Convolutional Long Short Term Memory based IOT node for Violence Detection

Nouar AlDahoul*
Faculty of Engineering
Multimedia University
Cyberjaya, Selangor, Malaysia
corresponding author:
nouar.aldahoul@live.iium.edu.my

Hezerul Abdul Karim
Faculty of Engineering
Multimedia University
Cyberjaya, Selangor, Malaysia
hezerul@mmu.edu.my

Rishav Datta
Electrical and Electronics Engineering
Birla Institute of Technology and Science, Pilani, Hyderabad Campus
Hyderabad, India
f20180910@hyderabad.bits-pilani.ac.in

Shreyash Gupta
Electronics, and Communications Engineering
Birla Institute of Technology and Science, Pilani, Hyderabad Campus
Hyderabad, India
f20180444@hyderabad.bits-pilani.ac.in

Kashish Agrawal
Electrical and Electronics Engineering
Birla Institute of Technology and Science, Pilani, Hyderabad Campus
Hyderabad, India
f20180685@hyderabad.bits-pilani.ac.in

Ahmad Albunni
Artificial Intelligence Department
Yo-Vivo Corporation
Philippines
ahmadbunni@gmail.com

Abstract— Violence detection has been investigated extensively in the literature. Recently, IOT based violence video surveillance is an intelligent component integrated in security system of smart buildings. Violence video detector is a specific kind of detection models that should be highly accurate to increase the model's sensitivity and reduce the false alarm rate. This paper proposes a novel architecture of ConvLSTM model that can run on low-cost Internet of Things (IOT) device such as raspberry pi board. The paper utilized convolutional neural networks (CNNs) to learn spatial features from video's frames that were applied to Long Short-Term Memory (LSTM) for video classification into violence/non-violence classes. A complex dataset including two public datasets: RWF-2000 and RLVS-2000 was used for model training and evaluation. The challenging video content includes crowds and chaos, small object at far distance, low resolution, and transient action. Additionally, the videos were captured in various environments such as street, prison, and schools with several human actions such as playing football, basketball, tennis, swimming and eating. The experimental results show high performance of the proposed violence detection model in terms of average metrics having an accuracy of 73.35 %, recall of 76.90 %, precision of 72.53 %, F1 score of 74.01 %, false negative rate of 23.10 %, false positive rate of 30.20 %, and AUC of 82.0 %.

Keywords— Convolutional Neural Network – Internet of Things - Violence Video Detection -Long Short-Term Memory - Surveillance System

I. INTRODUCTION

Over the last few decades, the rate of violent criminal activities has been increased and made the businesses, government and law enforcement agencies motivated to use surveillance systems to identify dangerous environments and respond immediately to the violent actions [1]. The traditional surveillance solution is based on CCTV cameras for human supervision and monitoring purposes. The previous task was performed by a limited number of security staff members to monitor huge quantities of CCTV footage [1]. Therefore, several factors such as worker fatigue, boredom, and discontinuity of observation make the traditional solution unreliable. One of the challenging issues of security systems in commercial buildings such as shopping malls [2] is violence detection that should be able to operate in various environments to detect fights from surveillance cameras and control the aggressive and violent incidents. The environments include outdoor (street and public spaces) and indoor (markets, banks, schools, and prison) with several human actions such as playing football, basketball, tennis, swimming and eating.

"Eye in the Sky" which is a camera-equipped drone for automated surveillance technology can identify violent behavior in crowds and transmit video footage for real-time analysis [3,4]. This drone has an algorithm trained using deep learning to estimate the poses of humans in the video and match them to "violent" postures. It can detect crime in public spaces and at large events [3,4].

Violence could happen in different scenarios and places. Therefore, violence could be detected in various methods. Several research works have been investigated to find efficient solutions to detect humans and recognize their activities by applying computer vision and artificial intelligence methods such as the HOG algorithm with SVM for object "human" detection [5] and spatiotemporal analysis of MoSIFT action descriptors with k-mean clustering for human action recognition in surveillance system [6]. Additionally, Bag-of-Words framework has been used with two action descriptor: STIP and MoSIFT for action recognition and it was proved that MoSIFT was better than STIP in all conditions [7]. To address the problem of abnormal behavior detection, a brute force detection method based on the combination of convolutional neural network and trajectory features was demonstrated to extract the spatiotemporal features from the video [8]. More traditional feature methods such as Improved Fisher Vectors (IFV) were used to represent a video utilizing both local features and their Spatio-temporal positions, for detecting violence in videos [9].

Support vector machine (SVM) was used with a linear kernel to detect the violence in short videos. The inputs of SVM were features (25 skeleton keypoints, 6 angels and human contact detection) extracted from pose estimation algorithm [10], which is a key tool for analyzing human action from video. SVM with an RBF kernel was also used to classify the human poses and identify individuals' actions such as hands up and lying down [11]. To learn complex motion structures, recurrent pose-attention network (RPAN) was used to learn human-part features by sharing attention parameters partially on the semantically related human joints [12].

The 3D point clouds method [13] was used to extract action features from human skeleton points which were applied at the SPIL module to model the interactions between skeleton points. This SPIL module performed information propagation based on the assigned different weights to different skeleton points. Finally, a global feature was used for final classification [13].

Deep learning was utilized for monitoring the behavior of players and watchers to prevent and control the violence inside the stadium directly [14]. There are probabilities of fight occurrence between the players from one side and between watchers from the other side. A camera sensor was used to capture big size data captured during long time to be utilized with spark framework. HOG was able to extract the features from the frames which were labeled for training the Bidirectional LSTM model [14]. Additionally, Flow Gated network utilized 3D-CNNs with optical flow to detect violence using surveillance cameras of the public streets [15]. To avoid the high computational cost of optical flow, the dynamic images were used by convolutional networks instead of optical flow to represent motion for detecting and localizing the violence [16]. When the weak supervision with week labels are available, the violence can be detected by three parallel branches neural network [17]. Both audio and visual data can be used for violence detection [17].

Convolutional LSTM has been developed recently for detecting the violence from surveillance video, Convolutional neural network was used to extract the features from the video's frames. The whole set of extracted features were aggregated by a long short-term memory for violence/non-violence classification [18,19,20,21]. Both CNN and LSTM were capable to capture spatiotemporal features that enable the analysis of local motion in the video. Various Pre-trained CNNs such as VGG16 [19,21], VGG19 [20,21] and ResNet50 [21] were utilized to extract the spatial features before adding LSTM for violence/non-violence classification. The integration of Xception model, BiLSTM, and attention was found to improve the state-of-the-art accuracy for fight scene classification [22].

3D CNNs with 3D convolutional filters were found to replace 2D CNNs within the framework of residual learning (ResNet) [23]. Both spatial and temporal components produced significant gains in accuracy. End-to-end 3D CNN method was also used for detecting fights scenes in videos [24]. The architecture of DenseNet was utilized to represent abstract spatiotemporal features with a fewer number of parameters to develop a computational efficient and real-time processing model [24].

RNN-based (ST-LSTM) for 3D action recognition was explored to analyze the 3D location of each individual joint in each video frame network [25]. A skeleton tree traversal algorithm was found to take the adjacency graph of body joints.

Public datasets that have small size of samples were used extensively in the literature. The datasets are summarized as follows: Hockey Fight with 1000 videos [26], movie dataset with 200 videos [27], and violence-flow with 246 videos [28] datasets. Recent challenging datasets such as RWF with 2000 videos [15] and RLVS with 2000 videos [19] have not demonstrated widely with recent methods. This paper utilized a combination of these two datasets for training and evaluation of the proposed ConvLSTM.

Despite the promising results of accuracy rates, current existing methods are still memory demanding, and computational expensive. In other words, they are inapplicable for real purposes, particularly surveillance, where low memory and high detection speed solutions are required. This paper addresses this research gap and proposes a specific architecture of ConvLSTM that can run on IOT node such as Raspberry PI with limited memory and computation.

This paper is organized as follows: In section 2, the methodology including datasets, CNN model, LSTM model, and ConvLSTM architecture was described. Section 3 discusses the experimental setup, and results. Section 4 summarizes the conclusion and future work.

II. MATERIALS AND METHODS

In this section, the datasets that were used in this work are reported. Additionally, the methods employed are described. A brief review of End-to-End ConvLSTM is highlighted to shed light on the approach of training deep learning model such as ConvLSTM from scratch with violence datasets.

A. Dataset Overview

In this research work, two datasets including RWF-2000 [15] and Real-Life Violence Situations (RLVS) [19] were merged to compose a complex violence dataset with a large number of videos (4000 videos). These videos were divided into 2000 violence and 2000 non-violence samples. These videos were utilized to train, evaluate, and test the End-to-End ConvLSTM model.

These two datasets are summarized as follows:

1) RWF-2000 which is an open large scale video database including 2,000 videos captured by surveillance cameras in real-life situations. It contains 1000 instances class of violence and another 1000 of non-violence videos.

2) Real life violence Situations (RLVS) Dataset [19] which has 2000 videos: 1000 of violence and 1000 of non-violence. It contains videos taken from YouTube consisting of several street fights situations for people varied in race, age, and gender with different environments. Furthermore, it includes non-violence videos taken from many different human actions like sports, eating, walking etc. Fig.1 and Fig. 2 illustrate a few samples of violence and non-violence classes.

The 4000 videos violence dataset was shuffled and divided into 2400, 800 and 800 videos for training, validation, and testing, respectively. The dataset used in this work is summarized in TABLE I.

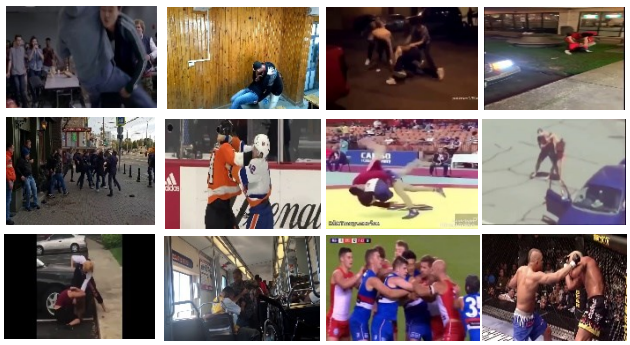

Fig. 1. Samples of violence frames from the used dataset [15,19].

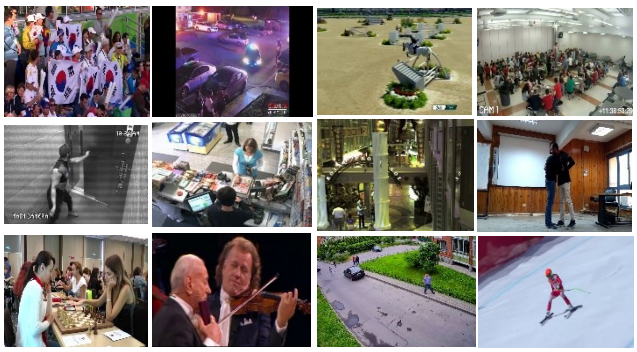

Fig. 2. Samples of non-violence frames from the used dataset [15,19].

TABLE I. Summary of dataset

Datasets	Usage	Number of videos
RWF-2000 [15] + RLVS-2000 [19]	Training	2400
	Validation	800
	Testing	800

B. Data pre-processing and Augmentation

The videos in training datasets (RWF-2000 and RLVS dataset) were converted to a sequence of frames or images before being passed to the proposed ConvLSTM model**.** As the dataset contained videos from the internet, they were limited to 30 frames per second. Each video contains an average of 150 frames. From each video, 10 frames were taken in a periodic interval and then resized to 112 × 112 pixels. Furthermore, the extracted frames were converted from BGR colour space to RGB space.

Data augmentation was used to increase the variability of training data and prevent the model from learning irrelevant patterns. The pixel intensities of each frame were rescaled by a factor of 1/255. After that, the rescaled images in the dataset were then flipped both horizontally and vertically, further they were randomly rotated and magnified for broadening the diversity of the dataset.

C. Convolutional Neural Networks (CNNs)

CNNs are a subset of Deep Neural Network, mostly applied in the problems of image classification. They are recognized as regularized versions of multi-layer perceptron. CNN algorithms take image or frame as an input and prioritize various aspects in the image by assigning weights and biases to differentiate image from another [29]. CNN was designed to learn spatial hierarchies of features automatically and adaptively through a backpropagation algorithm [29]. Each input image is passed through a series of layers including convolution with filters, pooling, and fully connected layers (FC). The Softmax function is applied in the final classification layer to classify an image with probabilistic values between 0 and 1.

D. LSTM

LSTM is a special type of Recurrent Neural Network (RNN) that has been used for long-range sequence modeling [30]. LSTM has a memory cell, which acts as an accumulator of state information, supported by control gates. The advantage of this structure is that it speeds down the gradient vanishing [30].

E. ConvLSTM architecture

In the convolutional neural network, ten convolutional (conv) layers were utilised. The number of filters in the first two conv layers were 64. Additionally, the number of filters in other conv layers were 128. At the end of each convolutional layer, ‘ReLU’ activation function was used. After every two layers, batch normalization and pooling were added in order to minimize the overfitting. To treat the sequence of the feature maps from CNN model in a chronological notion, LSTM layer was applied to filter out useful values from our sequential input. The LSTM layer has 64 nodes with two fully connected layers including 1024 nodes with ReLU function and 2 nodes with softmax function. Fig.3 shows the architecture of ConvLSTM.

Best Architecture of CNN
Conv2D(64, (3,3),'relu')
Conv2D(64, (3,3),'relu')
BatchNormalization(momentum=0.9)
MaxPool2D
Conv2D(128, (3,3), 'relu')
Conv2D(128, (3,3), 'relu')
BatchNormalization(momentum=0.9)
MaxPool2D
Conv2D(128, (3,3), 'relu')
Conv2D(128, (3,3), 'relu')
BatchNormalization(momentum=0.9)
MaxPool2D
Conv2D(128, (3,3), 'relu')
Conv2D(128, (3,3), 'relu')
BatchNormalization(momentum=0.9)
MaxPool2D
Conv2D(128, (3,3), 'relu')
Conv2D(128, (3,3), 'relu')
BatchNormalization(momentum=0.9)
GlobalAveragePooling2D

Best Architecture of LSTM
time_distributed
LSTM (64)
Dense (1024, 'relu')
Dropout (0.5)
Dense (2, 'softmax')

Fig. 3. Best Architecture of CNN and LSTM (ConvLSTM)

The ConvLSTM model was trained using the following settings:

1) number of epochs is 50
2) batch size is 8

3) Optimizer is Adam
4) Learning rate is 10^{-4}
5) Momentum is 0.9

Total number of parameters in ConvLSTM was 1,265,986 which can fit the limited memory size in Raspberry PI.

F. Raspberry pi as IOT node

Raspberry pi is a low-cost and programmable computer that includes a set of General-Purpose Input Output pins and can be used to connect and control external electronic devices and create Internet of Things (IoT) solutions. The final version is Raspberry Pi 3 Model B+ includes 1.4GHz 64-bit quad-core processor, dual-band wireless LAN, Bluetooth 4.2/BLE, and faster Ethernet. In this research work, Raspberry pi 3 Model B+ was used to deploy and run the proposed ConvLSTM for surveillance purposes. The proposed architecture was selected carefully to consider the limited memory and computational capability of the Raspberry PI board.

In this paper, we used RWF- RLVS-4000 videos dataset for model training and testing. The training was done as follows: Firstly, spatial features were learned from violence frames using CNN. Secondly, the LSTM was trained to fit the extracted spatial features, learn temporal features, and map videos to two categories: violence and non-violence. The complete End-to-End ConvLSTM model was tested on Raspberry pi which is connected to Camera to capture series of pictures and Internet to deliver alarm in real time. The block diagram of the proposed violence detection system is shown in Fig. 4.

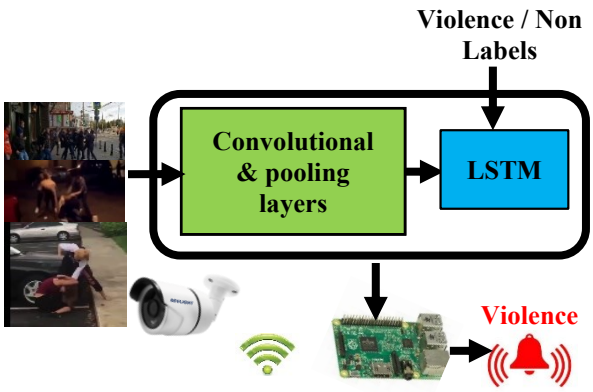

Fig. 4. The block diagram of the proposed violence detection system

III. EXPERIMENTS AND RESULTS DISCUSSION

A. Experimental Setup

The experiments discussed in this paper were performed on Google Colab Pro based GPU Tesla P100. These Tensor Processing Units are used to accelerate operations and each TPU has 180 teraflops of floating-point performance and 26 GB of high bandwidth memory. The frameworks used in this paper are sklearn, tensorflow, keras and OpenCV. Furthermore, matplotlib and seaborn were used for visualization purposes. Additionally, numpy and pandas were used for basic arithmetic and data manipulation tasks.

The average accuracy, recall, precision, F1 score, FPR, FNR. And AUC were calculated for each fold of dataset using k=5 cross validation. In other words, the dataset was divided into k=5 sets. Four sets out of five were utilized after being augmented for model training and the remaining set was used for testing. This process was repeated five times with each used 3200 augmented videos for training and validation and 800 videos for testing. From 3200 videos of training, 800 videos were used for validation. The average evaluation metrics of five experiments were considered as shown in figures 5,6,7 and Table II.

B. Evaluation Merics

Several performance metrics were used to validate the performance of the violence/non-violence classification model. A brief description of the Accuracy, Precision, Recall, F1 score, FPR and FNR, and AUC terms are as follows:

1) Accuracy is the ratio of the correctly labelled subjects to the whole pool of subjects.
2) Precision is the ratio of the correctly labelled "Violent videos" to the total number of labelled "Violent Videos".
3) Recall is the ratio of the correctly labelled "Violent Videos" to all the videos which are actually violent.
4) F1- score is the harmonic mean of Precision and Recall.
5) False Positive Rate is the ratio of the number of videos marked wrongly as Violent over the total number of "Non-Violent Videos".
6) False Negative Rate is the ratio of the number of videos marked wrongly as non-violent over the actual number of "Violent Videos".
7) Area Under the Curve (AUC): this metric is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. Higher AUC implies that the model performs better at distinguishing between non-violence and violence videos.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 score} = \frac{2 \times precision \times recall}{precision + recall} \qquad FNR = \frac{FN}{TP+FN} \qquad \text{FPR} = \frac{FP}{TN+FP}$$

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

C. Experimental Results

In this section, we examine the performance of ConvLSTM for the classification of videos into violence and non-violence categories.

The experiment was carried out to demonstrate the good performance of the proposed ConvLSTM that has about 1.2 million parameters and can run on low-cost Raspberry PI. The model was trained five times within 50 epochs. The learning curve of accuracy vs epochs was illustrated in Fig 5. The five trained models were tested with testing videos. The results show the average of 5 folds cross validation in Fig 7 and Table II.

This paper replicated the same experiment targeting k=5 cross validation to find the average evaluation metrics. Fig 6.

shows confusion matrix of the ConvLSTM model with the best fold. It is obvious that 327 videos from 400 were classified correctly as violence. In other words, only 73 videos were misclassified as non-violence.

Fig. 5. Learning convergence curve with accuracy vs epochs

TABLE II. Summary of dataset

K =5 cross validation	Accuracy	Precision	Recall	F1-Score	FPR	FNR
Fold 1	76.38	72.98	83.75	78.00	31.00	16.25
Fold 2	72.50	72.28	73.00	72.64	28.00	27.00
Fold 3	71.13	77.89	59.00	67.14	16.75	41.00
Fold 4	69.38	64.33	87.00	73.96	48.25	13.00
Fold 5	77.38	75.17	81.75	78.32	27.00	18.25
Average	**73.35**	**72.53**	**76.90**	**74.01**	**30.20**	**23.10**
Standard Deviation	**3.060**	**4.545**	**10.08**	**4.088**	**10.22**	**10.08**

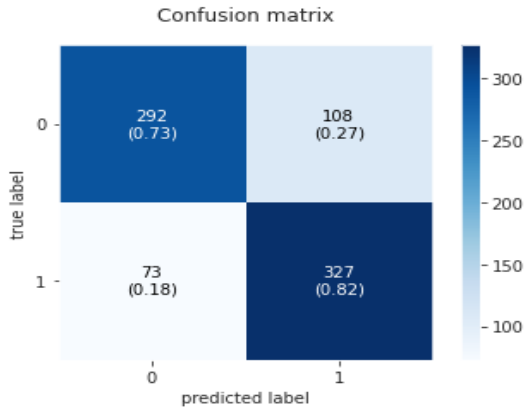

Fig. 6. Confusion Matrix of the proposed ConvLSTM model with the best fold.

From the results, we can deduce that high-level features or representations (objects and shapes) and low-level features (colours, edges, textures) were learned from violence and non-violence frames. It was found that the ConvLSTM model trained using violence dataset has good performance with an accuracy of 73.35 %, recall of 76.90 %, precision of 72.53 %, F1 score of 74.01 %, false negative rate of 23.10 %, false positive rate of 30.20 %, and AUC of 82.0 % as shown in Table II. Fig.7 demonstrates the ROC curve and AUC of the five folds with their average.

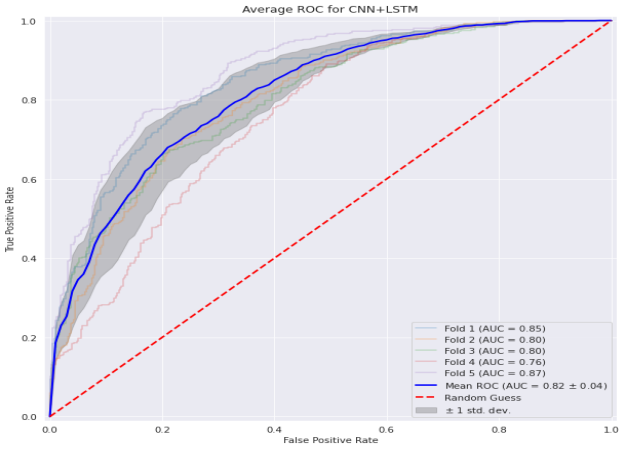

Fig. 7. ROC curve and AUC for the five folds with their average

IV. CONCLUSION AND FUTURE WORK

Most of the existing violence detection methods in literature review depend on using very deep learning models which are computationally expensive and memory hungry to tune and store a large number of model's parameters, respectively. This paper contributed to the body of knowledge and proposed a novel deep learning ConvLSTM architecture that can run on IOT node including raspberry pi board connected to a video camera to detect violent behavior and deliver detection results and alarms in real-time.

This paper proposed a novel architecture of convolutional neural network to learn spatial features from video's frames. Additionally, Long Short-Term Memory was utilized to learn the temporal features to classify the learned spatio-temporal features into violence and non-violence classes.

The contribution of this paper is summarized as follows:

- A novel video surveillance system for violence detection evaluated using challenging public video datasets: RWF-2000 and RLVS 2000. A complex dataset which combines these two datasets and includes crowds and chaos, small object at far distance, low resolution, and transient action. Additionally, the videos were captured in various environments such as street, prison, and schools with several human actions such as playing football, basketball, tennis, swimming and eating. The results showed good performance with an accuracy of 73.35 %, recall of 76.90 %, precision of 72.53 %, F1 score of 74.01 %, false negative rate of 23.10 %, false positive rate of 30.20 %, and AUC of 82.0 %.
- The proposed architecture of ConvLSTM has only 1,265,986 parameters which fit the limited memory size and conventional CPU in the Raspberry pi used as IOT node.

The CNN utilized in this paper can be designed in future works with recent types of pre-trained deep CNN such as MobileNet [31], and EfficientNet B0 [32] which have good performance and low number of parameters at the same time.

ACKNOWLEDGMENT

This research was fully funded by Multimedia University, Malaysia

REFERENCES

[1] Abto Software, "Violence Detection for Smart Surveillance Systems", Online Available: https://www.abtosoftware.com/blog/violence-detection , Accessed 26 May 2021

[2] S. Mirgani, "Target Markets – International Terrorism Meets Global Capitalism in the Mall". *Chapter Title: "Securing the Shopping Mall", Book Title: "Target Markets",* 2017, https://www.jstor.org/stable/j.ctv1fxdv4.7l

[3] Vincent, J, "Drones taught to spot violent behavior in crowds using AI", 2018, Online Available: https://www.theverge.com/2018/6/6/17433482/ai-automated-surveillance-drones-spot-violent-behavior-crowds, Accessed 26 May 2021

[4] C. Z. Gao and W. Bao, "Research on Video Violence Detection Technology of UAV on Cloud Platform," in *Communications in Computer and Information Science*, 2020, vol. 1252 CCIS, doi: 10.1007/978-981-15-8083-3_33.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, 2005, vol. 1, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[6] M. Chen and A. Hauptmann, "MoSIFT : Recognizing Human Actions in Surveillance Videos," *Informedia@TRECVID*, 2009.

[7] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Lecture Notes in Computer Science*, 2011, vol. 6855 LNCS, no. PART 2, doi: 10.1007/978-3-642-23678-5_39.

[8] P. Wang, P. Wang, and E. Fan, "Violence detection and face recognition based on deep learning," *Pattern Recognit. Lett.*, vol. 142, 2021, doi: 10.1016/j.patrec.2020.11.018.

[9] P. Bilinski and F. Bremond, "Human violence recognition and detection in surveillance videos," 2016, doi: 10.1109/AVSS.2016.7738019.

[10] D. Nova, A. Ferreira, and P. Cortez, "A Machine Learning Approach to Detect Violent Behaviour from Video," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, 2019, vol. 273, doi: 10.1007/978-3-030-16447-8_9.

[11] P. G. S. do C. Soares, A. B. Da Silva, and L. F. A. Pereira, "An assault detection system based on human Pose Tracking for video surveillance," 2019, doi: 10.5753/sibgrapi.est.2019.8327.

[12] W. Du, Y. Wang, and Y. Qiao, "RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-October, doi: 10.1109/ICCV.2017.402.

[13] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12349 LNCS, doi: 10.1007/978-3-030-58548-8_5.

[14] S. R. Dinesh Jackson *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Comput. Networks*, vol. 151, 2019, doi: 10.1016/j.comnet.2019.01.028.

[15] M. Cheng, K. Cai, and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection," 2021, doi: 10.1109/icpr48806.2021.9412502.

[16] D. G. C. Roman and G. C. Chavez, "Violence Detection and Localization in Surveillance Video," 2020, doi: 10.1109/SIBGRAPI51738.2020.00041.

[17] P. Wu *et al.*, "Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision," in *Lecture Notes in Computer Science*, 2020, vol. 12375 LNCS, doi: 10.1007/978-3-030-58577-8_20.

[18] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 2017, doi: 10.1109/AVSS.2017.8078468.

[19] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," 2019, doi: 10.1109/ICICIS46948.2019.9014714.

[20] U. M. butt, S. Letchmunan, F. H. Hassan, S. Zia, and A. Baqir, "Detecting video surveillance using VGG19 convolutional neural networks," *Int. J. Adv. Comput. Sci. Appl.*, no. 2, 2020, doi: 10.14569/ijacsa.2020.0110285.

[21] S. A. Sumon, R. Goni, N. Bin Hashem, T. Shahria, and R. M. Rahman, "Violence Detection by Pretrained Modules with Different Deep Learning Approaches," *Vietnam J. Comput. Sci.*, vol. 07, no. 01, 2020, doi: 10.1142/s2196888820500013.

[22] S. Akti, G. A. Tataroglu, and H. K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras," Ninth International Conference on Image Processing Theory, Tools and Applications, 2019, doi: 10.1109/IPTA.2019.8936070.

[23] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Lecun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition,* 2018, pp. 6450– 6459, doi: 10.1109/CVPR.2018.00675.

[24] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3D convolutional neural networks," 2019, doi: 10.1109/AVSS.2019.8909883.

[25] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Lecture Notes in Computer Science*, 2016, vol. 9907 LNCS, doi: 10.1007/978-3-319-46487-9_50.

[26] E. B. Nievas, et al, "Hockey fight detection dataset," *in Computer Analysis of Images and Patterns. Springer*, 2011, pp. 332–339.

[27] E. B. Nievas, et al, "Movies fight detection dataset," *in Computer Analysis of Images and Patterns. Springer*, 2011, pp. 332–339.

[28] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," 2012, doi: 10.1109/CVPRW.2012.6239348.

[29] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., 1997, doi: 10.1162/neco.1997.9.8.1735.

[31] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861,* 2017.

[32] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019*.