Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

A 4D Convolutional Neural Networks for Video Violence Detection

Mai Magdy[1,*], Fahima A. Maghraby[1], Mohamed Waleed Fakhr[1]

[1] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Cairo P.O. Box 2033, Egypt

ARTICLE INFO

Article history:
Received 2 February 2023
Received in revised form 18 October 2023
Accepted 2 November 2023
Available online 25 December 2023

Keywords:
Surveillance Cameras; Computer Vision; Deep Leaning; Violence Detection

ABSTRACT

As global crime has escalated, surveillance cameras have become widespread and will continue to proliferate. Due to the large amount of video, there must be systems that automatically look for suspicious activity and send out an online alert if they find it. This paper presents a deep learning architecture based on video-level four-dimensional convolution neural networks. The suggested architecture consists of residual blocks, which are combined with three-dimensional Convolutional Neural Networks (3D CNNs). The architecture aims to learn short-term and long-term representations of spatiotemporal from video, in addition to interactivity between clips. ResNet50 serves as the foundation for three-dimensional convolution networks and Dense optical flow in the region of concern. The proposed architecture is tested on the RWF2000 dataset with a test accuracy of 94.75. This research achieved higher results compared to other methods in the state of the art.

1. Introduction

Surveillance Cameras are becoming more common in a variety of regions around the globe. so that organisations can keep an eye out for those with malicious intentions. Violence is on the rise, and surveillance cameras can help prevent it by detecting it. Violence detection is becoming increasingly popular in the computer vision world because surveillance cameras can gather evidence and identify suspected acts of violence. Automatically identifying recording crime scenes has become required since personnel monitoring in real time of a large volume of video material is prohibitively expensive. Due to the strong connection between frames, a series of successive frames can show a continuous movement. This means that video data has more time sequences than still images.

For recognizing human activity and detecting aggression in videos, there are essentially two kinds of strategies: traditional feature extraction and deep learning methods [1]. There is a drawback to some traditional approaches, though, in that they are tied to specific geographic places; therefore, this data can't be used. The cost of calculation is also a problem, as is the lack of workplace flexibility. Deep learning approaches for action recognition have been widely discussed due to their higher overall accuracy. Because of the complexity of the scene, it is challenging to extract a portion of the

* *Corresponding author.*
E-mail address: maimagdy@adj.aast.edu

https://doi.org/10.37934/araset.36.1.1625

most relevant features from a video using various deep learning approaches. Clip-based models usually ignore the long-range spatiotemporal dependence and video-level structure during training.

In this paper, a four-dimensional convolutional neural network is presented to detect violent content in videos, namely "Violent 4D." The main contribution of this work is a method for identifying violent content in videos to model long-term reliance more precisely. Violent 4D is comprised of two parts:

i. a uniform, comprehensive sampling technique that captures a series of short-term units across the entire video
ii. a four-dimensional convolutional interaction that evaluates long-range spatial correlation by applying a 4D residual block to collect inter-clip interactions.

Violent 4D achieved better performance than previous studies on the RWF2000 benchmark dataset [2]. The remainder of the study is arranged as follows:

i. Section 2: studies on violence detection
ii. Section 3: proposed architecture
iii. Section 4: dataset definitions and tests
iv. Section 5: discussion
v. Section 6: conclusion and future work.

2. Related work

This section summarises the current state-of-the-art strategies for identifying violence using traditional and deep learning algorithms. Deep learning has shown superior results, making it the more widely used approach. This section contains deep learning algorithms and traditional techniques, such as spatiotemporal descriptors and optical flow-based features, for recognising motion and aggression in video.

2.1 Traditional Approaches

2.1.1 Spatiotemporal descriptors

To detect violence, traditional methods rely on spatial-temporal descriptors and visual features generated by algorithms utilizing the Kohonen self-organizing map [4]. Action recognition hierarchies include initialization, recognition, tracking, and posture estimation [5]. There are several methods in [6] that use spatial and temporal descriptors, like Motion Scale-Invariant Feature Transform (MOSIFT), Space-Time Interest Points (STIPs), and Scale-Invariant Feature Transform (SIFT), which use a bag of features to characterise every video and SVM to classify it with independent kernels. These methods are referred to as spatiotemporal descriptors. Bag-of-Features frameworks have a high computational cost for retrieving spatial-temporal features. Disadvantageously, different space-time descriptors relate to the regions of important data that are skipped. This decreases the actual video output.

2.1.2 Optical flow-based features

With the use of spatiotemporal identifiers, researchers have been able to look at novel characteristics such as Lagrangian directed fields. Features from motion blobs include low-contrast

characteristics like the local histogram of optical flow [7], the local histogram of directed gradient [8], as well as the difference in magnitude between subsequent frames [9]. Optical flow intensity and direction histograms were included as new feature descriptors for this technique [10]. The VIF [3] has been proposed for the detection of violence in real time. For classification, traditional techniques employ SVMs, which are regarded as state-of-the-art.

2.2 Deep Learning Approaches

Convolutional networks [11] assess visual spatiotemporal properties, whereas LSTM layers [12] incorporate temporal information in deep learning algorithms for the recognition of violence. To obtain spatial information, [13] demonstrated a model including convolutional, batch normalisation, and pooling layers, in addition to minimization methods. Using a recurrent Convolutional Layer Long Short-Term Memory (ConvLSTM), the level changes or temporal properties that identify violent events are encoded. Then, a three-component method consisting of a spatial encoder, a classifier, and a temporal encoder was presented [14]. The Bidirectional Xception LSTM Attention model, a CNN structure with a bigger kernel size related to the structure of the Xception [15], is then used as a tool for extracting and recording the spatial properties of violent scenarios because a bidirectional LSTM can determine how data from the previous and current are connected, while taking temporal characteristics into consideration. In addition, an attention layer [16] is added to aid in distinguishing vital areas. In conjunction with bidirectional LSTM layers, the attention layer determines how likely it is that the videos show violent content. One-stream sequential methods have been examined where an optical flow or RGB format may be used as the input. Other methods, on the other hand, use convolutional multi-stream architectures; each one focuses on a different type of visual feature and accepts input in both forms. To extract spatial meanings [20], used a low-frame-rate slow pathway, while a high-frame-rate fast pathway recorded motion in high-resolution time-lapse [17] used VGG19, ResNet50, and VGG19 to capture attributes from video frames, then Long Short-Term Memory (LSTM) with an attention layer. To detect aggression, a unique schema of a convolutional recurrent neural network for long-short-term memory (ConvLSTM) was described in [18]. This approach makes use of ResNet50 to extract crucial information from the input from each frame. There were various issues with this approach in the clip, and in some cases, it couldn't see anybody. Transfer learning was used for the purpose of identifying aggressive human behaviour. 3D convolutional neural networks and optical flow are combined in [2] to create a flow-gated network. It's hard to pull out features from complicated situations, which makes it impossible to get some of the most important ones from the video. Different space-time descriptors have the drawback of being tied to those spots of key points, over which content is disregarded according to earlier techniques. As a result of a lack of workplace adaptability, coping with scenarios such as camera motions, ever-changing backdrops, and variances in appearance from the same category hinders productivity. 3D CNNs provide many parameters, each of which needs a massive quantity of training data for optimal performance.

3. Proposed Method

This part presents the Violent 4D architecture for the automatic violence detection in video. The primary objective is to construct a network that can recognise violent clips. Violent 4D uses four-dimensional residual blocks [19] to model three-dimensional spatiotemporal characteristics. As seen in Figure 1, Violent 4D consists of three primary sections: video sampling, learning spatiotemporal interactions, and training utilising the V4D CNN architecture. Converting three-dimensional CNNs to

four-dimensional enables the learning of long-range three-dimensional feature interactions and the capturing of inter-clip interactions. This allows existing three-dimensional CNNs to represent video at a higher level.

3.1 Video Sampling

A video sampling approach was employed to generate a representation for violent recognition. As both the short-term action information and total length of each input video have been included in the input. U-segments of action units are used to portray the full action in a video. Then, frames in each portion of the action unit are cropped based on the region of interest, utilising Gunner Farneback's [21] method for determining intensive optical flow between successive frames. Within training, each unit of action is selected randomly over every U segment. Then, a sparse sampling approach is employed, and RGB frames are cropped and resized to 224x224x3, where the third dimension comprises RGB channels. The centre of each action unit matches precisely to the segment that refers to it during testing. Accordingly, the input video is obtained by V= {a1, a2,…, au}, as ai ∈ $R^{CxTxHxW}$, where C indicates channel number, T denotes the temporal duration, and H and W represent, respectively, the height and width of the action unit.

3.2 Learning Spatiotemporal Interaction

Convolutions in four dimensions have been used in the study of long-range spatiotemporal interactions. In the Violence 4D tensors, convolution is passed as (C, U, T, H, W). Three-dimensional convolutions can be used to construct 4D convolutions using Eq. (1), according to [21]. This allows for the implementation of four-dimensional convolutions using three-dimensional convolutions.

$$o_j^{uthw} = b_j + \sum_{s=0}^{S-1}\left(\sum_c^{C_{in}} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} W_{jc}^{spqr} U_c^{(u+s)(t+p)(h+q)(w+r)}\right) \quad (1)$$

V, S, P, Q, and R represent the form of the kernel of a 4D convolutional, whereas S, P, Q, and R of W_{jc}^{spqr} represent the weight of the kernel at each position. Using a kernel of convolution, the short-term three-dimensional properties of a specific action unit can be expressed concurrently with the long-term temporal evolution of several units of action in four-dimensional space. Videos can be modelled in a 4D feature space, which enables them to understand more intricate links between long-range three-dimensional spatiotemporal representations.

3.3 Training using V4D CNNs Architecture

For violence recognition, the ResNet50 CNN architecture with 4D convolutions has been employed. The 4D residual block enables the collection of both short-term three-dimensional features and the long-term evolution of representations of spatiotemporal events, which is useful for video action identification.

The residual four-dimensional convolution block is defined as follows in Eq. (2):

$$y_{3D} = x_{3D} + \varphi_{(U,C)}\left(F_{4D}\left(\varphi_{(C,U)}(X_{3D}); W_{4D}\right)\right) \quad (2)$$

Three-dimensional CNNs can instantly process Y3D and X3D where X3D and Y3D ∈ RUxCxTxHxW. Using the permutation formula, φ, dimensions can be permuted. φ has been used for the dimension permutation of X3D's U C T H W to C U T H W so that it can be handled by 4D convolutions. Then the outputs of the 4D dimensions are permuted back to 3D to match X3D. Then, batch normalisation and Relu activation are implemented. Each action unit is individually and concurrently trained in three-dimensional convolution layers with identical parameter values. The 3D Feature action units are subsequently transmitted to the Residual 4D Block, which mimics the long-term temporal development of subsequent action units. As shown in Figure 1, global average pooling is a combination of all units of action from the video representation that are sent to two fully connected layers and softmax to detect violence.

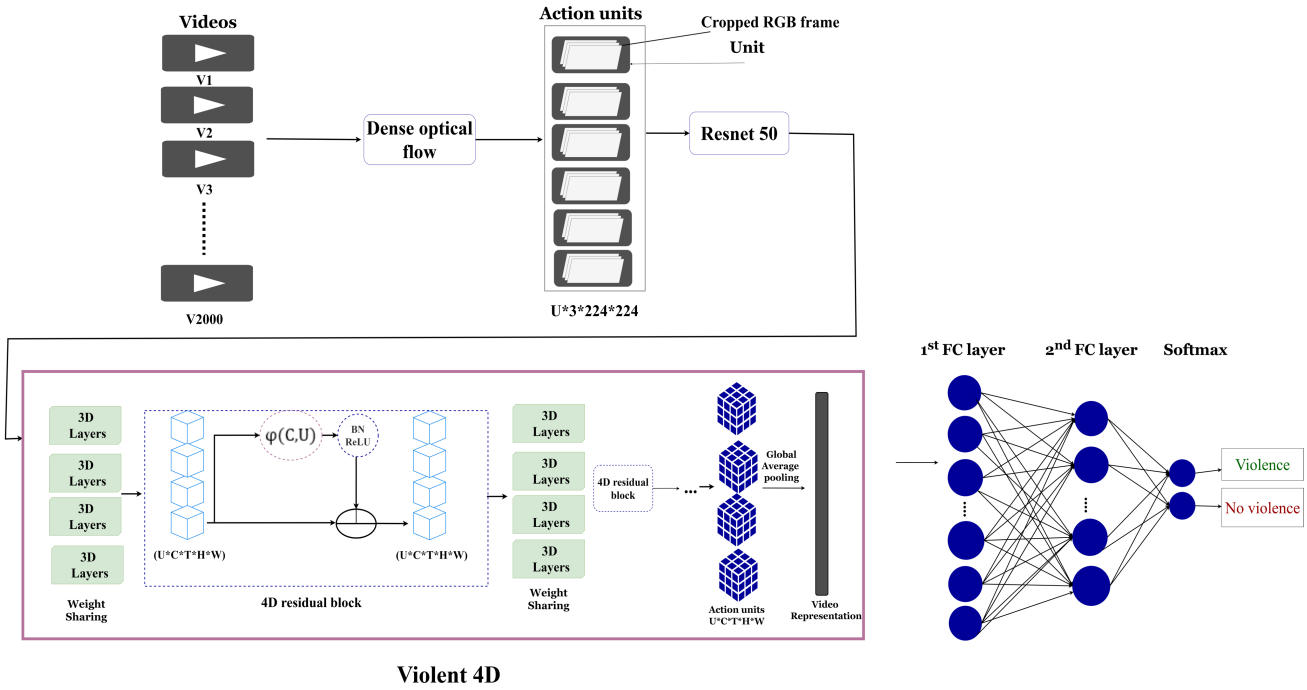

Fig. 1. Violent 4D architecture which illustrates the stages of: (1) Video sampling; (2) learning spatiotemporal interaction; and (3) training using 4D CNNs

4. Experimental Methodology

This part examines the effectiveness of Violent 4D in detecting violence in videos from the RWF2000 dataset.

4.1 Dataset

Various video datasets exist for the detection of violence, but they suffer from restrictions such as small size, limited variation, and inadequate imaging resolution. There are other datasets that come from movies that aren't close enough to the current world to be considered good quality. The RWF2000 [2] dataset consists of 2000 videos taken from YouTube in the real world. There are a broad variety of indoor and outdoor locations included in the RWF2000 dataset, including the street, numerous sporting event locations, and various rooms in a house, as well as a variety of severe weather situations. The videos are classified into two categories: 1000 violent and 1000 nonviolent clips with variable resolution. Figure 2 shows sample in violence in video and Figure 3 shows sample of non-violence in video in RWF2000 dataset.

Fig. 2. Sample of violence frames in video from RWF2000

Fig. 3. Sample of non-violence frames in video from RWF2000

4.2 Experiments and Results

First, RWF2000 was split into 80% and 20% for training and testing, respectively. with 1600 videos for training, divided into 800 violent and non-violent; 400 videos for testing, divided into 200 violent and non-violent. The primary step was to divide each video into U-action units, each with a specific number of frames. When adopting Farneback's approach [22], to get a region of interest, the output RGB frames are cropped and resized to 224x244. A Stochastic Gradient Descent (SGD) optimizer with varying learning rates and weight decay was then implemented with differing momentum. The model has been trained for 70 epochs with 3D ResNet50 as the basis with 8-frame inputs, and then the weights from 3D ResNet50 have been transferred to V4D ResNet50 with all 4D Blocks assigned to zero. A high accuracy of 94.75% was obtained after several tests in which the number of action units was set to 4, the number of frames in each action unit was set to 8, and the learning rate was set to 0.001, with weight decay equal to 5e-3 and momentum equal to 0.9. The 8 x 4 input frames are used to further refine the V4D ResNet50. After that, optimization is done on all 4D blocks. Figure 4 depicts the outcomes of training and validation.

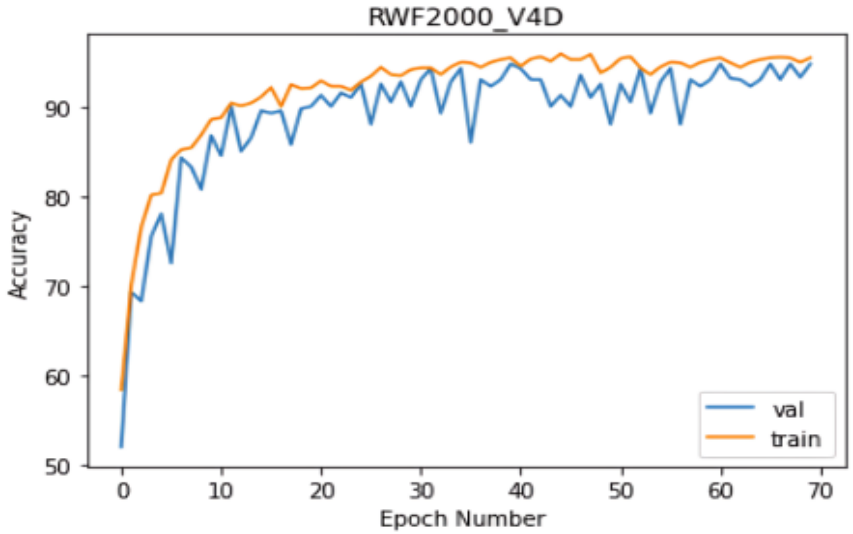

Fig. 4. Training and validation accuracies using RWF2000 with Violent 4D

Figure 5 illustrates testing losses.

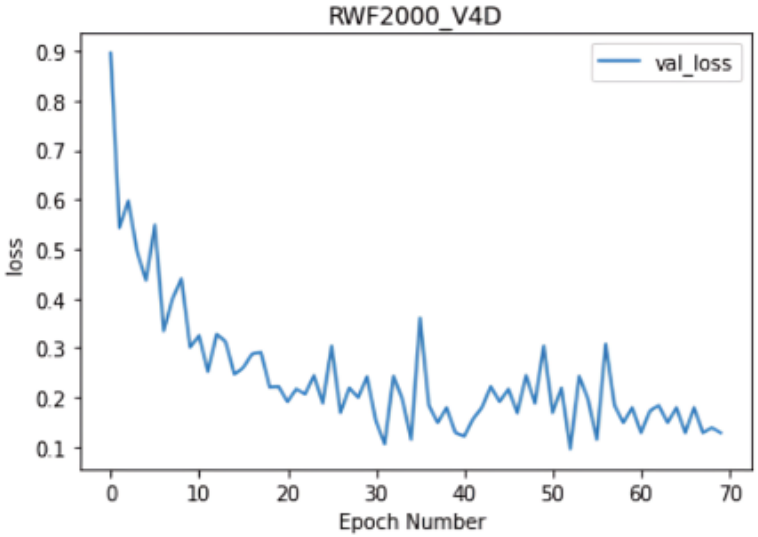

Fig. 5. Losses appeared in validation RWF2000 with Violent 4D

Figure 6 illustrates a confusion matrix of test data with accuracy of 0.9475, precision of 0.9408, recall of 0.9550, F1-Score of 0.9478, and specificity of 0.9400.

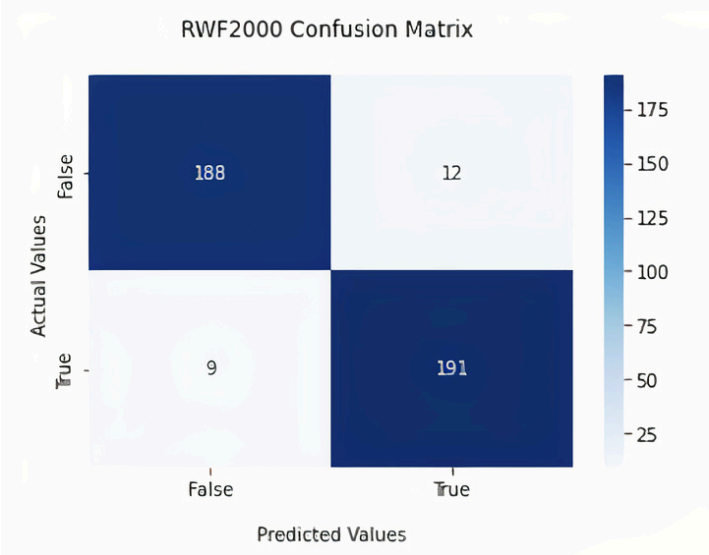

Fig. 6. Confusion matrix of RWF2000 dataset on test data

Table 1 compares the efficiency of several violence detection methods with a violent 4D model, which were evaluated using the RWF2000 dataset. By comparing violent 4D technique to the reported literature, certain previously used models, such as Flow Gated Net [23], are substantially heavier since the usage of two-stream convolution neural networks is difficult and costly owing to parallel optimization. Violent 4D can typically compete with other techniques in terms of accuracy. Since the proposed method integrates four-dimensional residual blocks with three-dimensional convolutional neural networks to train both short-term and long-term representations of spatiotemporal information, the RWF2000 dataset produced the most effective results. The Violent 4D technique is computationally efficient and accurate. The results obtained from constructing the confusion matrix for the testing videos from the RWF2000 dataset are represented as follows: 191 true positives (TP), 12 true negatives (TN), 188 false positives (FP), and 9 false negatives (FN). The following values were generated using the confusion matrix: accuracy = 0.9475, precision = 0.940886, recall = 0.95, F1 score = 0.9478, and specificity = 0.94. In conclusion, the suggested method yielded high result when examined against a complex dataset comprised of real-world information.

Table 1
Comparison between methods and Violent 4D on RWF2000 dataset

Method	RWF2000 dataset
Flow Gated Net [23]	87.25%
SepConvLSTM-M [24]	89.75%
SepConvLSTM-A [24]	87.75%
SepConvLSTM-C [24]	89.25%
SPIL Convolution [25]	89.30%
2D CNNs+LSTM [26]	92.00%
2D Spatio-Temporal Representations [27]	93.80%
Violence 4D	**94.75%**

5. Discussion

Violent 4D demonstrated more accuracy than comparable state-of-the-art approaches. The proposed Violent 4D method has obtained the best performance on the RWF2000 dataset, which is the largest surveillance violent video collection to date. A substantial amount of training data is

required for 3D CNNs to provide satisfactory outcomes. Parallel optimization makes using two-stream convolution neural networks challenging and costly. This promising result of Violent 4D on RWF2000 is due to using 3D CNNs (ResNet50) with four-dimensional residual blocks. This mixture facilitates the acquisition of both long-term and short-term representations of spatiotemporal information. Even though there are other video datasets that can be used to find violence, they are limited in size, variety, and image quality.

6. Conclusion and Future Work

Short- and long-term representations of spatiotemporal information utilising Violent 4D are presented in this article. The model uses ResNet50 as its basis and intensive optical flow to get the RGB frames' region of interest. 4D residual blocks are combined with 3D CNNs for long-term modelling. At the clip level, using 4D residual blocks improve the capability of representation of three-dimensional convolution neural networks. As a potential future development to Violent 4D, optical flow data can be combined with RGB channels and then mixed, utilising two streams and four-dimensional video convolution.

Acknowledgement
This research was not funded by any grant.

References
[1] Senst, Tobias, Volker Eiselein, Alexander Kuhn, and Thomas Sikora. "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation." *IEEE transactions on information forensics and security* 12, no. 12 (2017): 2945-2956. https://doi.org/10.1109/TIFS.2017.2725820

[2] Cheng, Ming, Kunjing Cai, and Ming Li. "RWF-2000: an open large scale video database for violence detection." In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4183-4190. IEEE, 2021. https://doi.org/10.1109/ICPR48806.2021.9412502

[3] Hassner, Tal, Yossi Itcher, and Orit Kliper-Gross. "Violent flows: Real-time detection of violent crowd behavior." In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 1-6. IEEE, 2012. https://doi.org/10.1109/CVPRW.2012.6239348

[4] Ramzan, Muhammad, Adnan Abid, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood. "A review on state-of-the-art violence detection techniques." *IEEE Access* 7 (2019): 107560-107575. https://doi.org/10.1109/ACCESS.2019.2932114

[5] Beddiar, Djamila Romaissa, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. "Vision-based human activity recognition: a survey." *Multimedia Tools and Applications* 79, no. 41-42 (2020): 30509-30555. https://doi.org/10.1007/s11042-020-09004-3

[6] Xu, Long, Chen Gong, Jie Yang, Qiang Wu, and Lixiu Yao. "Violent video detection based on MoSIFT feature and sparse coding." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3538-3542. IEEE, 2014. https://doi.org/10.1109/ICASSP.2014.6854259

[7] Zhou, Peipei, Qinghai Ding, Haibo Luo, and Xinglin Hou. "Violence detection in surveillance video using low-level features." *PLoS one* 13, no. 10 (2018): e0203668. https://doi.org/10.1371/journal.pone.0203668

[8] Das, Sunanda, Amlan Sarker, and Tareq Mahmud. "Violence detection from videos using hog features." In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1-5. IEEE, 2019. https://doi.org/10.1109/EICT48899.2019.9068754

[9] Serrano, Ismael, Oscar Deniz, Jose Luis Espinosa-Aranda, and Gloria Bueno. "Fight recognition in video using hough forests and 2D convolutional neural network." *IEEE Transactions on Image Processing* 27, no. 10 (2018): 4787-4797. https://doi.org/10.1109/TIP.2018.2845742

[10] Mahmoodi, Javad, and Afsane Salajeghe. "A classification method based on optical flow for violence detection." *Expert systems with applications* 127 (2019): 121-127. https://doi.org/10.1016/j.eswa.2019.02.032

[11] Meng, Zihan, Jiabin Yuan, and Zhen Li. "Trajectory-pooled deep convolutional networks for violence detection in videos." In *Computer Vision Systems: 11th International Conference, ICVS 2017, Shenzhen, China, July 10-13, 2017,*

Revised Selected Papers 11, pp. 437-447. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-68345-4_39

[12] Dong, Zhihong, Jie Qin, and Yunhong Wang. "Multi-stream deep networks for person to person violence detection in videos." In *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part I 7*, pp. 517-531. Springer Singapore, 2016. https://doi.org/10.1007/978-981-10-3002-4_43

[13] Sudhakaran, Swathikiran, and Oswald Lanz. "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)*, pp. 1-6. IEEE, 2017. https://doi.org/10.1109/AVSS.2017.8078468

[14] Hanson, Alex, Koutilya Pnvr, Sanjukta Krishnagopal, and Larry Davis. "Bidirectional convolutional lstm for the detection of violence in videos." In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0-0. 2018. https://doi.org/10.1007/978-3-030-11012-3_24

[15] Aktı, Şeymanur, Gözde Ayşe Tataroğlu, and Hazım Kemal Ekenel. "Vision-based fight detection from surveillance cameras." In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1-6. IEEE, 2019. https://doi.org/10.1109/IPTA.2019.8936070

[16] Rendón-Segador, Fernando J., Juan A. Álvarez-García, Fernando Enríquez, and Oscar Deniz. "Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence." *Electronics* 10, no. 13 (2021): 1601. https://doi.org/10.3390/electronics10131601

[17] Sumon, Shakil Ahmed, Raihan Goni, Niyaz Bin Hashem, Tanzil Shahria, and Rashedur M. Rahman. "Violence detection by pretrained modules with different deep learning approaches." *Vietnam Journal of Computer Science* 7, no. 01 (2020): 19-40. https://doi.org/10.1142/S2196888820500013

[18] Vosta, Soheil, and Kin-Choong Yow. "A cnn-rnn combined structure for real-world violence detection in surveillance cameras." *Applied Sciences* 12, no. 3 (2022): 1021. https://doi.org/10.3390/app12031021

[19] Zhang, Shiwen, Sheng Guo, Weilin Huang, Matthew R. Scott, and Limin Wang. "V4d: 4d convolutional neural networks for video-level representation learning." *arXiv preprint arXiv:2002.07442* (2020).

[20] Haoqi Fan, Jitendra Malik, Christoph Feichtenhofer, and Kaiming He,2018. "Video recognition using slow-fast networks," in CoRR. https://arxiv.org/abs/1812.03982

[21] Yang, Tao, Dongdong Li, Yi Bai, Fangbing Zhang, Sen Li, Miao Wang, Zhuoyue Zhang, and Jing Li. "Multiple-object-tracking algorithm based on dense trajectory voting in aerial videos." *Remote Sensing* 11, no. 19 (2019): 2278. https://doi.org/10.3390/rs11192278

[22] Gupta, Arpan, and M. Sakthi Balan. "Action recognition from optical flow visualizations." In *Proceedings of 2nd International Conference on Computer Vision & Image Processing: CVIP 2017, Volume 1*, pp. 397-408. Springer Singapore, 2018. https://doi.org/10.1007/978-981-10-7895-8_31

[23] Cheng, Ming, Kunjing Cai, and Ming Li RWF. "2000: An open large scale video database for violence detection." In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4183-4190. 2021. https://doi.org/10.1109/ICPR48806.2021.9412502

[24] Islam, Zahidul, Mohammad Rukonuzzaman, Raiyan Ahmed, Md Hasanul Kabir, and Moshiur Farazi. "Efficient two-stream network for violence detection using separable convolutional lstm." In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE, 2021. https://doi.org/10.1109/IJCNN52387.2021.9534280

[25] Su, Yukun, Guosheng Lin, Jinhui Zhu, and Qingyao Wu. "Human interaction learning on 3d skeleton point clouds for video violence recognition." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 74-90. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-58548-8_5

[26] Kang, Min-Seok, Rae-Hong Park, and Hyung-Min Park. "Efficient spatio-temporal modeling methods for real-time violence recognition." *IEEE Access* 9 (2021): 76270-76285. https://doi.org/10.1109/ACCESS.2021.3083273

[27] Chelali, Mohamed, Camille Kurtz, and Nicole Vincent. "Violence detection from video under 2D spatio-temporal representations." In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2593-2597. IEEE, 2021. https://doi.org/10.1109/ICIP42928.2021.9506142