ORIGINAL RESEARCH

CNN-BiLSTM Model for Violence Detection in Smart Surveillance

Rohit Halder[1] · Rajdeep Chatterjee[1]

Received: 7 January 2020 / Accepted: 27 May 2020

Abstract

In this paper, a lightweight computational model has been introduced for the better classification of violent and non-violent activities. Nowadays, the occurrences of violent activities get increased in public places. It has various social and economic reasons behind the growth of violent actions. Therefore, the government agencies and public administrators need to check such incidents using smart surveillance. Deep learning-based an efficient violent activity detection model can help the authorities in detecting a violent activity in real-time. The evaluated results can be henceforth sent to store and analyze the captured video to automate the crime monitoring system. Convolutional Neural Network-based Bidirectional LSTM has been used to detect violent activities and also compared with other existing approaches. Our proposed model gives 99.27%, 100% and 98.64% classification accuracies for the widely used standard Hockey Fights, Movies and Violent-Flows video datasets, respectively.

Keywords Deep learning · Convolutional neural network · Detecting video sequences · Image processing · LSTM · Smart society · Social security

Introduction

Violence has always been a serious social issue. There are different causes for the rise of violent activities in public places. Individual's greed, frustration, and hatred, as well as social and economic insecurities, are the major reasons behind an increase in violence. In recent years, it has been witnessed the study of human action behavior under the lights of computer vision and data science. In spite of being the most alarming social issue, there are not many works indulging in the automation of action detection [1], violence detection, protest detection. This field of study has huge applicability as far as social security and stability are concerned. Prevention of crime and violent activities are not possible unless the brain signals are analyzed and detected the specific pattern inferred the criminal thoughts in real-time [2–6]. It is yet to be achieved due to its technical feasibility. However, we can detect violent activities in public places using deep learning-based computer vision. Surveillance cameras are already deployed in most of the public places and private institutes. The efficient violent detection technique can help the government or authorities to take a fast and formalized approach to identify the violence and to prevent the destruction caused to human life and public property. As we all are a part of society, we want safe streets, neighborhoods, and work around us. Deep learning is better than the machine learning technique as it does not require any explicit feature engineering. There are some drawbacks as well, such as high computational cost and large training datasets. These technical aspects motivate us to develop a model that takes less training time as well as a moderate number of training samples.

Our approach involves a Convolutional Neural Network Bidirectional LSTM model (CNN-BiLSTM) architecture to predict violence in the sequential flow of frames. Firstly, we breakdown a video into several frames. We pass each frame through a convolutional neural network, to extract the information present in that current frame. Then we use a Bidirectional LSTM layer to compare the information of the current frame once with the previous frames and once with the upcoming frames to identify any sequential flow of events. Finally, the classifier is used to identify whether an action is violent or not. Hence, this architecture uses spatial

✉ Rajdeep Chatterjee
cse.rajdeep@gmail.com

Rohit Halder
rhaldar9@gmail.com

[1] School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar 751024, Odisha, India

Published online: 12 June 2020

Table 1 Results obtained after fivefold cross-validation

Reference	Used methodology
[14]	CNN-LSTM
[15]	Three streams + LSTM
[16]	Motion Scale-Invariant Feature Transform+HIK
[17]	Violent Flow descriptors (ViF)
[9]	Motion Scale-Invariant Feature Transform + KDE + Sparse Coding
[18]	Substantial Derivative
[19]	Motion Weber Local Descriptor (MoWLD)
[20]	Violent Flow descriptors + Oriented VIolent Flows(ViF + OViF)
[21]	Spatiotemporal Encoder
[22]	Conv 3D
[23]	Space Time Interest Points

features as well as temporal features in both the directions for prediction analysis. The detailed description of the model architecture has been discussed in the later sections.

Our paper has been organized as follows. In “Related Works”, we discuss various algorithms that were used to detect for violent and non-violent activities in the past years. The section is further followed by “Model Architecture”, which gives a detailed description of our approach and proposed model. In “Experimental Setup”, we have given brief descriptions of the used datasets, data preprocessing and training methodology. In “Results and Analysis”, we have elaborately analyzed the obtained results. Finally, we conclude the paper in “Conclusion”.

Related Works

Previously, violent and non-violent activities were recognized using the presence of blood, degree of motion, even characteristics of sound relating to violent activities. The surveillance cameras are not very effective in recording sounds related to certain activities [7]. On the other hand, frame-based video analysis is solely based on a sequence of frames (that is, image) and not on audio. Violence can be categorized into many types, including one to one person violence, crowd violence, family violence, sports violence, violence with guns and many more. The crowd violence has been identified using Latent Dirichlet Allocation (LDA) and Support Vector Machines (SVMs) in [8]. Different methods for violence detection proposed by researchers includes, Motion Scale-Invariant Feature Transform (MoSIFT) [Histogram of Gradients (HoG) + Histogram of optical Flow (HoF)] [9], Harris corner detection [10], Long Short Term Memory (LSTM) cells [11] shown in Fig. 2, Convolutional Neural Networks (CNN), ConvLSTM networks (CNN+LSTMS) [12] and weakly supervised semantic segmentation [13]. The proposed approach of the researchers are shown in the Table 1.

The concept of Bidirectional Recurrent Neural Networks was later introduced in the field of Natural Language Processing and gained popularity. We have implemented the model by combining the ideas of ConvLSTM and Bidirectional RNNs to have better accuracy while predicting violence activity.

Model Architecture

To classify violent or non-violent actions, our model must be able to predict sequences in consecutive frames, that is a pattern in the movement of the subjects or a degree of their motion, etc. This is not possible by considering only the spatial features (features belongs to a particular frame) of the frames. The temporal or time-related features must also be considered while detecting sequences in the frames. The temporal features may be processed in the direction of upcoming frames or reverse order. Our model processes the temporal features in both the direction in addition to the spatial features, which helps the model to become more accurate at the same time consumes less computational time. The lightweight models are always preferred in surveillance due to its low-cost structure. The model consists of three sub-parts.

CNN

The Convolutional Neural Network (CNN) used in this paper, comprises of an input convolutional layer followed by three layers of convolution and max pooling. The kernel size for each convolutional layer is 3×3. 64 kernels are used in each convolutional layer. The output from each convolutional layer after passing through “relu” activation

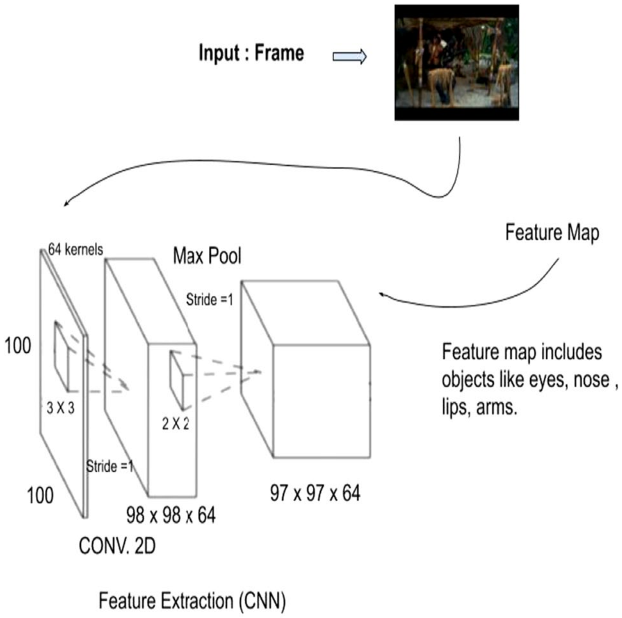


Fig. 1 General CNN structure

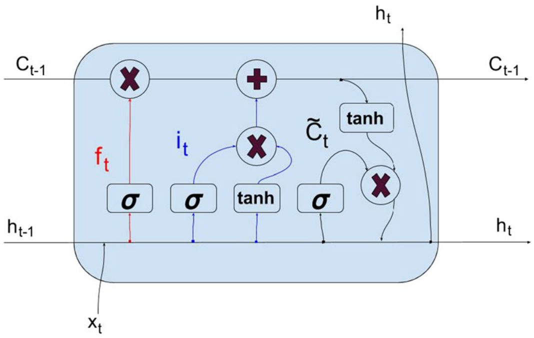


Fig. 2 The Basic LSTM cell

function is max pooled to extract the features. The filter size used in each max pooling is 2 × 2. Finally, the features are flattened and sent to the next model. Tensorflow[1] and Keras[2] API have been used to deploy the convolutional neural networks. The basic CNN functionality is shown in Fig. 1.

The Bidirectional LSTM Cells

The basic LSTM cell is shown in Fig. 2. The Long Short Term Memory cells are generally used to reconsider a part of previously trained features. LSTM mimics the activity of the human brain to remember the previously trained event. The first layer in an LSTM cell is known as the forgetting gate layer denoted by f_t. It is passed through a sigmoid function to get an output of either 0 or 1. The value 0 indicates a forget state and 1 denotes a remember state. The equation of the forget gate layer is given as,

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

The next layer is called the input gate layer i_t , in this layer, the remember state data are retrained with the new features.

$$i_t = \sigma(W_i.[h_{t-1}, x_t], b_i), \tag{2}$$

The output from the forget gate layer is multiplied to the cell state vector (c_t) of the previous LSTM cell (c_{t-1}). The result is added to the output from the input gate layer, multiplied to the hidden state vector of the last state upon passing through a “tanh” function, to form a cell state vector for the next LSTM cell. This vector upon passing through a “tan h” function, is multiplied to the hidden state vector of the previous state(h_{t-1}) upon passing through a “sigmoid” function to form a hidden state vector for the next LSTM cell (h_t).

Hence, in the final layer C_t^r, a part of the features from the previous state and the newly constrained features of the current cell are added up and passed to the next state.

$$C_t^r = tanh(W_c.[h_{t-1}, x_t], b_c), \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * C_t^r, \tag{4}$$

$$O_t = \sigma((W_o.[h_{t-1}, x_t], b_o)), \tag{5}$$

$$h_t = O_t * \tanh(C_t), \tag{6}$$

where, x_t is an input vector to the LSTM unit and b_f, b_i and b_o are the weight vectors for the forget gate layer, input gate layer and the output gate layer, respectively. In the LSTM, the features are remembered and passed from state 1 to state 2 to state n. The LSTM can also work in reverse direction as well, the features will be remembered and passed from state n to state 2 to state 1. By combining both these mechanisms, we build a bidirectional LSTM layer as shown in Fig. 3. The bidirectional LSTM cells are more accurate in storing data. For violence detection, a bidirectional LSTM will compare the sequence of frames once in the forward direction and once in the reverse direction, this mechanism adds on various cell states and training features which add robustness to our model.

The Dense Layers

The dense layers are omnipresent when it comes to Deep Learning. Here, the fully connected dense layers help to

[1] https://www.tensorflow.org/.

[2] https://keras.io/.

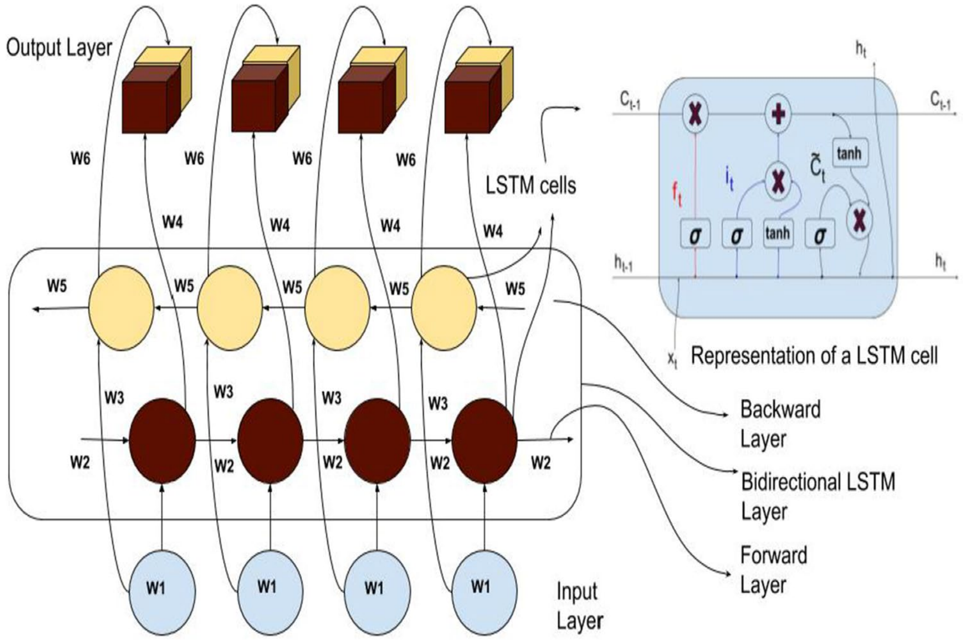


Fig. 3 The bidirectional LSTM cells

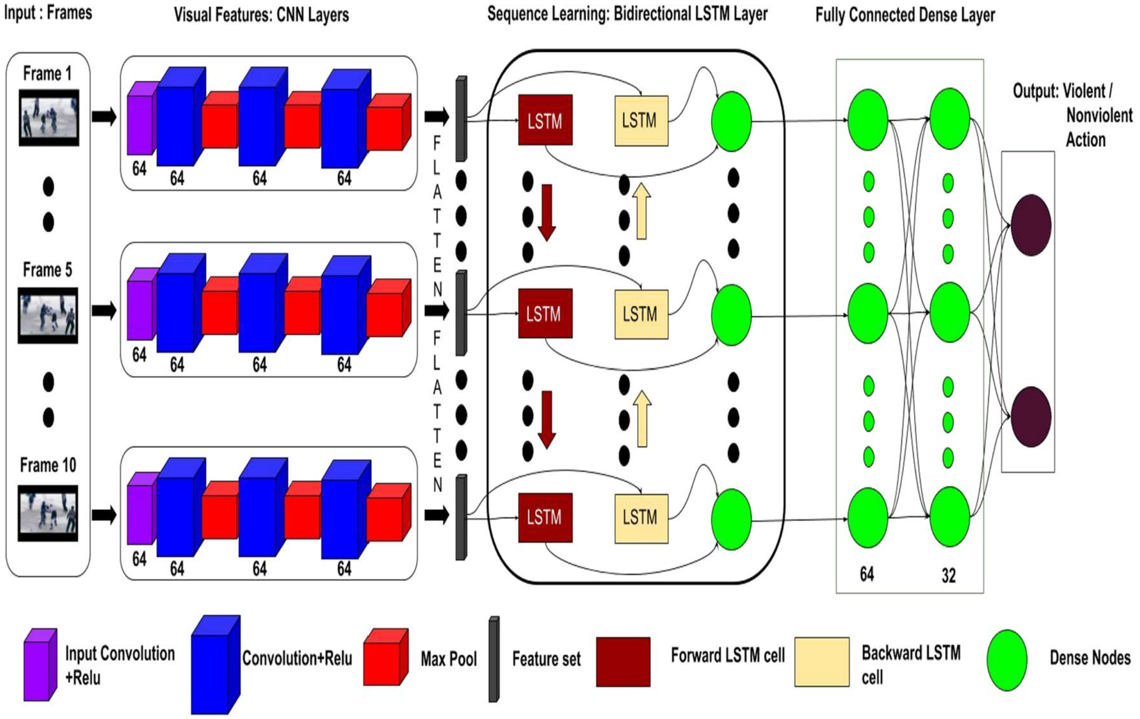
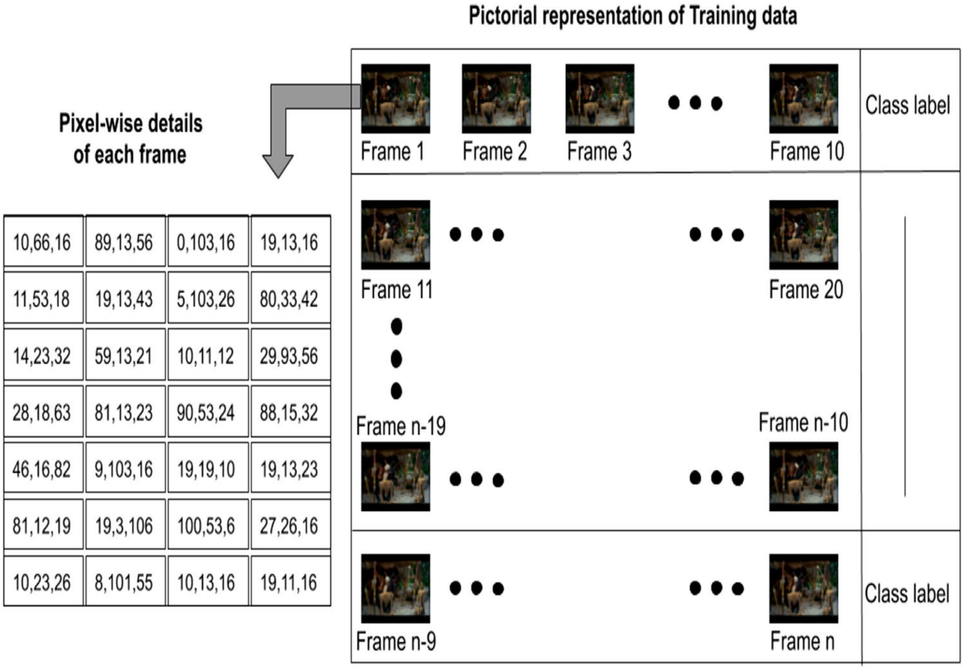


Fig. 4 Model architecture

add on random weights W_i to random features X_i, and test which set of features give the best accuracy over a certain number of epochs by passing through an activation function ν. In Fig. 4, the entire architecture of our proposed model has been shown.



10,66,16	89,13,56	0,103,16	19,13,16
11,53,18	19,13,43	5,103,26	80,33,42
14,23,32	59,13,21	10,11,12	29,93,56
28,18,63	81,13,23	90,53,24	88,15,32
46,16,82	9,103,16	19,19,10	19,13,23
81,12,19	19,3,106	100,53,6	27,26,16
10,23,26	8,101,55	10,13,16	19,11,16

Fig. 5 Visualization of the training data

Experimental Setup

Dataset

The effectiveness of the CNN Bidirectional LSTM model (CNN-BiLSTM) architecture has been validated by running on the standard datasets for violent and non-violent action detection, namely the *Hockey Fights* dataset [16], the *Movies* dataset [16] and the *Violent Flows* [24] dataset.

Hockey Fights Dataset

The Hockey Fights dataset contains clips from ice-hockey matches. The dataset has 500 violent clips and 500 non-violent clips of average duration of 1 s. The clips had a similar background and subjects.

Movies Dataset

The Movies dataset contains clips from different movies for action sequence whereas the non-fight sequences consist of clips from action recognition datasets. The dataset has 100 violent clips and 100 non-violent clips of average duration of 1 s. Unlike the Hockey Fights dataset, the clips of movies have different backgrounds and subjects.

Violent Flows Dataset

The Violent Flows data deal with crowd violence. The dataset consists of videos of human actions from the real world, CCTV footage of crowd violence, YouTube videos, properly maintaining the standard bench mark protocols. The dataset consists of 246 videos, with properly biased samples. The duration of the videos range from 1.04 to 6.52 s, with an average video length of 3.60 s.

Data Preprocessing

Frames have been extracted from the videos. The extracted frames are reshaped to 100 × 100 pixels (denoted as $x \times y$). The training data are a Numpy[3] array, with each of its row representing a sequence or pattern in video. A sequence might include a degree of movement and actions, whether a movement of the arm is a punch or a handshake, etc. The minimum number of frames required to extract a sequence is 2. However, we have used 10 consecutive frames (denoted as n) to extract the temporal features (that is, time-related features). The total number of samples (denoted by N) is the number of such sequences present in the dataset ((total number of frames)/(number of frames to be considered in a sequence)). For a simple implementation, numpy allows an arbitrary value of -1 to be used. Hence, a structure containing a sequence of 10 consecutive frames with their respective class labels is prepared. The shape of the training data is (-1, N, x, y, c)[4]. Here, c represents the number of channels in each frame. The pictorial representation of the training data is shown in Fig. 5.

[3] https://www.numpy.org/.

[4] N=total number of frames/n, n=10, x=100, y=100, c=3.

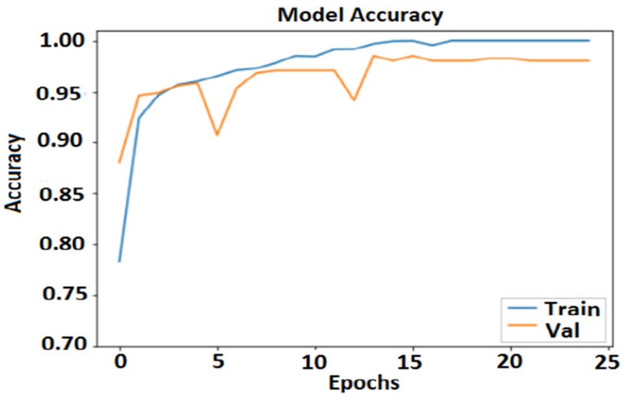


Fig. 6 Accuracy obtained form the Hockey fights dataset using hold out technique

System Configuration

The paper is implemented using Python 3.6 and Tensorflow (GPU) 1.14 on an Intel(R) Core(TM) $i7-9750H$ CPU (9^{th} Gen.) 2.60 GHz, 16 GB RAM and 6GB NVIDIA GeForce RTX 2060 with 64 bits Windows 10 Home operating system.

Training Methodology

A group of 10 consecutive frames, 100 × 100 dimensions, was passed to the model with a shape as shown in Fig 5 to extract the spatial and temporal features. Stochastic gradient descent has been used as an optimizer with a learning rate of 0.01 and decay=$1e^{-6}$. The loss function used in this paper is "sparse categorical crossentropy". In this multi-class classification problem, we have used "0 or 1" as class labels, instead of one-hot encoding, in a batch size of 5 samples at an instant. The datasets are divided into a 9 : 1 ratio, for training and testing purposes. The entire model has been build and trained from scratch for 25 epochs only to maintain its lightweight computation cost.

Results and Analysis

The following subsections give a detailed analysis of the model and the obtained results.

Accuracy Evaluation

The size of the datasets are relatively small, but the bidirectional LSTM model (CNN-BiLSTM model) addresses this issue quite well. The dataset has been divided into 10 parts. The 9 parts are used to train our model, whereas a single part has been used to achieve the validation accuracy (test accuracy). The in-sample accuracy (training sample accuracy) and out-sample accuracy (validation accuracy) have been averaged per epoch. A mean has been calculated for the epochs corresponding to the maximum accuracy region is taken with ±2 nearest epochs.

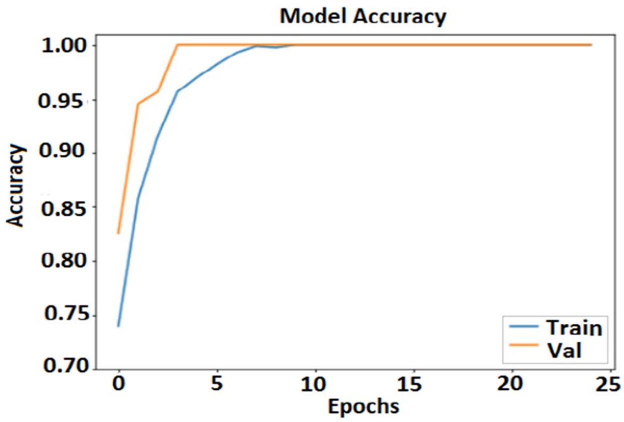


Fig. 7 Accuracy obtained form the Movies dataset using hold out technique

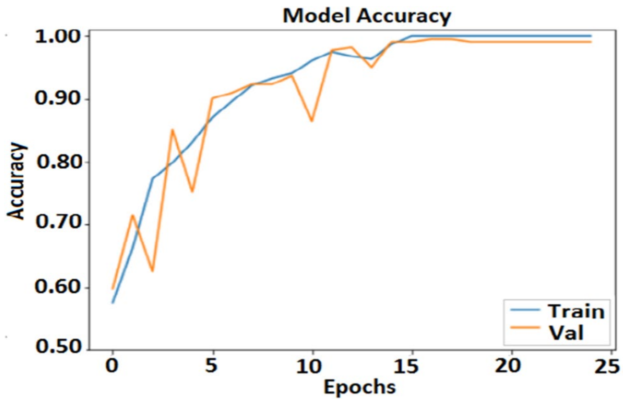


Fig. 8 Accuracy obtained form the Violent-Flows dataset using hold out technique

The Hockey Fights Dataset

In Fig. 6, we can see that the maximum validation accuracy is obtained for the 20th epoch with an accuracy of 99.27% and remained converged for the rest of the epochs in both the graphs. Hence, the overall accuracy is considered to be 99.27±0%[5].

The Movies Dataset

In the Fig. 7, we can see that the maximum validation accuracy is obtained for the 3rd epoch with an accuracy of 100%

[5] https://github.com/tintybot/CNN-BiLSTM-Model.

Table 2 Results obtained from different violent/non-violent classification models

Method	Hockey Fights	Movies	Violent-Flows
MoSIFT+HIK [16]	90.9%	89.5%	–
ViF [17]	82.9 ± 0.14%	–	81.3 ± 0.21%
MoSIFT+KDE+Sparse Coding [9]	94.3 ± 1.68%	–	89.05 ± 3.26
Deniz et al. [25]	90.1 ± 0%	98.0 ± 0.22%	–
Gracia et al. [26]	82.4 ± 0.4%	97.8 ± 0.4%	–
Substantial derivative [18]	–	96.89 ± 0.21%	85.43 ± 0.21%
Bilinski et al. [27]	93.4%	99%	96.4%
Three streams + LSTM [15]	93.9%	–	–
SELayer-3D CNN (C3D) [28]	99.0%	–	98.08%
3D CNN [22]	98.3%	100%	97.0%
[29]	96.33%	100%	95.71%
[30]	89.0%	-	92.0%
CNN-BiLSTM (our model)[a]	**99.27 ± 0%**	**100 ± 0%**	**98.64 ± 0%**

[a] https://github.com/tintybot/CNN-BiLSTM-Model

and remained converged for the rest of the epochs in both the graphs. Hence, the overall accuracy is considered to be 100%[5].

The Violent-Flows Dataset

In Fig. 8, we can see that the maximum validation accuracy is obtained for the 19th epoch with an accuracy of 98.64% and remained converged for the rest of the epochs in both the graphs. Hence, the overall accuracy is considered to be 98.64±0[5].

Accuracy Comparison

Our proposed model has obtained the best-reported results with an accuracy of 99.27% and 100% and 98.64% for the Hockey Fights dataset, the Movies dataset and the Violent-Flows dataset, respectively. It outperforms most of the best existing methods applied to the same datasets. Our model is lightweight and takes less training time with a less number of epochs than most of the previously used models. A comparative performance analysis for Hockey Fights, Movies and Violent-Flows datasets using hold out technique has been given in Table 2.

To examine the effectiveness of our model, we have implemented a fivefold cross-validation technique on the said datasets. In the method, after well shuffling the dataset, we have divided the entire dataset, comprising of sample sequences, into five equal folds. Onefold has been kept for validation while the model is trained on the rest of the fourfolds. Our model is prepared from scratch up to 25 epochs, keeping each fold once for validation (that is, a total of 5 times). For each trial, the validation accuracy has been calculated by taking a mean of maximum consistent accuracy for the consecutive 3 best performing epochs. We have calculated the mean and standard deviation for all the 5 trials. An accuracy-based comparison of our proposed models with other best practices obtained from the movies, hockey fights and violent-flows datasets has been shown in Table 3.

The detailed comparison of our proposed model with the few best performing previous models are shown in the Fig. 9 for 5-fold cross-validation. The proposed model performs slightly low (0.34%) in the Movies dataset than the other existing methods. However, our proposed approach achieves 98.87% mean-accuracy from all the three used datasets. This performance is also reportedly the combined best for these three violent activity detection standard datasets till date.

Furthermore, our proposed classification model has also been validated with a test video (of length 1 s) from the Movies dataset. The test is done with 10 frames per sequence and 4 sequences per second (non-overlapping), that is, 40 FPS. The time is taken to classify a given video input of length 1 s is 0.923 s.

Table 3 Accuracy obtained after fivefold cross-validation from our proposed classification model

Method	Hockey	Movies	Violent-Flows
MoIWLD [19]	96.8±1.04%	–	93.19±0.12%
ViF+OViF [20]	87.5±1.7%	–	88±2.45%
Spatiotemporal Encoder [21]	98.1±0.58%	**100±0%**	93.87±2.58%
Conv 3D [22]	98.3±0.81%	**100±0%**	97.17±0.95%
CNN-LSTM [14]	97.1±0.55%	**100±0%**	94.57±2.34%
CNN-BiLSTM (our model)	**98.78±0.38%**	99.66±0.31%	**98.18±1.12%**

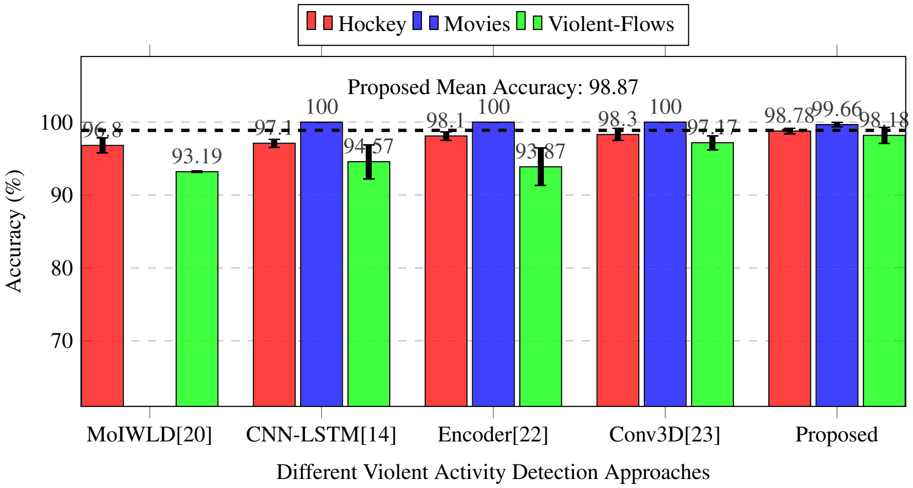


Fig. 9 Graphical presentation of the results obtained from different approaches (Table 3)

Conclusion

Our proposed CNN-BiLSTM variant provides the reportedly best results for the used datasets. Information about both the past trajectory and future trajectory of a video clip helps in better prediction and localization of occurrence of a violent event in a frame. Despite the satisfactory performance of our proposed model, it needs to be further validated with more standard datasets where identification of one to many or many to many violent activities including weapons are tough to detect.

Again, the detection model can be extended to a prevention model by analyzing the past sequence of events for an individual or a group of people. In the future, we will extend this work to address the said challenges in detecting violent and non-violent activities.

Compliance with Ethical Standards

Funding Nil.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Yuan J, Liu Z, Wu Y. Discriminative subvolume search for efficient action detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009; pp. 2442–2449. IEEE.
2. Chatterjee R, Bandyopadhyay T. Eeg based motor imagery classification using svm and mlp. In: 2016 2nd International Conference on Computational Intelligence and Networks (CINE), 2016 pp. 84–89. IEEE.
3. Neshige R, Kuroda Y, Kakigi R, Fujiyama F, Matoba R, Yarita M, Lüders H, Shibasaki H. Event-related brain potentials as indicators of visual recognition and detection of criminals by their use. Forensic Sci Int. 1991;51(1):95–103.
4. Datta A, Chatterjee R. Comparative study of different ensemble compositions in eeg signal classification problem. In: Emerging Technologies in Data Mining and Information Security. Berlin: Springer; 2019. p. 145–54.
5. Abootalebi V, Moradi MH, Khalilzadeh MA. A new approach for eeg feature extraction in p300-based lie detection. Comput Methods Programs Biomed. 2009;94(1):48–57.
6. Chatterjee R, Maitra T, Islam SKH, Hassan MM, Alamri A, Fortino G. A novel machine learning based feature selection for motor imagery eeg signal classification in internet of medical things environment. Future Gener Comput Syst. 2019;98:419–34.
7. Nam J, Alghoniemy M, Tewfik AH. Audio-visual content-based violent scene characterization. In: Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269), volume 1, 1998; pp. 353–357. IEEE,
8. Mousavi H, Mohammadi S, Perina A, Chellali R, Murino V. Analyzing tracklets for the detection of abnormal crowd behavior. In: 2015 IEEE Winter Conference on Applications of Computer Vision, 2015; pp. 148–155. IEEE
9. Xu L, Gong C, Yang J, Wu Q, Yao L. Violent video detection based on mosift feature and sparse coding. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014; pp. 3538–3542. IEEE.
10. Chen D, Wactlar H, Chen M, Gao C, Bharucha A, Hauptmann A. Recognition of aggressive human behavior using binary local motion descriptors. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008; pp. 5238–5241. IEEE.
11. Xingjian SHI, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W. Convolutional lstm network: a machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, 2015. pp. 802–810.
12. Medel JR, Savakis A. Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv preprint arXiv:1612.00390, 2016.
13. Patraucean V, Handa A, Cipolla R. Spatio-temporal video autoencoder with differentiable memory. arXiv preprint arXiv:1511.06309, 2015.

14. Sudhakaran S, Lanz O. Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017; pp. 1–6. IEEE.
15. Dong Z, Qin J, Wang Y. Multi-stream deep networks for person to person violence detection in videos. In: Chinese Conference on Pattern Recognition. Berlin: Springer; 2016. p. 517–31.
16. Nievas EB, Suarez OD, García GB, Sukthankar R. Violence detection in video using computer vision techniques. In: International conference on Computer analysis of images and patterns. Berlin: Springer; 2011. p. 332–9.
17. Hassner T, Itcher Y, Kliper-Gross O. Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012; pp. 1–6. IEEE
18. Mohammadi S, Kiani H, Perina A, Murino V. Violence detection in crowded scenes using substantial derivative. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2015; pp. 1–6. IEEE
19. Zhang T, Jia W, He X, Yang J. Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans Circuits Syst Video Technol. 2016;27(3):696–709.
20. Gao Y, Liu H, Sun X, Wang C, Liu Y. Violence detection using oriented violent flows. Image Vis Comput. 2016;48:37–41.
21. Hanson A, Pnvr K, Krishnagopal S, Davis L. Bidirectional convolutional lstm for the detection of violence in videos. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018.
22. Li J, Jiang X, Sun T, Xu K. Efficient violence detection using 3d convolutional neural networks. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019; pp. 1–8. IEEE,
23. Laptev I. On space-time interest points. Int J Comput Vision. 2005;64(2–3):107–23.
24. Itcher Y, Hassner T, Kliper-Gross O. Violent flows: Real-time detection of violent crowd behavior. In: 3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012.
25. Deniz O, Serrano I, Bueno G, Kim T-K. Fast violence detection in video. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), 2014. volume 2, pp. 478–485. IEEE.
26. Gracia IS, Deniz Suarez O, Garcia GB, Kim T-K. Fast fight detection. PLoS One. 2015;10(4):e0120448.
27. Bilinski P, Bremond F. Human violence recognition and detection in surveillance videos. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 30–36. IEEE, 2016.
28. Jiang B, Xu F, Tu W, Yang C. Channel-wise attention in 3d convolutional networks for violence detection. In: 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp. 59–64. IEEE.
29. Abdali A-MR, Al-Tuma RF. Robust real-time violence detection in video using cnn and lstm. In: 2019 2nd Scientific Conference of Computer Sciences (SCCS), 2019; pp. 104–108. IEEE,
30. Sharma M, Baghel R. Video surveillance for violence detection using deep learning. In: Advances in data science and management. Berlin: Springer; 2020. p. 411–20.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.