Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 2 Issue 2, October 2020***
(https://fcihib.journals.ekb.eg)

Violence Detection in Surveillance Videos Using Deep Learning

Mostafa mohamed moaaz
Computer science department,
Faculty of Computers and Artificial Intelligence,
Helwan university, Cairo, Egypt
mostafa_20160431@fci.helwan.edu.eg

Ensaf Hussein Mohamed*
Computer science department,
Faculty of Computers and Artificial Intelligence,
Helwan university, Cairo, Egypt
ensaf_hussein@fci.helwan.edu.eg

***Abstract*— Nowadays computer technologies are flowering especially the artificial intelligence field. It lives its prosperous years. Recently it closes the gap between humans and machines with the facilitation of supporting decisions. One of these gaps is the surveillance cameras labors' attentiveness and the lack of instantaneous detection of violence actions on the scenes of such cameras. In this paper we present an end to end deep neural network to detect the violence scenes in the surveillance cameras, the proposed system composed of set of phases. It extracts a set of selectively distributed frames of the video clip, performs spatio-temporal features, and passes them to a fully connected neural to classify the video to violence or non-violence action. The model is evaluated on different datasets; like Real Life Violence Situations aka RLVS and Hockey Fight Detection datasets. The accuracy was 92% and 94.5% respectively, which outperformed the previous related works.**

Keywords: Anomaly Detection, Deep Learning, Surveillance Video, Feature Extraction, Classification.

I. INTRODUCTION

Anomaly is that some instance deviates from the standard format of the majority of the other instances. In the field of pattern, recognition the term "anomaly detection" was introduced to detect an anomaly using the study of the characteristics of the instances and classify what is normal - major- characteristic from what is not [1].

Nowadays surveillance cameras capture almost all events during the day. It provides safety to roughly all the streets, city squares, hospitals, banks, and courts. Violence events threatens the safety of everybody and it requires to be detected instantaneously and with quite good precision in order to take reaction to these unwanted actions. However, with these huge video records and the lack of labor to monitor them. A great challenge raised in the Era of computer technology. Automated monitoring established numerous techniques to overcome the problem of recognizing anomaly actions such as bare hands fights and weapon abuse. In this paper, we propose a model of end-to-end deep net consists of three main stages.

- Frames extraction and preprocessing; we selectively extract frames from video clips and preprocess them to be prepared for the upcoming phases.
- Features extraction; apply CNN model on the extracted framed to detect the objects and their locations represented as spatial features. Then apply LSTM recurrent model to add a time domain to the objects' locations during the passage of time to form temporal features. Combine these two features to form Spatio-temporal features.
- Classification; we feed the features vector to a fully connected neural network that classify the videos to two classes; violence or non-violence action.

The rest of the paper is organized as follows. Section II explores a survey on the previous related work. Section III presents the proposed model with a detailed description. Section IV discusses the experiments and results. Finally, section V concludes the work.

II. RELATED WORK

There are a few researches published in surveillance anomaly detection, because of the unavailability of the datasets; the limitation of the computational power, and the disability of the deep learning techniques. But in the last few years due to the huge increase of the computational power especially the breaking technologies of the GPUs many universities and research laboratories introduced video datasets violence actions related.

Nguyen et al. [2] propose a model that is designed as a combination of a network for reconstruction and a model for image translation, which shares the same encoder. The former sub-network identifies the most significant structures that appear in video frames, and the latter attempts to connect movement models to those structures. The model achieved 86.9% accuracy on avenue dataset and 96.2% on ped2 dataset.

Vu et al. [3] presents a model of robust anomaly recognition utilizing multi-level portrayals of both intensity and movement information. The proposed multilevel detector shows a critical improvement in pixel-level Equal Error Rate, to be specific 11.35%, 12.32% and 4.31% improvement in UCSD Ped 1, UCSD Ped 2 and Avenue dataset individually.

Wu et al. [4] propose a Fast-Sparse Coding Network (FSCN) in the view of High-level Features. First a two-stream neural network to extract Spatial-Temporal Fusion Features (STFF) in hidden layers. With the output, a utilization of Fast-Sparse Coding Network to construct a dictionary. By leveraging the indicator to create inexact sparse coefficients, the FSCN produces sparse coefficients inside a forward pass, which is basic and computationally proficient. Compared with traditional sparse coding-based methods, FSCN is much more faster at the test stage.

Duman et al. [5] propose a system (OF-ConvAE-LSTM) to distinguish anomalies utilizing Convolutional Autoencoder and Convolutional Long Short-Term Memory in an unsupervised way. Other than the deep learning model, the feature extraction stage dependent on thick optical flow is applied in the structure to get the speed and heading data of closer view objects which accomplished accuracy of 89.5 % on avenue dataset ,92.4% on ped1 and 92.9 % on ped2.

Kang et al. [6] propose a neural network-based technique that joins the concept of area under curve (AUC) with the multiple-instance learning (MIL) approach. it formulates the multiple instance AUC (MIAUC) model that predicts high anomaly scores for irregular segments. Moreover, sparsity and temporal smoothness constraints are used in the loss function for better detection. it accomplishes the AUC of 84.4 on their dataset and it accomplishes 93.2 on ped1 dataset.

Yang et al. [7] propose an anomaly discovery approach by learning a generative model utilizing deep neural network system a weighted convolutional autoencoder-(AE-) long short term memory (LSTM) network is proposed to reconstruct data and perform anomaly location dependent on reconstruction errors to determine the current difficulties of irregularity recognition in complicated definitions and background impact. this methodology accomplished an accuracy of 85.7% on CUHK Avenue Dataset ,85.1% on UCSD Ped1 dataset and 92.6% on Ped2 Dataset.

Sultani et al.[8] propose a model to learn anomaly through the deep multiple instance ranking system that utilizes weakly labeled training videos i.e. the training labels (anomalous or normal) are at video-level rather than clip-level. In our approach, the concept about normal and anomalous videos as bags and video fragments as multiple instance learning (MIL), and naturally learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video fragments. This approach accomplished a accuracy of 74.44% on 13 real-world anomalies dataset.

Akcay et al. [9] present a novel encoder-decoder-encoder design model for general anomaly identification empowered by an adversarial training system. Experimentation across dataset benchmarks of shifting complexity, and inside the operational anomaly recognition setting of X-ray security screening, shows that the proposed strategy outperforms both state-of-the-art cutting edge GAN-based and conventional autoencoder-based abnormality identification approaches with speculation ability to any irregularity location task. They method was tested on two datasets: University Baggage Anomaly Dataset — (UBA) and Full Firearm versus Operational Benign — (FFOB) accomplished precision of 64,3% with UBA and 88,2% with FFOB.

Gao et al. [10] Experimental results show that utilizing consolidated highlights with AdaBoost +Linear-SVM accomplishes improved performance over the state-of-the-art on the Violent-Flows benchmark. This technique accomplished an accuracy of 92.81 % on datasets Hockey Fight database and 94.84 % on Violent-Flows dataset.

Das et al. [11] propose techniques for violence detection. It is classified into three categories: visual based methodology, which utilize SVM classifier. Sound based methodology, which utilize progressive methodology. It is dependent on Gaussian mixture models and Hidden Markov models (HMM). And hybrid method is utilizing k-Nearest Neighbor classifier to choose whether the given sequence is violent or not. Its accuracy is 88.19% on the KTH dataset.

Isupova et al. [13] Anomaly detection can be applied by using the probabilistic framework that based on topic modeling. In this framework, the data is considered abnormal if it has a low value of likelihood. The proposed framework uses Batch and online Gibbs samplers. This method achieved an accuracy of 72.80 % on QMUL datasets.

Ahmed et al. [13] Anomaly recognition is a significant data analysis task, which is helpful for identifying the network interruptions. This model presents an inside and out analysis of four major classes of anomaly detection procedures which incorporate classification, statistical, information theory and grouping. This algorithm accomplished a precision 57.81% by using k-means, 65.40% by using improve k-means and 80.15% by using Distance-base anomaly recognition on DARPA/KDD datasets.

Bilinski et al. [14] propose an expansion of the Improved Fisher Vectors (IFV) for videos, which permits to represent a video using both local features and their spatio-temporal positions. At that point, the famous sliding window approach is used for violence identification. It is used to re-figure the Improved Fisher Vectors and utilize the summed area table data structure, which accelerate this approach. They are utilized direct Support Vector Machines. This technique accomplished an accuracy 96.4 % on violent-flows dataset, 93.7 % on Hockey Fight dataset and 99.5% Movies dataset.

Xu et al. [15] propose Appearance and Motion DeepNet (AMDN), which uses deep neural systems to naturally learn feature representations. To abuse the corresponding data of both appearance and motion patterns, the addressed model present a novel double fusion framework, joining both the advantages of conventional early combination and late fusion strategies. In particular, stacked de-noising auto encoders are proposed to independently learn both appearance and motion features includes just as a joint representation (early fusion). In view of the learned representations, numerous one-class SVM models are utilized to anticipate the oddity scores of each input, which are then coordinated with a late fusion technique for definite anomaly detection. They evaluate the proposed method on two openly accessible video surveillance datasets.

Jiang [16] initially proposed a solution depends on an unsupervised learning approach. To start with, all the video occasions are represented by directions of moving objects. At that point they are clustered into a few behavior patterns under a probabilistic system. Those examples with low recurrence of event (not many directions uphold) are recognized as abnormal examples.

Piciarelli's [17] introduces application domains that are unique, running from video surveillance to automatic video annotation for sport videos or TV shots. The vast majority of the works in occasion analysis depend on two main approaches: the previous dependent on explicit occasion recognition, focused on finding high-level, semantic

translations of video sequences, and the last dependent on abnormality location. The application manages the subsequent approach, where the last objective is not the explicit labeling of perceived occasions, however the detection of anomalous occasions varying from regular occasions. The proposed approach is based on single-class support vector machine (SVM) clustering, where the novelty detection SVM capabilities are used for the identification of anomalous trajectories. Particular attention is given to trajectory classification in absence of a priori information on the distribution of outliers.

III. Proposed Model

The architecture of the model is presented in Fig. 1. The video is extracted into frames and the most important framed are picked; ensuring that the frames are informative and contain all the events of the video. Then the frames are preprocessed. The next two successive processes are feature extraction using both space and time dimensions to create Spatio-temporal features through a custom build a convolutional neural network and long short term memory LSTM recurrent neural network. The output from the last two processes is a features vector that is fed into a classifier. The last classification process is performed by building 5 layers fully connected neural network to classify the video into violence or non-violence action.

A. Preprocessing

This process consists of two sub-processes. The first is generating informative frames of the addressed video. The frames are extracted uniformly across the video clip timeline. We have to ensure that the extraction process caught an anomaly event, then we end up with 10 frames along with the video clip. Supposedly if the video is 100 frames the video frame generator task is to capture 1 frame and skip the next 9 frames so the index of the frames will be 01, 11, 21, ..,91. Fig. 2 illustrates the extraction process. The output of this sub-process is set of frames, the shape is (V,F) where V represents the number of videos in the dataset, F represents the number of frames extracted of each video.

Fig. 2. Illustration of picking the frames from a video clip

The second sub-process is the augmentation. The extracted frames are augmented. Each patch of the video frames is resized to 112× 112, randomly zoomed in the range of 10%, flipped horizontally, rotated in the range of 8 degrees, shifted width and height 20% of the total width and height. The shape of the output is (V, F, w, h, 3), where V and F are the number of videos and the number of frames. W and h represents the width and the height of each frame consecutively, and 3 for the color channel RGB of the frames. The augmented dataset is split to ratio of 80% for training and 20% for testing.

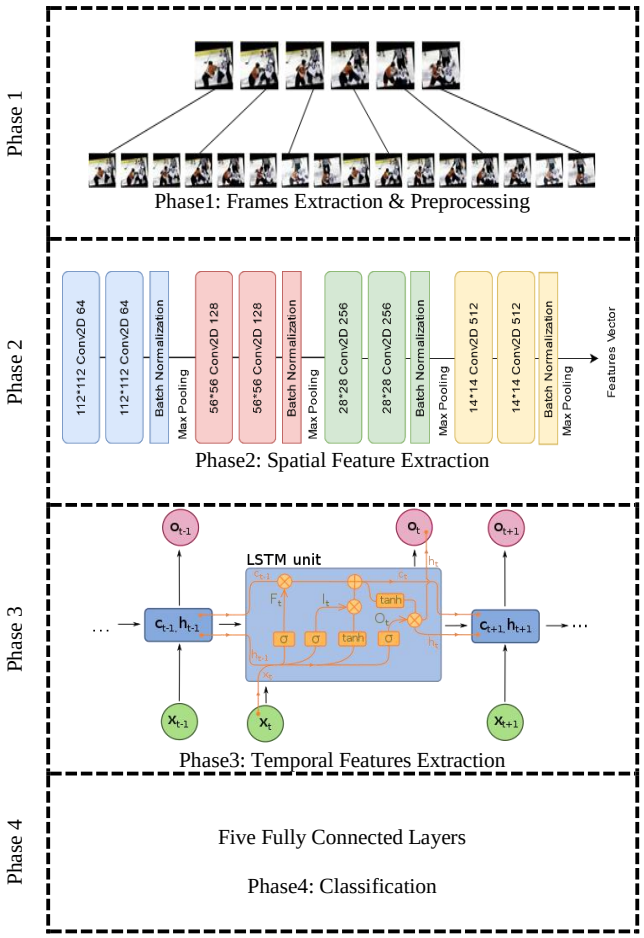

Fig. 1. The Proposed Model Architecture.

B. Feature extraction

The features of each frame are obtained by using a CNN model, which detect the crucial objects and their position in the frame. The frames are relatively connected, so we added a time domain by using LSTM recurrent model, which catch and store relevant features and forgets irrelevant ones.

Fig. 3 shows the introduced CNN model. It is constructed of four blocks of convolution layers every block contains two Conv2D layers, one Batch Normalization layer and one MaxPool2D layer, The total parameters used are 4,689,216 The output size after this step is (V, F, 512) where 512 is the number of features extracted from each frame.

LSTM model is used to extract the temporal features that keep track of the changes over time. The LSTM model consists of 128 units as the dimensionality of the output space and the tanh function as an activation function. In addition, the hard-sigmoid function is used as recurrent

activation, glorot uniform is used as a kernel initializer and the initial values of bias are set to zero.

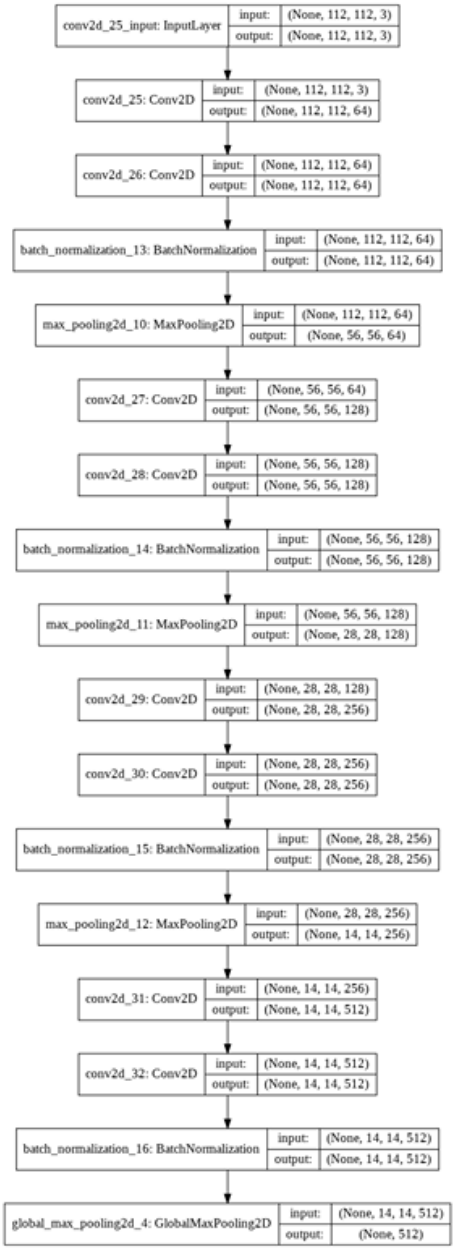

Fig. 3. The architecture of the CNN model

C. Classification

In Fig. 4 the block diagram of the classifier, which is the final phase for the proposed model. It classifies the video into non-violence or violence video. It is fed with the features vectors, which produced from the previous feature extraction phase. The classification phase consists of 4 fully connected hidden layers. The layers consist of 1024, 512, 128, 64 neurons respectively. Each layer has a ReLU activation function and between each layer, a drop-out layer with drop value 0.5. Finally, the last layer consists of 2 neurons and SoftMax function as an activation function.

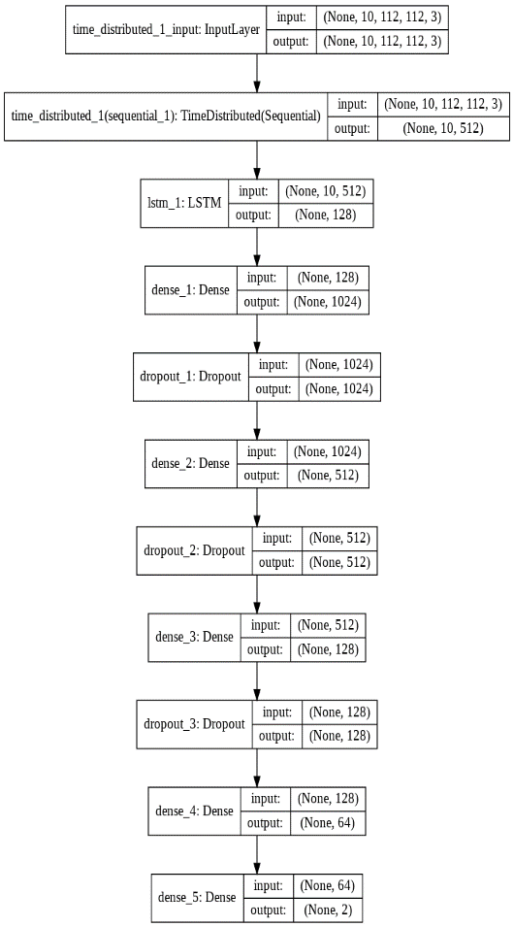

Fig. 4. The architecture of the classifier

IV. EXPERIMENTAL RESULTS

The proposed model is evaluated on two state-of-the-art benchmarks. One of them is hockey fights dataset [18]. The other is the Real-Life Violence Situations dataset [1].

A. Datasets

i) Real-Life Violence Situations dataset

RLVS is the main benchmark dataset. Soliman et al. [1] introduce it. The dataset consisted of 2000 trimmed videos of full length of 3 hours. The average length of each video is 4 seconds long, which is suitable informative length for violence. As the expected violence event takes no time to be initiated. The dataset has 2 classes a normal activities class labeled as non-violence which contains nearly all the human normal daily activities like exercising, playing sports, eating, etc., the other class is for the violence actions which were labeled violence. The violence class contained videos of bare hands fights, non-projectile weapon abuse fights that are sticks, knives, and big throw-able objects. Generally, the fights are near distance fights. The dataset was distributed equally on the 2 classes, 1000 video for each class.

ii) Hockey fight detection dataset

It is composed of 1000 videos divided into 500 violence videos and 500 nonviolence videos. It was taken from National Hockey League hockey games where each video consists of 50 frames and each frame has a size of 720×576. All videos have the same background, whereas players in ice hockey only appear in all videos.

Fig. 5. Samples of the RLVS dataset

Fig. 6. Samples of the hockey fights dataset

Figure 5, shows sample of vilence and non-violence scences in RLVS dataset. While Figure 6, shows sample of volence and non-violence scences in hocky fights dataset.

B. Environment Settings

The system is run on Colab [19] that have GPU NIVIDIA K80, 12 GB RAMs and 68 GB as hard disk.
The proposed model is implemented in python using Keras library [20] with backend TensorFlow [21]. And some libraries like OpenCV [22] and matplotlib [23].

C. Results and Discussion

In this section, the model performance is discussed. In the first experiment, the model is trained and tested for 50 epochs on the Real-Life Situations data set. The training accuracy was 99.8% while the validation accuracy was 92%.
The second experiment, the same model is trained and tested for 50 epochs on Hockey fights dataset. The training accuracy was 96.38% while the validation accuracy was 94.5%.

Figure 7, shows the training and validation accuracies for both RLVS and hockey fights datasets.

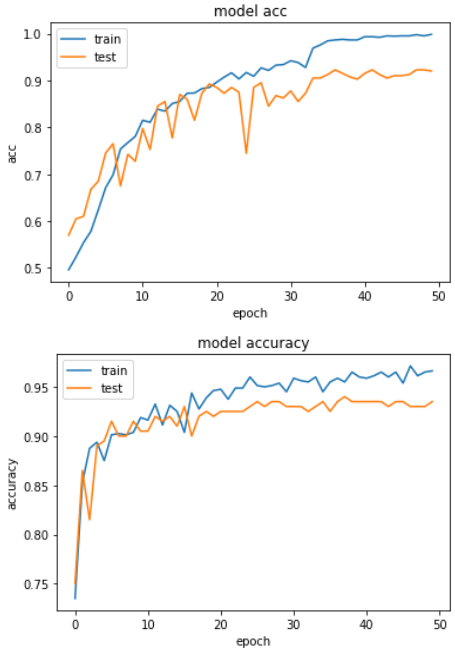

Fig. 7. The training and validation accuracies over 50 epochs (a) the RLVS, (b) hockey datasets

In Table 1, the proposed model accuracies are compared with three previous related work. Soliman et al. [1], Gao et al. [11], and Bilinski et al. [15].

Table 1: Performance Accuracy of the proposed mode in comparison with related work

Method	Hockey fight dataset	RLVS dataset
Soliman et al. [1]	95.1%	94.9%
Gao et al. [11]	92.81%	--
Bilinski et al. [15]	93.7%	--
Proposed model	94.5%	92%

Although many related research faced a problem with hockey fights dataset where the standard background of all the videos is white ice. Their models could not be generalized. The accuracy of their models on this specific dataset is considerably bad. However, our proposed model succeeded to generalize and achieve good accuracy.

V. CONCLUSION

In this paper, we proposed an end-to-end deep learning neural network for the violence action recognition and detection. The proposed model is consisted of four phases; preprocessing through distributed selective frames across the video clip. Spatial feature extraction using the convolutional neural network. Temporal feature extraction using LSTM that output the Spatio-temporal features from the frames. Finally, the classification phase that implement a fully connected neural network to classify the videos into violence or non-violence clip.

The model was evaluated on one of the widely used hockey fights datasets, and a recently introduced RLVS dataset. The model showed competitive performance in comparison with related work.

VI. ACKNOWLEDGEMENT

The authors are grateful to each of Mohamed A Gohar, Mahmoud A Abbas, Hassan M Kamel, Ali M Ali for their contributions in an early version of this work.

VII. REFERENCES

[1] Soliman, M. M., Kamal, M. H., Nashed, M. A. E. M., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019, December). Violence Recognition from Videos using Deep Learning Techniques. In 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 80-85). IEEE.

[2] Nguyen, T. N., & Meunier, J. (2019). Anomaly detection in video sequence with appearance-motion correspondence. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1273-1283).

[3] Vu, H., Nguyen, T. D., Le, T., Luo, W., & Phung, D. (2019, July). Robust anomaly detection in videos using multilevel representations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 5216-5223).

[4] Wu, P., Liu, J., Li, M., Sun, Y., & Shen, F. (2020). Fast Sparse Coding Networks for Anomaly Detection in Videos. Pattern Recognition, 107515.

[5] Duman, E., & Erdem, O. A. (2019). Anomaly detection in videos using optical flow and convolutional autoencoder. IEEE Access, 7, 183914-183923.

[6] Kang, L., Liu, S., Zhang, H., & Gong, D. (2020). Person anomaly detection-based videos surveillance system in urban integrated pipe gallery. Building Research & Information, 1-14.

[7] Yang, B., Cao, J., Ni, R., & Zou, L. (2018). Anomaly detection in moving crowds through spatiotemporal autoencoding and additional attention. Advances in Multimedia, 2018.

[8] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6479-6488).

[9] Akcay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2018, December). Ganomaly: Semi-supervised anomaly detection via adversarial training. In Asian conference on computer vision (pp. 622-637). Springer, Cham.

[10] Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using oriented violent flows. Image and vision computing, 48, 37-41.

[11] Das, S., Sarker, A., & Mahmud, T. (2019, December). Violence Detection from Videos using HOG Features. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE.

[12] Isupova, O., Kuzin, D., & Mihaylova, L. (2016). Anomaly detection in video with Bayesian nonparametrics. arXiv preprint arXiv:1606.08455.

[13] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.

[14] Bilinski, P., & Bremond, F. (2016, August). Human violence recognition and detection in surveillance videos. In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 30-36). IEEE.

[15] Xu, D., Ricci, E., Yan, Y., Song, J., & Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510. 01553.Anomalous Event Detection.

[16] Jiang, F. (2011). Anomalous event detection from surveillance video. Northwestern University.

[17] Piciarelli, C., Micheloni, C., & Foresti, G. L. (2008). Trajectory-based anomalous event detection. IEEE Transactions on Circuits and Systems for video Technology, 18(11), 1544-1554.

[18] Gracia, I. S., Suarez, O. D., Garcia, G. B., & Kim, T. K. (2015). Fast fight detection. PloS one, 10(4), e0120448.

[19] https://colab.research.google.com

[20] https://keras.io/

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

[22] https://opencv.org/

[23] https://matplotlib.org/