Values below the threshold indicate the same frames, and those above the threshold are considered as keyframes.

### *3.2 Features Selection and Classification*

In the violent activity detection framework, the next task is feature selection and classification. In trajectory-based and non-object-centric approaches, the task of classification and feature extraction is considered as two different modules. In deep learning methods, these are combined into a single module. In this study, sequential CNN and inception V4 networks were used for feature selection and classification.

#### *3.2.1 Sequential CNN Architecture*

As shown in Fig. 2, the sequential CNN architecture consists of three convolutional layers with a size of $64 \times 64 \times 3$. It uses the rectified layer unit as an activation function in these layers after the convolutional process max-pooling, which realizes the network's spatial variance property. Max-pooling is used to provide an abstract form of representation and avoid overfitting. In addition, it reduces the computational cost by reducing the number of parameters. The stride size also refers to the pool size $(2 \times 2)$ for all max-pooling functions in the entire network. After the third convolutional layer, the pooling function adds a flattening function that is used to convert the frame pixel into a vector column.

The proposed model uses two fully connected layers after flattening. The dense function is used in both fully connected layers, but both function parameters are changed. In the first fully connected layer, 128 units and rectified layer units are used as activation functions. In the second layer, only one unit with a sigmoid activation function is used. The last fully connected layer predicts the class of the input frames. After adding all the functions into a sequential model, the call model compiles the function using three parameters: optimizer, loss, and metrics. Adam optimizer is used to iteratively update the weights during data training. Subsequently, binary cross-entropy measures the loss and accuracy as evolution metrics for evaluation.

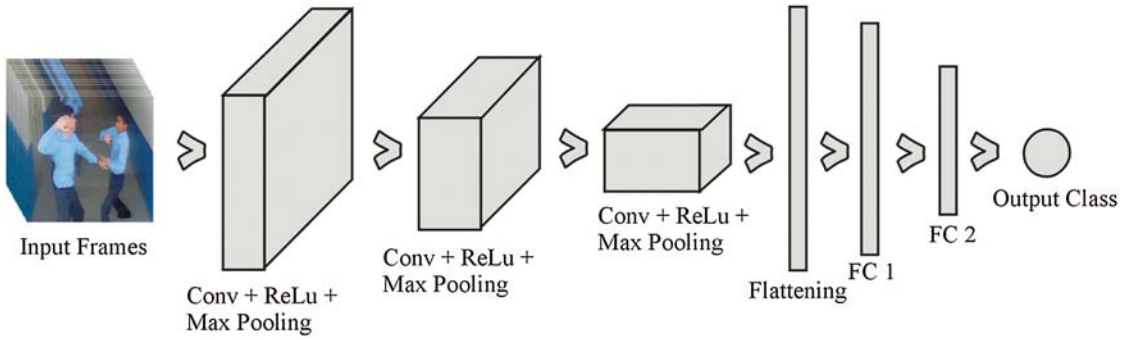


**Figure 2:** The proposed architecture of sequential CNN for violence detection

*3.2.2 CNN Inception V4 Architecture*

Tab. 4 illustrates the Inception v4 network architecture for recognizing violent activities from the video sequence. Inception is a deep architecture of CNN, wherein multidimensional convolutional layers are used in parallel. The Inception v4 architecture comprises four inception blocks, namely A, B, C, and base. Inception blocks A and B are followed by two reduction blocks, A and B. In inception block A, the input is divided into four branches, B0 to B3, and each branch has convolutional layers. Tab. 4 lists all blocks with their branches and the size of the convolutional layers. After merging all the blocks' outputs, the flatting function and fully connected layers predict the output class of the video frames.

**Table 4:** The proposed architecture of CNN Inception v4 for violence detection

| Block | Branches | Conv layers | Layers dimension |
|---|---|---|---|
| Inception A | B0 | 1 | (96, 1, 1) |
| | B1 | 2 | (64, 1, 1) (96, 3, 3) |
| | B2 | 3 | (64, 1, 1) (96, 3, 3) (96, 3, 3) |
| | B3 | 1 | (96, 1, 1) |
| Reduction A | B0 | 1 | (384, 3, 3) |
| | B1 | 3 | (192, 1, 1) (223, 3, 3) (256, 3, 3) |
| | B2 | 1 | max pooling layer ($3 \times 3$) |
| Inception B | B0 | 1 | (384, 1, 1) |
| | B1 | 3 | (192, 1, 1) (224, 1, 1) (256, 7, 1) |
| | B2 | 5 | (192, 1, 1) (192, 7, 1) (224, 1, 7) (224, 7, 1) (256, 1, 7) |
| | B3 | 2 | (128, 1, 1) and one average pooling layer ($3 \times 3$) |
| Reduction B | B0 | 2 | (192, 1, 1) (192, 3, 3) |
| | B1 | 4 | (256, 1, 1) (256, 1, 7) (320, 7, 1) (320, 3, 3) |
| | B2 | 1 | max pooling layer ($3 \times 3$) |

(Continued)

**Table 4:** Continued

| Block | Branches | Conv layers | Layers dimension |
|---|---|---|---|
| Inception C | B0 | 1 | (256, 1, 1) |
| | B1 | 2 | (256, 1, 1) (384, 1, 1), |
| | B10 | 1 | (512, 1, 3) |
| | B11 | 1 | (256, 1, 1) |
| | B2 | 3 | (384, 1, 1) (448, 3, 1) (512, 1, 3) |
| | B20 | 1 | (256, 1, 3) |
| | B21 | 1 | (256, 1, 3) |
| | B3 | 1 | average pooling layer ($3 \times 3$) |
| Inception base block | B0 | 1 | max-pooling layers ($3 \times 3$) |
| | B1 | 1 | (96, 3, 3) |
| | B0 | 2 | (64, 1, 1) (96, 3, 3) |
| | B1 | 4 | (64, 1, 1) (64, 1, 7) (64, 7, 1) (96, 3, 3) |
| | B0 | 1 | (192, 3, 3) |
| | B1 | 1 | max-pooling layers ($3 \times 3$) |

## 4 Datasets

The performance of a classification model also depends on the quality of the learning content. For image classification, we used an image dataset that contained images of each class to train the classification model. Four video datasets, namely, hockey fights, violent crowd detection, movies, and BEHAVE, are widely used for violence detection [41]. These datasets contain videos collected from different sources, such as the fight and non-fight actions of movies, fight scenes in national hockey matches, self-made videos, and videos collected from social media, and the implementation of the surveillance place was neglected. These datasets are more general and do not target specific public places such as markets, highways, banks, and educational institutes. Violent recognition systems target surveillance cameras placed in public places. However, in the BEHAVE dataset, the angle of the camera for capturing the videos is considered similar to that of surveillance cameras. However, dataset videos are extremely long and contain both violent and non-violent activities in a single video overhead during the training of the model.

Another major contribution of this study is the preparation of a dataset for violence detection. The dataset focuses on the violent activities of students in educational institutes. Surveillance cameras are placed in educational institutes to track and monitor students' activities. In this study, we collected these videos from CCTV cameras installed in educational institutes. Different possible violent and non-violent actions performed by students were recorded to maintain the training quality of the dataset. In this step, the distance of the object from the camera and the camera angle are considered such that the video is recorded with sufficient light intensity. In preprocessing, a surveillance camera recording is first converted into a normal video format. Surveillance cameras record videos in 'dav' format, which cannot be used directly for training. Subsequently, the level of lightness, hue, and saturation was adjusted for all videos. The videos were split into durations of 3 s. The dataset contains 320 videos divided into 172 videos of the violent class and 148 videos of the non-violent class (Fig. 3).

(a) (b)

**Figure 3:** Sample frames from education institutes violence detection video dataset (a) frames of non-violent class (b) frames of violent class

## 5 Experimental Setup and Results Discussion

This section describes the experimental setup and results of the proposed framework for violence detection from video sequences. The implementation of the framework was accomplished in Python. The deep learning architecture sequential CNN and inception v4 used the Keras open-source library and tensor flow as the backend. In the experiment, the keyframe extraction technique was implemented on four video datasets. The frame rate was 5 fps, and the adjusted threshold for keyframing was 300000 for all datasets. Tab. 5 presents the results of the keyframe extraction technique; the last column presents the number of eliminated frames for each dataset, which is approximately 25% of all frames in the dataset. These results indicate that many frames are not necessary for the training of the classification model; these frames are generally not required for training, but their inclusion in the training increases the processing time. These eliminated frames save computational time, which reduces the complexity of the classification technique.

**Table 5:** Results of keyframe extraction technique on benchmark violence datasets

| Video Dataset | Total Videos | Total Frames | Key Frames | Eliminated Frames |
|---|---|---|---|---|
| Movies Dataset | 201 | 2512 | 703 | 1809 |
| Crowd Violence Dataset | 248 | 3100 | 868 | 2232 |
| BEHAVE Dataset | 4 | 44400 | 12432 | 31968 |
| Hockey Fight Dataset | 1000 | 11352 | 9000 | 2352 |
| EIVD Dataset | 320 | 40000 | 1120 | 2880 |

After implementing keyframe extraction, sequential CNN and Inception v4 were used for feature selection and classification. For an empirical analysis, the proposed framework was applied to the four video datasets for evaluation. We used accuracy as an evaluation metric to measure the classification model's performance. The accuracy of the model was calculated using Eq. (1). Accuracy refers to the correct identification of a portion of the entire prediction. In Eq. (1), TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative,

respectively. Four parameters were used to compile the classification model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

These parameters include the training data, validation or testing data, number of epochs, and steps per epoch. In this framework, we split the video data with a training : testing ratio of 0.3. The number of epochs was adjusted to 25, with 8000 steps per epoch. In Inception, the number of network steps per epoch depends on the number of training frames. Tab. 6 presents the results of the proposed CNN architecture in terms of recognizing violent activities from an input video sequence.

**Table 6:** The accuracy of sequential CNN on the state of the art violence detection dataset

| No of Epochs | Movies % | Hockey Fight % | CVD % | EIVD % |
|---|---|---|---|---|
| 1 | 96.52 | 86.70 | 79.37 | 85.00 |
| 5 | 96.71 | 93.70 | 86.28 | 91.65 |
| 10 | 97.40 | 94.35 | 94.00 | 95.12 |
| 15 | 97.91 | 95.10 | 94.71 | 96.97 |
| 20 | 98.20 | 97.60 | 95.22 | 98.66 |
| 25 | 98.40 | 98.12 | 97.40 | 99.20 |

Tabs. 6 and 7 show the accuracy of the proposed framework for the violence detection dataset. Fig. 4 shows that the accuracy of the sequential CNN and Inception v4 increase with the number of epochs. However, in this classification, the models attained a higher accuracy with a smaller number of epochs.

**Table 7:** The accuracy of inception v4 CNN on the state of the art violence detection dataset

| No of Epochs | Movies % | Hockey Fight % | CVD % | EIVD % |
|---|---|---|---|---|
| 1 | 88.33 | 94.52 | 83.32 | 92.11 |
| 5 | 93.23 | 95.66 | 86.76 | 93.71 |
| 10 | 95.00 | 97.45 | 93.43 | 95.65 |
| 15 | 96.90 | 97.78 | 94.91 | 96.23 |
| 20 | 97.89 | 98.43 | 95.80 | 97.32 |
| 25 | 99.03 | 98.11 | 97.65 | 98.55 |

A comparison between the proposed model and existing studies is shown in Tab. 8. In existing studies, datasets of videos from movies, hockey fights, and CVD have been commonly used for violence detection but with different approaches. The proposed model achieves a higher accuracy for violence detection as compared to other models.

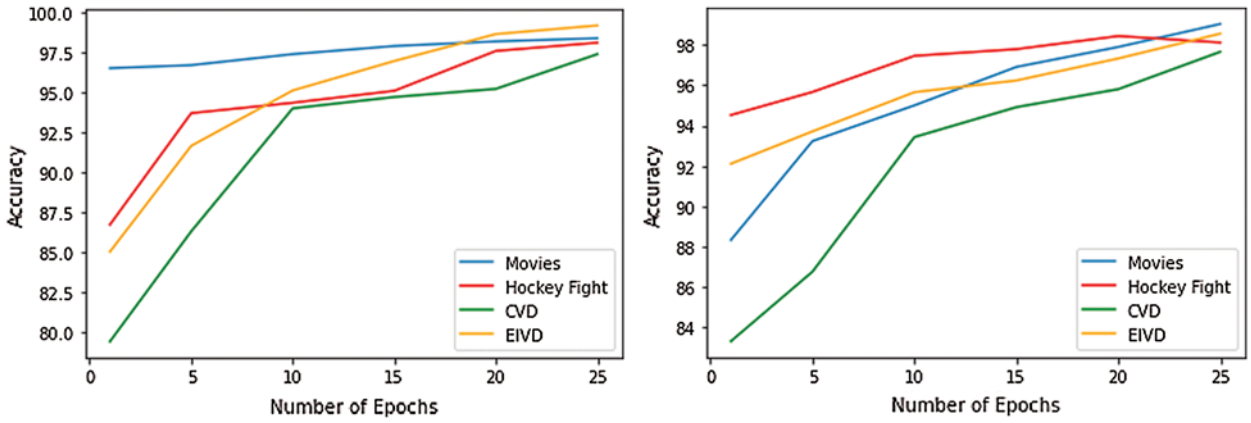


**Figure 4:** The accuracy of proposed model with respect to number of epochs on state of art violence detection datasets (a) Sequential CNN (b) Inception v4 CNN

**Table 8:** The comparison of the proposed framework with state of the art violence detection techniques on movies, hockey fight, and CVD violence detection video dataset

| Methodology | Movies % | Hockey Fight % | CVD % |
|---|---|---|---|
| ViF (Violent Flow) | 82.40 | 88.90 | 82.00 |
| ViF+ OViF (Optical Violent Flow) | 87.50 | 96.70 | 84.00 |
| ADMN | 89.00 | - | - |
| CNN Hough Forest | 94.00 | - | 93.25 |
| Bi Channel CNN | 95.90 | - | 93.00 |
| BOW(MOSIFT) | 96.50 | - | - |
| Convolutional LTSM | 97.50 | 98.00 | - |
| Substantial Derivation | - | 97.00 | - |
| MoSIFT+HIK | - | 96.89 | - |
| MOIWLD+SRC | - | - | 93.00 |
| Deep Neural Network using WLD | - | - | 94.00 |
| **Proposed S-CNN** | **98.19** | **98.40** | **97.40** |
| **Proposed Inception v4 CNN** | **99.11** | **99.03** | **97.65** |

## 6 Conclusion

This study proposed a deep learning framework for recognizing violent activity from a video. The proposed framework used the keyframe extraction technique to eliminate duplicate frames and employed S-CNN and inception v4 CNN for feature selection and classification. Detailed experiments were performed to validate the proposed model. The results show that keyframe extraction eliminates up to 25% of duplicate frames. The classification model attained an accuracy of approximately 98%. Thus, sequential CNN and inception v4 are more effective in detecting violent activities from videos. The proposed technique will be used to recognize other abnormal activities in a future study.

**Funding Statement:** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A1B07042967) and the Soonchunhyang University Research Fund.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] T. Akram, M. Raza, T. Saba and A. Rehman, "Hand-crafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition," *Applied Soft Computing,* vol. 87, pp. 105986, 2020.

[2] K. Aurangzeb, I. Haider, T. Saba, K. Javed, T. Iqbal *et al.*, "Human behavior analysis based on multi-types features fusion and von nauman entropy based features reduction," *Journal of Medical Imaging and Health Informatics,* vol. 9, pp. 662–669, 2019.

[3] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail *et al.*, "A review on state-of-the-art violence detection techniques," *IEEE Access,* vol. 7, pp. 107560–107575, 2019.

[4] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria and A. F. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," *Applied Artificial Intelligence,* vol. 34, pp. 329–344, 2020.

[5] A. Sharif, K. Javed, H. Gulfam, T. Iqbal, T. Saba *et al.*, "Intelligent human action recognition: A framework of optimal features selection based on Euclidean distance and strong correlation," *Journal of Control Engineering and Applied Informatics,* vol. 21, pp. 3–11, 2019.

[6] M. Rashid, M. Raza, M. M. Sarfraz and F. Afza, "Object detection and classification: A joint selection and fusion strategy of deep convolutional neural network and SIFT point features," *Multimedia Tools and Applications,* vol. 78, pp. 15751–15777, 2019.

[7] R. Brito, R. P. Biuk-Aghai and S. Fong, "GPU-Based parallel shadow features generation at neural system for improving gait human activity recognition," *Multimedia Tools and Applications,* vol. 3, pp. 1–16, 2021.

[8] P. Zhou, Q. Ding, H. Luo and X. Hou, "Violence detection in surveillance video using low-level features," *PLoS ONE,* vol. 13, pp. e0203668, 2018.

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

[10] M. Raza, M. Sharif, M. Yasmin, T. Saba and S. L. Fernandes, "Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning," *Future Generation Computer Systems,* vol. 88, pp. 28–39, 2018.

[11] G. Tripathi, K. Singh and D. K. Vishwakarma, "Violence recognition using convolutional neural network: A survey," *Journal of Intelligent & Fuzzy Systems,* vol. 11, pp. 1–22, 2020.

[12] S. Siddiqui, K. Bashir, F. Azam and M. Y. Javed, "Human action recognition: A construction of code-book by discriminative features selection approach," *International Journal of Applied Pattern Recognition,* vol. 5, pp. 206–228, 2018.

[13] M. Sharif, T. Akram, M. Y. Javed, T. Saba and A. Rehman, "A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection," *EURASIP Journal on Image and Video Processing,* vol. 2017, pp. 1–18, 2017.

[14] D. Song, C. Kim and S. -K. Park, "A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance," *Information Sciences,* vol. 447, pp. 83–103, 2018.

[15] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib *et al.*, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimedia Tools and Applications,* vol. 9, pp. 1–27, 2020.

[16] M. A. Khan, T. Akram, M. Sharif, N. Muhammad and S. R. Naqvi, "Improved strategy for human action recognition; experiencing a cascaded design," *IET Image Processing,* vol. 14*,* pp. 818–829, 2019.
[17] F. Zahid, J. H. Shah and T. Akram, "Human action recognition: A framework of statistical weighted segmentation and rank correlation-based selection," *Pattern Analysis and Applications,* vol. 23*,* pp. 281–294, 2020.
[18] B. Boufama, P. Habashi and I. S. Ahmad, "Trajectory-based human activity recognition from videos," in *2017 Int. Conf. on Advanced Technologies for Signal and Image Processing*, Fez, Moroco, pp. 1–5, 2017.
[19] A. A. Mishra and G. Srinivasa, "Automated detection of fighting styles using localized action features," in *2018 2nd Int. Conf. on Inventive Systems and Control*, Coimbatore, India, pp. 1385–1389, 2018.
[20] W. Lejmi, A. B. Khalifa and M. A. Mahjoub, "Fusion strategies for recognition of violence actions," in *2017 IEEE/ACS 14th Int. Conf. on Computer Systems and Applications*, Hammamet, Tunisia, pp. 178–183, 2017.
[21] J. Xie, W. Yan, C. Mu, P. Li and S. Yan, "Recognizing violent activity without decoding video streams," *Optik,* vol. 127*,* pp. 795–801, 2016.
[22] E. Y. Fu, M. X. Huang, H. V. Leong and G. Ngai, "Cross-species learning: A low-cost approach to learning human fight from animal fight," in *Proc. of the 26th ACM International Conference on Multimedia*, NY, USA, pp. 320–327, 2018.
[23] K. Deepak, L. Vignesh and S. Chandrakala, "Autocorrelation of gradients based violence detection in surveillance videos," *ICT Express,* vol. 6*,* pp. 155–159, 2020.
[24] N. Hussain, S. A. Khan, A. A. Albesher and T. Saba,"A deep neural network and classical features based scheme for objects recognition: An application for machine inspection," *Multimedia Tools and Applications,* vol. 6*,* pp. 1–23, 2020.
[25] S. Ramachandran and L. H. Palivela, "An intelligent system to detect human suspicious activity using deep neural networks," *Journal of Intelligent & Fuzzy Systems,* vol. 36*,* pp. 4507–4518, 2019.
[26] Y. D. Zhang, S. A. Khan, M. Attique, A. Rehman and S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network," *Multimedia Tools and Applications,* vol. 17*,* pp. 1–23, 2020.
[27] G. Basavaraj and A. Kusagur, "Vision based surveillance system for detection of human fall," in *2017 2nd IEEE Int. Conf. on Recent Trends in Electronics, Information & Communication Technology*, Bangalore, India, pp. 1516–1520, 2017.
[28] M. Lorbach, E. I. Kyriakou, R. Poppe, E. A. van Dam and R. C. Veltkamp, "Learning to recognize rat social behavior: Novel dataset and cross-dataset application," *Journal of Neuroscience Methods,* vol. 300*,* pp. 166–172, 2018.
[29] R. Nar, A. Singal and P. Kumar, "Abnormal activity detection for bank ATM surveillance," in *2016 Int. Conf. on Advances in Computing, Communications and Informatics*, Jaipur, India, pp. 2042–2046, 2016.
[30] E. Y. Fu, H. V. Leong, G. Ngai and S. C. Chan, "Automatic fight detection in surveillance videos," *International Journal of Pervasive Computing and Communications,* vol. 4*,* pp. 1–17, 2017.
[31] H. Arshad, M. I. Sharif, M. Yasmin, J. M. R. Tavares, Y. D. Zhang *et al.*, "A multilevel paradigm for deep convolutional neural network features selection with an application to human gait recognition," *Expert Systems,* vol. 9*,* pp. e12541, 2020.
[32] I. Ashraf, M. Alhaisoni, R. Damaševičius, R. Scherer, A. Rehman *et al.*, "Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists," *Diagnostics,* vol. 10*,* pp. 565, 2020.
[33] T. Zhang, W. Jia, C. Gong, J. Sun and X. Song, "Semi-supervised dictionary learning via local sparse constraints for violence detection," *Pattern Recognition Letters,* vol. 107*,* pp. 98–104, 2018.
[34] M. Baba, V. Gui, C. Cernazanu and D. Pescaru, "A sensor network approach for violence detection in smart cities using deep learning," *Sensors,* vol. 19*,* pp. 1676, 2019.
[35] A. Traoré and M. A. Akhloufi, "Violence detection in videos using deep recurrent and convolutional neural networks," in *2020 IEEE Int. Conf. on Systems, Man, and Cybernetics*, Toronto, Canada, pp. 154–159, 2020.

[36] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Lecce, Italy, pp. 1–6, 2017.

[37] S. Mondal, S. Pal, S. K. Saha and B. Chanda, "Violent/non-violent video classification based on deep neural network," in *2017 Ninth Int. Conf. on Advances in Pattern Recognition*, Bangalore, India, pp. 1–6, 2017.

[38] R. Halder and R. Chatterjee, "CNN-Bilstm model for violence detection in smart surveillance," *SN Computer Science,* vol. 1, pp. 1–9, 2020.

[39] D. Arifoglu and A. Bouchachia, "Activity recognition and abnormal behaviour detection with recurrent neural networks," *Procedia Computer Science,* vol. 110, pp. 86–93, 2017.

[40] Z. Wang and Y. Zhu, "Video key frame monitoring algorithm and virtual reality display based on motion vector," *IEEE Access,* vol. 8, pp. 159027–159038, 2020.

[41] C. Liu, J. Ying, F. Han and M. Ruan, "Abnormal human activity recognition using Bayes classifier and convolutional neural network," in *2018 IEEE 3rd International Conference on Signal and Image Processing*, Shenzhen, China, pp. 33–37, 2018.