Int. J. Nonlinear Anal. Appl. 13 (2022) 2, 1117–1130
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6369

A review on video violence detection approaches

Mohamed Safaa Mohamed Shubber*, Ziyad Tariq Mustafa Al-Ta'i

Department of Computer Science, College of Science, University of Diyala, Baqubah, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

A violent behaviour detection system (VBDS) is an important application of intelligent video surveillance that performs a critical role in the field of public security and safety VBDS is a sort of behaviour recognition that seeks to determine whether the behaviours observed in the situation are violent, such as fighting or assault. This paper presents a survey of the existing approaches to VBDS. In this paper, the existing VBDS techniques are classified based on their framework, which includes the old-fashion framework and the end-to-end state-of-the-art deep learning framework. Finally, the VBDS methods' performance is assessed and compared.

Keywords: Artificial intelligence, computer vision, Deep learning, Violent behavior detection system (VBDS)
2020 MSC: 68T05

1 Introduction

It's possible that violence will always be an aspect of the human experience. Its influence can be observed in a variety of ways all throughout the world. Self-inflicted, individual, or social violence claims the lives of over a million people each year, with many more suffering unlethal injuries. Therefore, violence is one of the major causes of death among people aged 15 to 44 years old around the world [47].

Violence might strike at any time in any place. For example, in universities, the phenomenon of student involved in a violent behavior has spread rapidly, whether among teaching staff, teachers with students, students with students, and other partners such as members of the family, administration staff, and others. University is among the establishments where social interactions are observed [26]. Also, Verbal and physical violence against healthcare workers (HCWs) has reached alarming levels around the world and according to the World Medical Association, it defined violence against health employees as "an international emergency that undermines the very foundations of health systems and impacts critically on patient's health" [75]. As a result of human interactions violence may take a place in any crowded area like prisons, sport stadiums, malls, etc.

Considering violence can strike at any moment, depending on a human to perform the process of monitoring and detecting violent situations solely is ineffective [63]. Usually, this duty has required security staff to keep an eye on many monitors continuously. Therefore, undesirable events such as fighting and violent behavior may indeed be missed due to human exhaustion, poor attention and inexperience. As a result, automated video surveillance systems that detect anomalies in an automatic ways are critical for ensuring safety and assisting security guards [9].

*Corresponding author
Email addresses: msm.shubber@gmail.com (Mohamed Safaa Mohamed Shubber), ziyad1964tariq@uodiyala.edu.iq (Ziyad Tariq Mustafa Al-Ta'i)

Received: January 2022 *Accepted:* April 2022

Human behaivior detection and recognition is a hot topic in the research area of artificial intelligence via computer vision especially monitoring suspicious activities [64]. Lately, Various deep learning methods have been used to develop machine learning studies with impressive results. Convolutional neural networks also known as (CNNs), for example, are used in computer vision projects [40], also recurrent neural networks (RNN) are used [60].

This overview paper sheds the light on violence detection frameworks. In section two a general overview of the approaches which are used in this area. In section three CNN and RNN architectures will be reviewed, respectively. Section four shows the types of datasets that are used in violence detection and the performance of previous results (ranged from 2017 to 2021) are compared. In section five, some conclusions about the reviewed works are summarized.

2 Violence Detection Approaches

Many techniques to recognizing and analyzing human behavior have been implemented in the research field, and they are mostly classed as machine learning or deep learning approaches. Machine learning (ML) is defined as the ability of a system to learn from a training data specified in a certain problem with the intention of automate the process of developing analytical models and performing related activities. Deep learning which is a type of machine learning that uses artificial neural networks for the purpose of learning. Deep learning models outperform superficial machine learning models and traditional data analysis methodologies in many situations. Traditional machine learning methods or approaches depends on handmade feature extraction derived from classifying techniques; however, deep learning approaches have recently proved to be more successful in terms of detection accuracy and performance [37, 62]. Generally, every action recognition framework comprises mainly of two parts which are feature extractor and a classifier. Feature extraction is crucial in the development of any video detection system [58].

2.1 Machine Learning Approach

Before introducing deep learning to this field, what is called a two-step machine learning approach was used to detect anomaly behavior like violence as shown Figure 1, firstly it learns (extracts) feature from training data that are previously labeled and then uses an anomaly measure to estimate the normal/anomaly results based on the extracted or learned features [54]. feature extraction was done using some traditional (hand-crafted) machine learning techniques. Table 1 lists some of these techniques with their main feature.

Table 1: Hand-crafted Feature Extractor Techniques

Feature Extractor Technique	Feature
violence flows (ViF) [27]	Based on statistics of how flow-vector magnitudes change over time, to detect violence in crowded scenarios
oriented violent flows (OViF) [22]	using both orientation and magnitude.
Histogram of oriented optical flow (HOF) [51]	offer a long-term temporal description of the motion trajectory and its surroundings
Histograms of Oriented Gradients (HOG) [69]	calculates the number of times a gradient orientation appears in a confined area of an image
Histogram-of-optical-flow-orientation(OHOF) and Gaussian mixed-model-optical-flow-domain(GMOF) [78]	locate the scene of a violent act Then, in each location, the histogram-of-optical-flow-orientation (OHOF) descriptor is used to calculate spatiotemporal features that distinguish between violent and nonviolent behaviors.
Space–time-interest-point(STIP) [51]	examining the variety in sequences' spatial and temporal region of interest (roi)
three-layered bag-of-visual-words (BoVW) [48]	a three-layered structure, The video description is on the bottom layer, coding and pooling is on the middle layer, while supervised learning approach is on the top layer that is used to complete the classification task (SVM)
Space–time interest points (STIP) and scale-invariant-feature-transform (MoSIFT) descriptors are used in a bag of words framework [34]	every sequence of the video is explained as a form of histogram which is made up from a Bag of words, resulting in a fixed-dimensional encoding that classifiers can process.
Motion binary pattern (MBP) [51]	as the pixel intensity changes over time, it represents motion in a certain pixel

As for classifying the obtained features, three machine learning models were used: RF [29], SVM [13], KNN [20] and adaBooster [7] classifiers.

2.2 Deep Learning Approach

Although the handcrafted approach performs well, extracting features in this way is still costly in applications of real-world since it is tailored to individual issues and datasets. Deep learning algorithms have recently received interest in the computer vision field, replacing handcrafted methods [34].

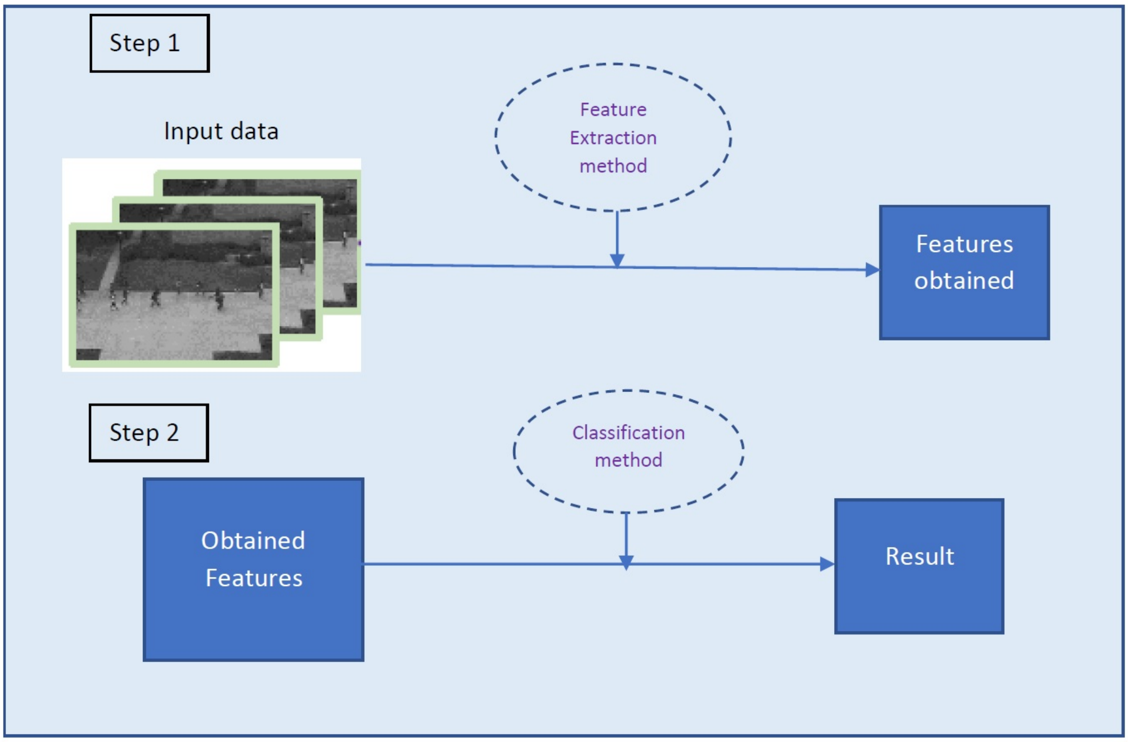

Figure 1: Two-step anomaly detection

3 Deep Learning

With the improvements in computing power and the availability of large-scale data, significant progress has been made in computer vision by using a deep-learning framework [42]. Deep learning techniques like Convolutional-Neural-Networks (CNN) have pushed the limits of what is possible by improving prediction accuracy with large amounts of data and huge processing power. Issues that were once thought to be unsolvable are now being resolved with remarkable precision [53]. The deep learning-based approaches can be classified as the end-to-end deep detection framework and the hybrid deep detection framework.

3.1 Convolutional Neural Network (CNN)

Convolutional-neural-networks (CNNs) are at the core of today's object detection algorithms. They are employed in features extraction. There are many CNNs architectures, such as LeNet, VGGNet and InceptionV3. These networks have been tested on a variety of commonly recognized benchmarks and datasets and are mostly used for object classification tasks. CNN contains three components which are convolution, polling and activation functions [8] as illustrated in **Error! Reference source not found.**. There are many types of convolution which are dilated, transposed module, tiled, Network in network and Inception module. The pooling part of the CNN has mixed, Lp pooling, stochastic, multiscale orderless spectral. The activation functions include rectified linear units (ReLU), parametric ReLU, randomized ReLU, Leakey ReLU, maxout, probout and exponential and linear unit (ELU) [35].

3.2 CNN Architecture

Several CNN models have been presented in the past decade. The architecture of a model is an important aspect in enhancing the performance of many systems. From 1989 to the present time, various adjustments to CNN architecture have been made. Structure transformation, regularization, optimizations in parameter, and other changes are examples of such modifications. On the other hand, it should be emphasized that the major improvement in CNN performance was due primarily to the restructuring of processing units and the development of new blocks. The most creative breakthroughs in CNN architectures have focused on the usage of network depth [5]. In Table 2 several common CNN architectures along with their accuracy results in ImageNet [19] which are also illustrated and in Figure 3.

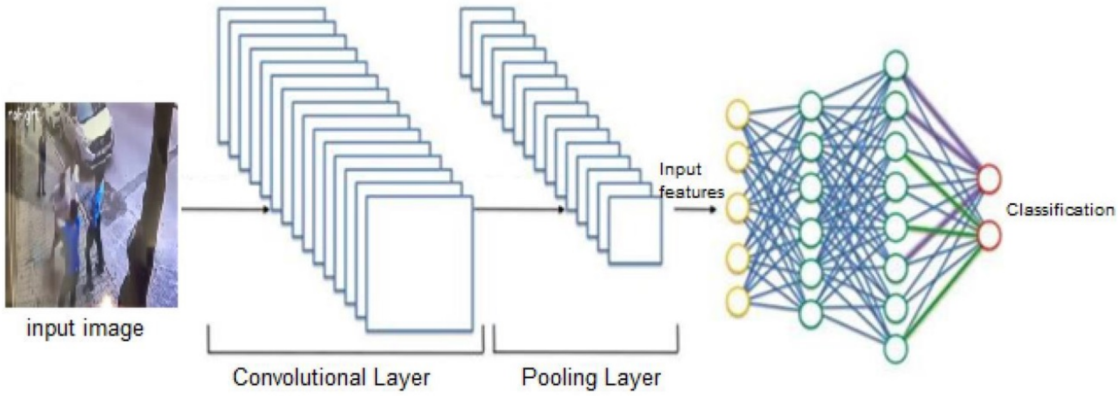

Figure 2: General CNN Architecture

Table 2: Common CNN Architectures

Model name	Parameters Number	ImageNet Accuracy	Year	Specification
Meta Pseudo Labels [57]	480 M	90.2%	2021	Design a feedback mechanism to correct the teacher's bias
Noisy Student EfficientNet-L2 [77]	480 M	88.4%	2020	It consists of two neural networks called teachers and students
BiT-L (ResNet) [41]	928 M	87.54%	2019	replaced the normalization of the batch (BN) with the normalization of the group (GN) and the weight standardization (WS).
Inception V3 [70]	27 M	78.8%	2015	Among the first to use batch normalization
ResNet-152 [16]	60 M	78.57%	2015	demands a large number of computations (approximately ten times that of AlexNet), implying greater training time and energy.
DenseNet-264 [32]	22 M	77.85%	2016	264 is number of layers with trainable weights
ResNet-50 [28]	26 M	77.15%	2015	Despite not the first to propose skip connections, this was the first one to promote them.
DenseNet-121 [32]	8 M	74.98%	2016	121 is number of layers with trainable weights
Inception V2 [70]	11.2 M	74.8%	2015	In the architecture of Inception V2 the 5 by 5 convolution is substituted by the two convolutions of 3 by 3. Because of this, computational time is reduced, and thus computational speed is increased. a 5 by 5 convolution is 2.78 more expensive than a 3 by s3 convolution.
VGG 19 [65]	144 M	74.5%	2014	VGG19 is slightly better than VGG16 but requests more memory
VGG 16 [65]	138 M	74.4%	2014	Has a contribution in the design of deeper networks
InceptionV1 [70]	5 M	69.8%	2014	Convolutional layers were stacked within modules/blocks.
AlexNet [65]	60 M	63.3%	2012	It was the first to use ReLU as an activation function.
LeNet-5 [43]	60,000	N/A	1998	Standard template also Stacking convolutional layer and pooling layers, and ending with one or many layers that are entirely connected

3.3 Recurrent Neural Network

Detection a human behavior makes an extension over a period of time (aka frames in videos), therefore a convolutional neural network will not be sufficient for such purpose because it generates a prediction on each frame individually without taking the previous frames into account since it lacks the memory function. Recurrent neural networks (RNN) [18] which are the state-of-the-art algorithm specialized in sequential data. Due to its internal memory, this is the first algorithm that remembers its input, making it ideal for machine learning issues with sequential data. It's one of the algorithms that's been at the heart of deep learning's incredible progress over the last several years.

Yet, RNNs have issues with exploding and also vanishing gradients, which can be explained as the gradient norm increases (or decreased) significantly during training. The growth of long-term components, which can increase exponentially faster than short-term components, is causing these phenomena. As for the vanishing problem that hinders the learning of long data sequences. The gradients contain the information that is employed in the RNN (Recurrent

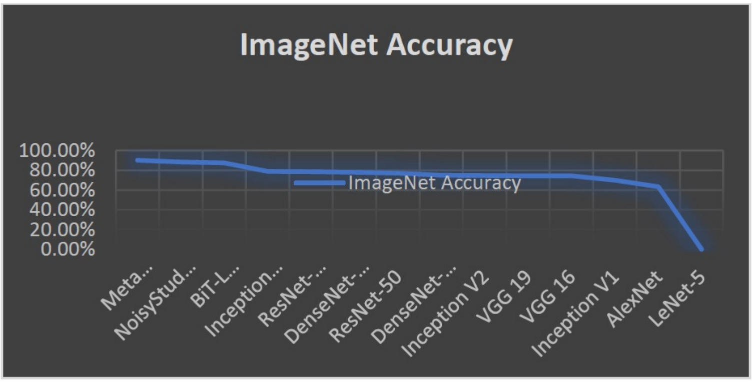

Figure 3: ImageNet accuracy results

Neural Network) parameter update, and as the gradient shrinks, the parameter changes become unimportant, implying that no significant learning is taking place [55].

3.3.1 LSTM

There have been several variants on the recurrent neural network architecture to solve the long-range dependency problem. One such architecture is the Long-Short Term Memory (LSTM [30]). The LSTM is a type of a recurrent neural network which has a sequence of inputs, an optional sequence of outputs, a hidden state, and a cell state.

While a vanilla RNN cell consists of a fully connected network with recurrent connections, an LSTM cell abstracts its mathematical operations into 'gates' which are a combination of weight multiplication and non-linearity. Information in the cell state, which is the main information highway of an LSTM, is modified according to the gates. LSTM can have different models such as:

A- Classic (or vanilla) LSTM [30]

The cell state is controlled by four gating levels in this architecture: two input gates, a forget gate, and output gates. The input gates collaborate to figure out which inputs should be added to the cell state. Based on the current cell state, the forget gate determines which former cell state to forget. The output gates decide what output to be sent through them.

B- Stacked LSTM [15]

An LSTM Model with many LSTM layers is known as a Stacked LSTM. The LSTM layer provides sequential output to the following LSTM layer.

C- Bi-directional LSTM [15]

Instead of training one input sequence, the Bi-directional LSTM trains two, with the first being the original and the second being its reversed replica. Which improves the learning rate of the model.

D- GRU (Gated Recurrent Unit) [21]

The Gated Recurrent Unit (GRU) is a type of recurrent unit in their most basic form, neural networks are made up of two gates (Reset Gate and Update Gate). Short-term dependencies in sequences are captured using Reset Gates, whereas long-term dependencies are captured using Update Gates. Both gates control how much information each concealed unit must remember or forget during the sequence's processing.

E- BGRU (Bidirectional GRU) [21]

The Bi-directional GRU, like the Bi-directional LSTM, is also a Bi-directional RNN, which means the BGRU is nothing more than a bi-directional GRU.

Figure 4 represents a simple CNN-LSTM architecture.

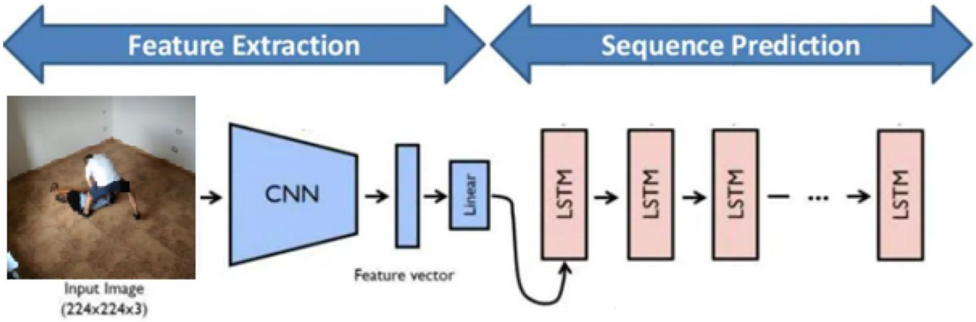

Figure 4: Simple CNN-LSTM architecture

Dataset and Performance used in Violence Detection Methods

3.4 Datasets

There are several datasets have been used in violence detection area:

A- RWF-2000 [**14**]: The most comprehensive dataset on violence detection, including over 2000 hours of real-time CCTV film. Each video is a 5-second clip with different resolutions and a 30-fps framerate. The videos contain a variety of settings and lighting.

B- Hockey dataset [**52**]: Comprises of 1000 clips gathered from various ice hockey footage. There are 50 frames in each video. The surroundings and violent activities in all the videos are the same.

C- Movies dataset [**52**]: Considerably a smaller dataset of 200 video clips in various resolutions were collected. The videos cover a broad range of topics. The 'violent' videos were compiled from a variety of movie clips.

D- Surveillance fight data set [**3**]: This data collection was gathered from YouTube and includes both violent and nonviolent incidents captured in the real world through regular and industrial surveillance. The data set contains a total of 300 videos, with 150 clips in each class, and resolution sizes ranging from 480×360 to 1280×720 pixels on average.

E- ViolentFlow data set [**27**]: This data set is containing a total of 246 videos with violent and nonviolent scenes. Each video clip has a resolution of 320×240 pixels, and the clip length varies between 50 and 150 frames.

F- Real-Life-Violence-Situations (RLVSs) [**66**]: The violence videos contain 1000 clips of real street combat situations in a variety of environments and conditions, and the nonviolence videos also contain 1000 nonviolent clips collected from YouTube. In addition, nonviolence clips are gathered from a variety of human activities such as sports, eating, walking, and so on.

G- Automatic Violence Detection in Videos Dataset [**10**]: This dataset contains 350 clips (MP4 video files with a resolution of 1920×1080 pixels and a frame rate of 30 frames per second). When portraying non-violent actions, 120 clips are labeled as non-violent, while when representing violent behaviors, 230 clips are labeled as violent. Due to fast movements and similarities with violent behaviors, the non-violent video contains behaviors like (hugs, claps, exulting, etc.) can result in false positives in the violence detection test.

3.5 Performance Comparisons

Table 3 represents some of the recent studies for violence detection which been using the datasets been mention above beside automatic violence detection in videos dataset [10] since no published paper stated the use of it yet.

Table 3: Recent violence detection studies according to the datasets

Year	Approach	Pre-processing method	Dataset	Results
2021	Dual Spatio-temporal Convolutional Network (DSTCN) [23]	Resize each frame to a fixed scale of 224 × 224, and sample 32 frames, generating input of shape 3 × 32 × 224 × 224 then various data augmentation approaches	Hockey dataset [52]	99%
			movies dataset [52]	100%
2021	Pre-trained CNN+NN [31]	resized frames to 224 × 224 × 3, which is the input size of used DNNs (deep)	Hockey dataset [52]	96%
			movies dataset [52]	100%
			violent flows dataset [27]	96%
			Real life Videos dataset [66]	97%
2021	Separable Convolutional LSTM Modules [33]	provide the difference of consecutive frames as inputs, which pushes the model to encode temporal changes between adjacent frames, improving motion capture.	Hockey dataset [52]	99.50%
			movies dataset [52]	100%
			rwf-2000 dataset [14]	89.75%
2021	two-cascade Temporal Shift Modules [44]	convert the temporal information that is not apparent in one frame into spatial function data that may be extracted	Hockey dataset [52]	98.995%
			violent flows dataset [27]	97.959%
			rwf-2000 dataset [14]	89.277%
2021	MSM + EfficientNet-B0 with frame-grouping + TSE Block [38]	RGB difference and morphological dilation	Hockey dataset [52]	99.6%
			movies dataset [52]	100%
			violent flows dataset [27]	98%
			rwf-2000 dataset [14]	92%
			Real life Videos dataset [66]	97.8%
			Surveillance Camera Fight Dataset [3]	92%
2021	Automated mobile neural architecture search network and ConvLstm along with machine learning models for classification [34]	Simple Data Augmentation	Hockey dataset [52]	99%
			movies dataset [52]	100%
			violent flows dataset [27]	96%
2021	Lightweight CNN for processing video stream acquired through vision sensor, and residential optical flow CNN used for extracting temporal optical flow features [73]	image augmentation (IMGAUG) technique [71]	Hockey dataset [52]	98%
			violent flows dataset [27]	98.21%
			Surveillance Camera Fight Dataset [3]	74%

Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, and Figure 5 illustrate the accuracy result of different methods on different datasets

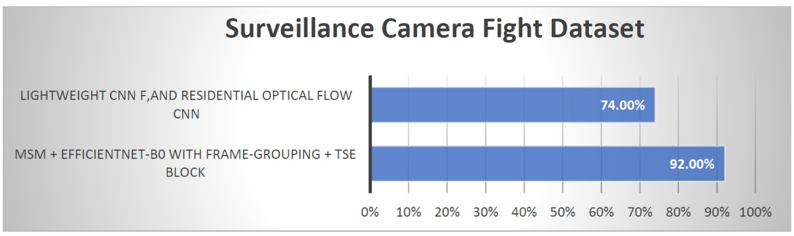

Figure 5: violent behavior detection accuracy results on Surveillance Camera Fight dataset

4 Conclusions

Deep learning-based detection of violent behavior does not necessitate the human construction of feature extraction algorithms that standard methods do. Alternatively, the video dataset can be used to train and learn to find the best

2021	VDstr: VIOLENCE DETECTION UNDER SPATIO-TEMPORAL REPRESENTATIONS [12]	linear interpolation on masked pixels to fill the missing values in the SITS [11]	Hockey dataset [52]	94.4%
			movies dataset [52]	99.5%
			violent flows dataset [27]	90.6%
			rwf-2000 dataset [14]	93.8%
2021	SAM-GhostNet-ConvLSTM [45]	SAM [76]	Hockey dataset [52]	97.5%
			rwf-2000 dataset [14]	87.50%
2021	Hybrid CNN + LSTM [56]	Video sampled to a frame-by-frame sequence then augmented	Hockey dataset [52]	86.70%
			movies dataset [52]	100%
			violent flows dataset [27]	91.40%
2021	Flow Gated RGB [17]	N/A	Hockey dataset [52]	93%
			movies dataset [52]	90%
			rwf-2000 dataset [14]	81%
			Real life Videos dataset [66]	87.25%
2021	CNN-LSTM that based on IOT node [4]	10 frames from each movie were grabbed at periodic intervals and then scaled to 112 × 112 pixels. Additionally changed the color scheme from BGR to RGB. The data was then augmented, and finally normalized by a factor of 1/255.	Mixed (rwf-2000 dataset [14]+ Real life Videos dataset [66])	73.35%
2021	3D DenseNet, multi-head self-attention mechanism, and BiConvLSTM [59]	N/A	Real life Videos dataset [66]	95.60%
2021	data-efficient video transformer (DeVTr) [1]	N/A	Real life Videos dataset [66]	96.25%
2021	CNN(VGG16) and ConvLSTM [49]	data augmentation techniques + resizing to 244 × 244 + normalization	Hockey dataset [52]	99.1%
			movies dataset [52]	100%
			violent flows dataset [27]	98.4%
			rwf-2000 dataset [14]	92.4%
2020	Extraction of motion features from RGB Dynamic Images [36]	N/A	Hockey dataset [52]	93.33%
			movies dataset [52]	100%
			Real life Videos dataset [66]	86.79%
2020	(RNNs) and (2D CNN) [72]	N/A	Real life Videos dataset [66]	96.74%
2020	multi-head Skeleton Points Interaction Learning (SPIL) [67]	N/A	rwf-2000 dataset [14]	89.3%
			Hockey dataset [52]	96.8%
			violent flows dataset [27]	94.5%
			movies dataset [52]	98.5%
2020	CNN BiLSTM [24]	Extracted frames are reshaped to 100 × 100 pixels	Hockey dataset [52]	99.27%
			movies dataset [52]	100%
			violent flows dataset [27]	98.64%
2019	Bidirectional Convolutional LSTM [25]	Resize and normalization and random cropping (RC) and random horizontal flipping (RHF)	Hockey dataset [52]	96.96%
			movies dataset [52]	100%
			violent flows dataset [27]	92.18%
2019	3D convolutional neural network [74]	Person shape detection and frame resizing	Hockey dataset [52]	96%
			movies dataset [52]	99.9%
			violent flows dataset [27]	98%
2019	CNN+LSTM [6]	Resize and normalization	Hockey dataset [52]	98%
			violent flows dataset [27]	92.19%

efficient video representation method. This method has a high degree of data adaptation and can produce improved detection results. The Convolutional Neural Network was successful, because the picture identification results are promising, the convolutional neural network-based technology has grabbed people's interest from the start. CNN has produced reliable results in video violence identification, but this method is supervised learning and requires a large number of training samples as well as expensive hardware. For a long time, the LSTM algorithm overcomes the phenomena of gradient extinction in video and can make greater use of the time dimension information. With ongoing study, it is expected that new video violence detection frameworks based on deep learning will emerge in the future. Although the methods of detecting violence in videos appeared some time ago, this research covers techniques in this field from 2017 until now. Most of these techniques are based on the concept of deep learning, because it (at least) does not need for using separate feature extraction algorithms that standard machine learning methods do.

2019	CNN (pre-trained vgg19) followed by LSTM [2]	N/A	Hockey dataset [52]	98%
2018	Deep CNN using transfer learning [50]	N/A	Hockey dataset [52]	99.28%
			movies dataset [52]	99.97%
2017	Convolutional Long Short-Term Memory [68]	During the training step, data augmentation techniques such as random cropping and horizontal flipping are used. A segment of the frame of size 224 × 224 is cropped during each training iteration.	Hockey dataset [52]	97.1%
			movies dataset [52]	100%
			violent flows dataset [27]	94.57%
2017	Temporal examination of texture measures based on the grey level co-occurrence matrix (GLCM) [46]	N/A	violent flows dataset [27]	86.03%
2017	Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation [61]	N/A	Hockey dataset [52]	94.42%
			movies dataset [52]	94.95%
			violent flows dataset [27]	93.12%
2017	Optical flow and pre-trained CNN [39]	N/A	Hockey dataset [52]	94.40%
			movies dataset [52]	96.5%
			violent flows dataset [27]	80.9%

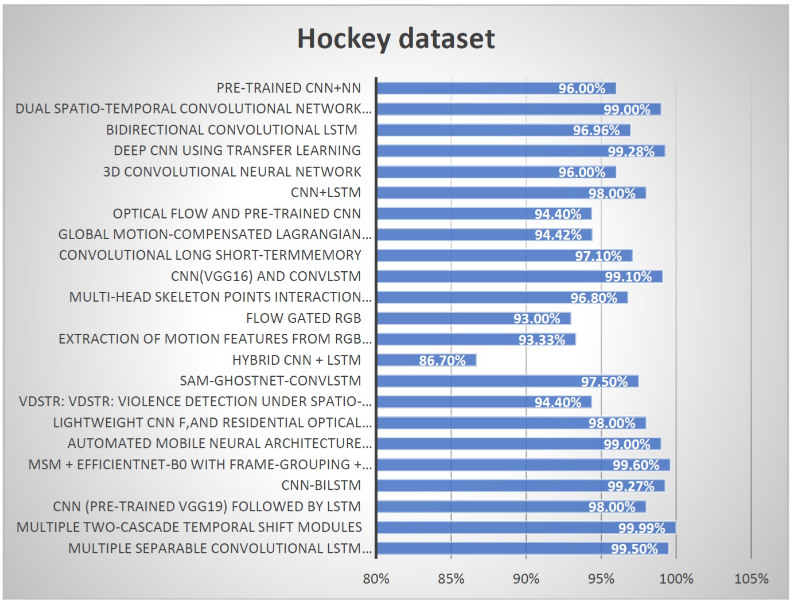

Figure 6: violent behavior detection accuracy results on Hockey dataset

Alternatively, deep learning-based detection of violent behavior uses the video dataset to train and learn in order to find the best efficient video representation method. The most reliable method in this review was Convolutional Neural Network (CNN), because it produced reliable results in video violence identification. This review also found that for long time videos the LSTM algorithm overcomes the phenomena of gradient extinction in video and can make greater use of the time dimension information. As a conclusion, this review expects that more new video violence detection frameworks based on deep learning will emerge in the future.

References

[1] A.R. Abdali, *Data efficient video transformer for violence detection*, IEEE Int. Conf. Commun. Networks Satell., 2021, p. 195–199.

[2] A.M.R. Abdali and R.F. Al-Tuma, *Robust real-time violence detection in video using CNN snd LSTM*, 2nd Sci. Conf. Comput. Sci. (SCCS), 2019, p. 104–108.

[3] S. Akti, G.A. Tataroglu and H.K. Ekenel, *Vision-based fight detection from surveillance cameras*, 9th Int. Conf. Image Process. Theory, Tools Appl., 2019, p. 1–6.

[4] N. Aldahoul, H.A. Karim, R. Datta, S. Gupta, K. Agrawal and A. Albunni, *Convolutional neural network-long*

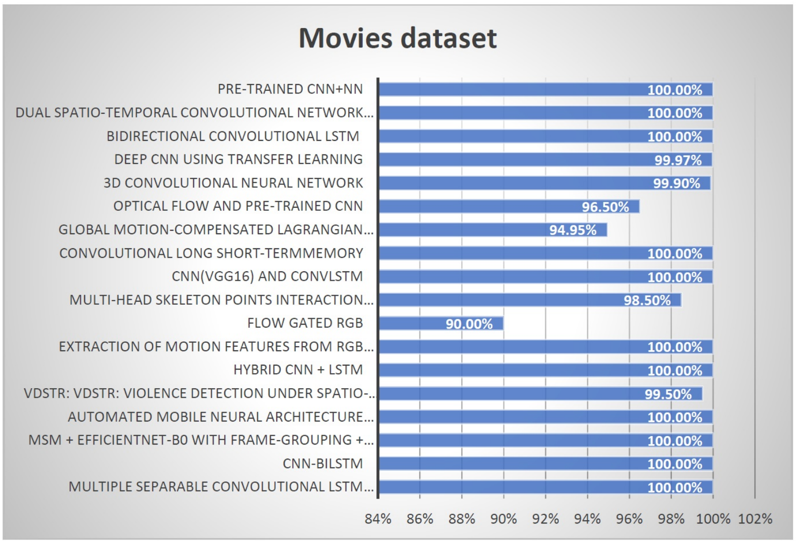

Figure 7: violent behavior detection accuracy results on Movies dataset

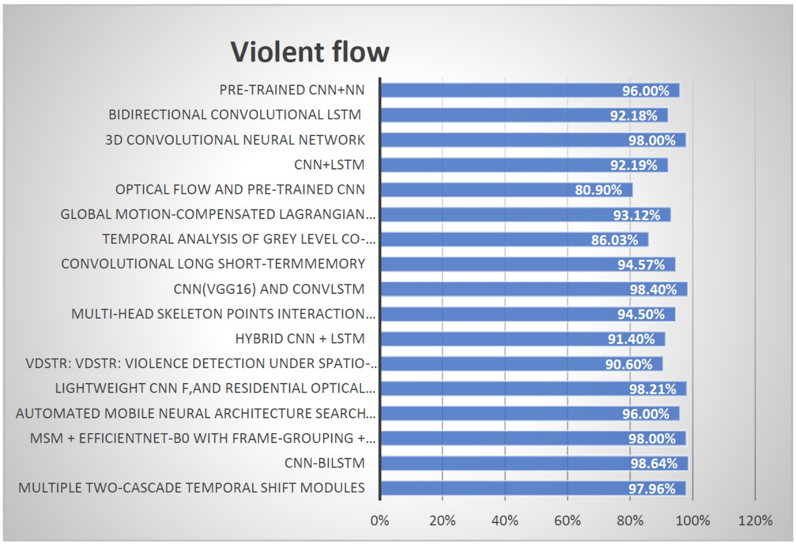

Figure 8: : violent behavior detection accuracy results on Violent flow dataset

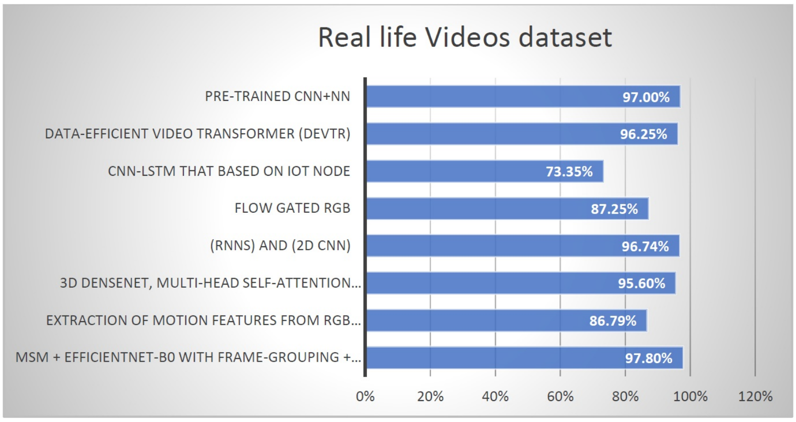

Figure 9: violent behavior detection accuracy results on Real life Videos dataset

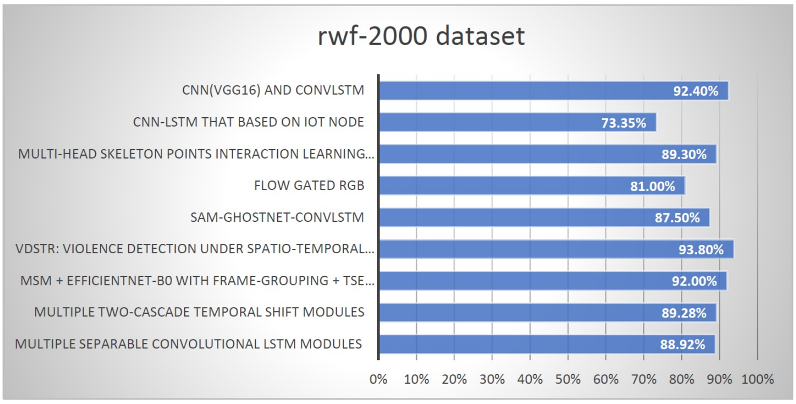

Figure 10: violent behavior detection accuracy results on rwf-2000 dataset

short term memory based IOT node for violence detection, IEEE Int. Conf. Artif. Intell. Eng. Tech. (IICAIET), 2021, p. 1–6.

[5] L. Alzubaidi, J. Zhang, A.J. Humaidi and A. Al-Dujaili, *Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions*, J. Big Data **8** (2021), no. 1.

[6] S.M.R. Ammar, M. Anjum, T. Rounak, M. Islam and T. Islam, *Using deep learning algorithms to detect violent activities*, Doctoral dissertation, BRAC University, 2019.

[7] R. Barmaki, *A decision-theoretic generalization of on-line learning and an application to boosting**, J. Comput. Syst. Sci. **55** (1996), no. 1, 119–139.

[8] A. Benali Amjoud and M. Amrouch, *Convolutional neural networks backbones for object detection*, Int. Conf. Image Signal Process., 2020, p. 282–289.

[9] A. Ben Mabrouk and E. Zagrouba, *Spatio-temporal feature using optical flow based distribution for violence detection*, Pattern Recog. Lett. **92** (2017), 62–67.

[10] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi and A.F. Dragoni, *A dataset for automatic violence detection in videos*, Data Br. **33** (2020), 106587.

[11] M. Chelali, C. Kurtz, A. Puissant and N. Vincent, *Classification of spatially enriched pixel time series with convolutional neural networks*, 25th Int. Conf. Pattern Recog. (ICPR), 2020, p. 5310–5317.

[12] M. Chelali, C. Kurtz and N. Vincent, *Violence detection from video under 2d spatio-temporal representations*, IEEE Int. Conf. Image Process. (ICIP), 2021, p. 2593–2597.

[13] H.F. Chen, *Support-vector networks CORINNA*, Chem. Biol. Drug Des. **74** (1995), no. 2, 142–147.

[14] M. Cheng, K. Cai and M. Li, *RWF-2000: An open large scale video database for violence detection*, 25th Int. Conf. Pattern Recog. (ICPR), 2020, p. 4183–4190.

[15] Z. Cui, R. Ke, Z. Pu and Y. Wang, *Stacked bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction*, Transp. Res. Part C Emerg. Technol. **118** (2020), p. 102674.

[16] B. Di Liu, J. Meng, W.Y. Xie, S. Shao, Y. Li and Y. Wang, *Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification*, Remote Sens. **11** (2019), no. 5, 1–18.

[17] D. Durães, F. Santos, F.S. Marcondes, S. Lange and J. Machado, *Comparison of transfer learning behaviour in violence detection with different public datasets*, EPIA Conf. Artif. Intell., 2021, p. 290–298.

[18] J.L. Elman, *Finding structure in time*, Cogn. Sci. A Multidiscip. **14** (1986), no. 2, 179–211.

[19] L. Fei-Fei, J. Deng and K. Li, *ImageNet: Constructing a large-scale image database*, IEEE Conf. Comput. Vis. pattern Recog., 2009, p. 248–255.

[20] E. Fix and J.L. Hodges, *Discriminatory analysis. Nonparametric discrimination: Consistency properties*, Consistency Prop. Int. Stat. Rev. Int. Stat. **57** (1989), no. 3, 238–247.

[21] Y. Gao and D. Glowacka, *Deep gate recurrent neural network*, Workshop Conf. Proc., 2016, p. 350–365.

[22] Y. Gao, H. Liu, X. Sun, C. Wang and Y. Liu, *Violence detection using oriented violent flows*, Image Vis. Comput. **48** (2016), 37–41.

[23] D.K. Ghosh, A. Chakrabarty, N. Mansoor, D.Y. Suh and J. Piran, *Learning-driven spatio-temporal feature extraction for violence detection in IoT environments*, Int. Conf. Inf. Commun. Technol. Converg., 2021, p. 1807–1812.

[24] R. Halder and R. Chatterjee, *CNN-BiLSTM model for violence detection in smart surveillance*, SN Comput. Sci. **1** (2020), no. 4, 1–9.

[25] A. Hanson, K. Pnvr, S. Krishnagopal and L. Davis, *Bidirectional convolutional LSTM for the detection of violence in videos*, in European Conference on Computer Vision (ECCV) Workshops, 2018, p. 280–295.

[26] A.E.H. Hassan and M.E.E. Ageed, *Student violence in universities (manifestation, causes, effects, and solution's) in Zalingei University-central Darfur State Sudan*, ARPN J Sci Technol. **5** (2015), no. 2, 80–86.

[27] T. Hassner, Y. Itcher and O. Kliper-Gross, *Violent flows: Real-time detection of violent crowd behavior*, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, 2012, p. 1–6.

[28] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, IEEE Conf. Comput. Vis. pattern Recog., 2016, p. 770–778.

[29] T.K. Ho, *Random decision forests*, 3rd Int. Conf. Doc. Anal. and Recog., **1** (1995), 278–282.

[30] S. Hochreiter and J. Schmidhuber, L*ong short-term memory*, Neural Comput. **9** (1997), no. 8, 1735–1780.

[31] N. Honarjoo, A. Abdari and A. Mansouri, *Violence detection using pre-trained models*, 5th Int. Conf. Pattern Recognit. Image Anal. (IPRIA), 2021, p. 1–4.

[32] G. Huang, Z. Liu, L. Van Der Maaten and K.Q. Weinberger, *Densely connected convolutional networks*, IEEE Conf. Comput. Vis. Pattern Recog., 2017, p. 4700–4708.

[33] Z. Islam, M. Rukonuzzaman, R. Ahmed, M.H. Kabir and M. Farazi, *Efficient two-stream network for violence detection using separable convolutional LSTM*, Int. Joint Conf. Neural Networks, 2021, p. 1–8.

[34] H.M.B. Jahlan and L.A. Elrefaei, *Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video*, Arab. J. Sci. Eng. **46** (2021), no. 9, 8549–8563.

[35] A. Jain and D.K. Vishwakarma, *State-of-the-arts violence detection using ConvNets*, IEEE Int. Con. Commun. Signal Process., 2020, p. 813–817.

[36] A. Jain and D.K. Vishwakarma, *Deep neuralNet for violence detection using motion features from dynamic images*, Third Int. Conf. Smart Syst. Invent. Technol., 2020, p. 826–831.

[37] C. Janiesch and K. Heinrich, *Machine learning and deep learning*, Electron. Mark. **31** (2021), 685–695.

[38] M.S. Kang, R.H. Park and H.M. Park, *Efficient spatio-temporal modeling methods for real-time violence recognition*, IEEE Access **9** (2021), 76270–76285.

[39] A.S. Keçeli and A. Kaya, *Violent activity detection with transfer learning method*, Electron. Lett. **53** (2017), no. 15, 1047–1048.

[40] K.E. Ko and K.B. Sim, *Deep convolutional framework for abnormal behavior detection in a smart surveillance system*, Eng. Appl. Artif. Intell. **67** (2018), 226–234.

[41] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly and N. Houlsby, *Big transfer (BiT): General visual representation learning*, Computer Vision–ECCV 2020: 16th European Conf. **16** (2020), 491–507.

[42] Y. Lecun, Y. Bengio and G. Hinton, *Deep learning*, Nature **521** (2015), no. 7553, 436–444.

[43] Y. Lecun, L. Bottou, Y. Bengio and P. Ha, *Gradient-based learning applied to document recognition*, Proc. IEEE, **86** (1998), no. 11, 2278–2324.

[44] Q. Liang, Y. Li, B. Chen and K. Yang, *Violence behavior recognition of two-cascade temporal shift module with attention mechanism*, J. Electron. Imag. **30** (2021), no. 04, 1–13.

[45] Q. Liang, Y. Li, K. Yang, X. Wang and Z. Li, *Long-term recurrent convolutional network violent behaviour*

recognition with attention mechanism, MATEC Web Conf. **336** (2021), p. 05013.

[46] K. Lloyd, P.L. Rosin, D. Marshall and S.C. Moore, *Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM) -based texture measures*, Mach. Vis. Appl. **25** (2017), no. 3–4, 361–371.

[47] C. Mencacci, *Violence: A global public health problem*, Quad. Ital. Psichiatr. **30** (2002), no. 1, 1–2.

[48] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein and A. Rocha, *Temporal robust features for violence detection*, IEEE Winter Conf. Appl. Comput. Vision (WACV), 2017, p. 391–399.

[49] 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*, 6th Int. Conf. Image, Vision and Comput. ICIVC, 2021, p. 55–60.

[50] A. Mumtaz, A.B. Sargano and Z. Habib, *Violence detection in surveillance videos with deep network using transfer learning*, 2nd Eur. Conf. Electr. Eng. Comput. Sci. (EECS), 2018, p. 558–563.

[51] A.J. Naik and M.T. Gopalakrishna, *Violence detection in surveillance video-A survey*, Int. J. Lat. Res. Engin. Technol. **2017** (2017), 11–17.

[52] E.B. Nievas, O.D.Suarez, G.B. García and R. Sukthankar, *Violence detection in video using computer vision techniques*, Int. Conf. Comput. Anal. Images and Patterns, 2011, p. 332–339.

[53] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G.V. Hernandez, L. Krpalkova, D. Riordan and J. Walsh, *Deep learning vs. traditional computer vision*, Adv. Intell. Syst. Comput. **943** (2020), 128–144.

[54] G. Pang, C. Yan, C. Shen, A. van den Hengel and X. Bai, *Self-trained deep ordinal regression for end-to-end video anomaly detection*, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., 2020, p. 12170–12179.

[55] R. Pascanu, T. Mikolov and Y. Bengio, *On the difficulty of training recurrent neural networks*, Int. Conf. Machine Learn., 2013, p. 1310–1318.

[56] M.B. Patel, *Real-time violence detection using CNN-LSTM*, arXiv Prepr. arXiv2107.07578, (2021), 1–6.

[57] H. Pham, Z. Dai, Q. Xie, M.-T. Luong and Q.V. Le, *Meta pseudo labels*, in IEEE/CVF Conf. Comput. Vis. Pattern Recog., 2021, p. 11557–11568.

[58] M. Ramzan, A. Abid, H.U. Khan, S.M. Awan, A. Ismail, M. Ahmed, M. Ilyas and A. Mahmood, *A review on state-of-the-art violence detection techniques*, IEEE Access **7** (2019), 107560–107575.

[59] F.J. Rendón-Segador, J.A. Álvarez-García, F. Enríquez and O. Deniz, *ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence*, Electron. **10** (2021), no. 13, 1601.

[60] D.E. Rumelhart, G.E. Hinton and R.J. Williams, *Learning internal representations by error propagation*, Calif. Univ San Diego La Jolla Inst Cogn. Sci. **1985** (1985), 399–421.

[61] T. Senst, V. Eiselein, A. Kuhn and T. Sikora, *Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation*, IEEE Trans. Inf. Foren. Secur. **2017** (2017), 2945–2956.

[62] S.R. Shakya, C. Zhang and Z. Zhou, *Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data*, Int. J. Mach. Learn. Comput. **8** (2018), no. 6, 577–582.

[63] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan and K. Jayavel, *A fully integrated violence detection system using CNN and LSTM*, Int. J. Electr. Comput. Eng. **11** (2021), no. 4, 3374–3380.

[64] C.S. Shivaraj, *Artificial intelligence for human behavior analysis*, Int. Res. J. Eng. Technol. **5** (2018), no. 6, 1863–1870.

[65] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 3rd Int. Conf. Learn. Represent. ICLR, Conf. Track Proc., 2015, p. 1–14.

[66] 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*, IEEE 9th Int. Conf. Intell. Comput. Info. Syst. ICICIS, 2019, p. 80–85.

[67] Y. Su, G. Lin, J. Zhu and Q. Wu, *Human interaction learning on 3d skeleton point clouds for video violence*

recognition, in European Conf. Comput. Vis. (2020), 74–90.

[68] S. Sudhakaran, O. Lanz and F.B. Kessler, *Learning to detect violent videos using convolutional long short-term memory*, 14th IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS), 2017, p. 1–6.

[69] T. Surasak, I. Takahiro, C.H. Cheng, C.E. Wang and P.Y. Sheng, *Histogram of oriented gradients for human detection in video*, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR'05), 2005, p. 886–893.

[70] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, *Rethinking the inception architecture for computer vision*, IEEE Conf. Comput. Vis. Pattern Recog., 2016, p. 2818–2826.

[71] R. Takahashi, T. Matsubara and K. Uehara, *Data augmentation using random image cropping and patching for deep CNNs*, IEEE Trans. Circuits Syst. Video Technol. **30** (2020), no. 9, 2917–2931.

[72] A. Traore and M.A. Akhloufi, *Violence detection in videos using deep recurrent and convolutional neural networks*, 2020 IEEE Int. Conf. Syst. Man. Cyber. (SMC), 2020, p. 154–159.

[73] F.U.M. Ullah, M.S. Obaidat, K. Muhammad, A. Ullah, S.W. Baik, F. Cuzzolin J.J. Rodrigues and V.H.C. de Albuquerque, *An intelligent system for complex violence pattern analysis and detection*, Int. J. Intell. Syst. **36** (2021), 1–23.

[74] F.U.M. Ullah, A. Ullah, K. Muhammad, I.U. Haq and S.W. Baik, *Violence detection using spatiotemporal features with 3D convolutional neural network*, Sensors (Switzerland), **19** (2019), no. 11, 1–15.

[75] S. Vento, F. Cainelli and A. Vallone, *Violence against healthcare workers: A worldwide phenomenon with serious consequences*, Front. Public Heal. **8** (2020), 541.

[76] S. Woo, J. Park, J. Lee and I.S. Kweon, *CBAM: Convolutional block attention module*, Eur. Conf. Comput. Vis. (ECCV), 2018, p.3–19.

[77] Q. Xie, M.T. Luong, E. Hovy and Q.V. Le, *Self-training with noisy student improves imagenet classification*, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2020, p. 10687–10698.

[78] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang and X. He, *A new method for violence detection in surveillance scenes*, Multimed. Tools Appl. **75** (2016), no. 12, 7327–7349.