Violent Physical Behavior Detection using 3D Spatio-Temporal Convolutional Neural Networks

Xiuhong Xu[1], Zhongming Liao[2*], Zhaosheng Xu[3]
College of Photovoltaic Power Generation, Jiangxi New Energy Technology Vocational College, Xinyu 338004 Jiangxi, China[1]
Academic Affairs Office, Xinyu College, Xinyu 338004, Jiangxi, China[2]
School of Mathematics and Computer Science, Xinyu College, Xinyu 338004, Jiangxi, China[3]

***Abstract*—The use of surveillance cameras has made it possible to analyze a huge amount of data for automated surveillance. The use of security systems in schools, hotels, hospitals, and other security areas is required to identify the violent activities that can cause social, economic, and environmental damage. Detecting the mobile objects on each frame is a fundamental phase in the analysis of the video trail and the violence recognition. Therefore, a three-step approach is presented in this article. In our method, the separation of the frames containing the motion information and the detection of the violent behavior are applied at two levels of the network. First, the people in the video frames are identified by using a convolutional neural network. In the second step, a sequence of 16 frames containing the identified people is injected into the 3D CNN. Furthermore, we optimize the 3D CNN by using the visual inference and then a neural network optimization tool that transforms the pre-trained model into an average representation. Finally, this method uses the toolbox of OPENVINO to perform the optimization operations to increase the performance. To evaluate the accuracy of our algorithm, two datasets have been analyzed, which are: Violence in Movies and Hockey Fight. The results show that the final accuracy of this analysis is equal to 99.9% and 96% from each dataset.**

***Keywords*—Violence detection; surveillance cameras; 3D Convolutional Neural Network (3D CNN); Spatio-temporal convolution; deep learning; abnormal behavior**

I. INTRODUCTION

The perception of human behavior and the analysis of its activities has been faced with many challenges so far [1]; therefore, the processing of movies and also the perception of the movies' content with the proper accuracy and on a large scale has particular importance [2]. In this regard, due to the use of cameras in the city, we are faced with a very large amount of video and its content. It is impossible to analyze this amount of information by humans [3]. By analyzing the recorded frames from the surveillance videos, it is possible to recognize the abnormal behavior and the violent activity that has occurred, and it is possible to make effective decisions at the appropriate time and conditions [4]. In the prior years, with the development in the area of computer vision, a large number of new methods have emerged and have attracted much attention from researchers because of their wide-ranging security applications. In 2017, for example, 954,261 CCTV cameras were fixed at the generic level in South Korea, with an increase equal to 12.9% from the prior year [5]. The target of the installation of the cameras is to provide security in generic locations. To this end, we concentrate on violence detection with the use of cameras. Violence is an unusual activity that includes the bodily power to harm something or to murder or injure a person or a brute. These operations can be detected by an intelligent monitoring model that can be applied to ban these incidents before they become more deadly. The main application of the security systems, which are fixed in various locations such as schools, hotels, streets, and so on, is to comfort the availability of security guards by alarming them to violent activities. However, the human performance monitor on the surveillance film is too slow, which causes life loss and property. Therefore, there is a request for an automatic violence recognition model [6]. Hence, this area of study is constantly expanding, and different techniques have emerged in this field.

Violence detection has shown its application in the modern systems used by humans due to its wide and significant applications in the field of human security and comfort. By reviewing the existing articles in the field of violence detection, we can refer to the presented method in [7]. This method extracts the proper features by combining the spatiotemporal features and also acceleration features, each of which is obtained by using a two-dimensional convolutional network and then a recursive LSTM network. The acceleration changes are the important components in the detection of person-to-person violence. In this article, this acceleration has been calculated and has been modeled by using the severity of the ocular stream changes in three consecutive frames. To calculate spatial features, this article uses the VGG19 network [8] trained on Imagenet and then selects the features from the penultimate layer as the feature vectors. Another network that considers the changes in the optical flow as a feature is a Tdd network [9] which was trained and created by using the UCF101 dataset.

In another research presented in [10], the idea of TSN, that is, the temporal division network, was introduced. This method was actually a new framework for the detection of movie-based performance, and it was based on the concept of long-range domain structure modeling. Their method was the extraction of the sparse temporal feature, which involves surface video monitoring to enable effective learning by using violent and action-packed videos that sparsely sample the input frames. It can be said that this architecture is segmental. In research [11], the authors used 3D ConvNet and the key-frame to extract the features from clips that contain violent scenes. They have used 3D ConvNet for the short clips, and also, they have used the

key-frame for the longer clips. The key-frame method divides the movie based on the extracted key-frames, and then it examines the similarity between adjacent frames by changing the position of the gray center.

In research [12], the authors present a method for detecting violent robberies from CCTV footage by using a deep end-to-end sequence model. They have used VGG-16 and a prior-trained CNN by the input video frames (which extract features). They process the features trail with two long-term and short-term memories (convLSTM) using LSTM convolutional, which receives the trail of the obtained features. After these steps, in the end, they used several fully-connected layers for the prediction and classification. In this method, the types of firearms and cold weapons in the image are recognized; in this way, the robberies that show different levels of aggression can be classified.

According to the different detection methods that were examined, it was found that there is a gap in the correct extraction of the features; thus, each of the methods has a high computational complexity and a high cost, as well as they have network overhead and the loss of the movement's features. So, to dissolve the moot point of violence detection, the scene information of both levels is needed (namely, the structure of the scene and the movement made by the people present in the scene). Therefore, by examining different methods, a network has been designed, which is fully explained in the following sections. Our contributions include:

- By taking into account the limitations of the presented methods in this field which are presented in the next section, we present a 3D CNN model for learning the complex sequential patterns to accurately predict the violence from video frames.
- A major limitation in the existed methods is the processing of the un-important frames, which leads to the use of the more memory and the very time-consuming. By considering this limitation, we first identified the people in the video stream by using a pre-trained MobileNet CNN model. Only a trail of 16 frames containing the individuals was transferred to the 3D CNN model for the final prediction, which has helped to achieve the efficient processing.
- Next, inspired by the concept of the transfer learning, 3D-CNN was set up by using the standard datasets to detect the violence in the internal and external surveillance.
- After obtaining the trained deep learning model, it was optimized by using the toolbox of OPENVINO to speed up and improve its performance in the phase of the model deployment. By using this strategy, the trained model was transformed into an average representation based on the trained weights and topology.

The research rest is as follows: Section II discovers our method. The empirical evaluation is considered in Section III. The conclusions and suggestions are presented in Section IV.

II. Proposed Method

In this part, we explain our method. In this method, violent behavior is recognized with the use of the end-to-end deep learning scheme. The general procedure is as follows: The camera records the film sequence. These film frames are sent directly to a trained MobileNet CNN model. This work is done to identify the persons in the film frame. When a person is detected in the film, a sequence of 16 frames is formed, and then it is sent to a 3D CNN to exploit the spatiotemporal features. These extracted spatiotemporal characteristics are fed to the Softmax classifier to examine the extracted characteristic associated with an activity, and then it provides the predictions. When violence is detected in the video frame, an alarm is sent to the nearest police stand. The suggested method is displayed in Fig. 1. In each of the following subsections we illustrate one step of our method.

A. Pre-Processing

To detect the violent behavior, the first step is the step of recognition of the persons in the video frame. Therefore, the first phase in the pre-processing stage is to use the methods to identify the persons. We only process the parts of the video that contain the persons, avoiding the irrelevant frames, instead of processing the total film. The input video is injected into the MobileNet-SSD CNN system [13], and the persons in the video are identified. The presented model uses CNN to identify the persons because it limits the delay and the size. The MobileNet model processes the separable deep convolutions for object recognition. If the deep and point convolutions are numbered separately, then there will be 28 layers, and each layer will have a non-linear Bacth of the ReLu type. Of course, the fully-connected layer will not have this feature. The first convolution layer will consist of 2-strides with a filter form equal to 3×3×3×32 and the size of input equal to 224×224×3. The subsequent deep convolution has one stride, its filter form is equal to 3×3×32, and its input size is equal to 112×112×32. Mainly, the MobileNet model is applied for the classification tasks. Meanwhile, its SSD is applied to put the multi-box recognizer, and it performs a combination of object recognition. This version, for this purpose, is added at the terminal of the network that performs the feed-forward convolution. Also, it generates a constant-size team of marginal boxes to certify the object detection in the video by extracting its feature maps. The boundary box convolution filters are formed using a predicted category and a certain probability for each category. The category which has the highest probability represents the existing object. An example of the detection of the persons existing in a video frame by using the described model is presented in Fig. 2.

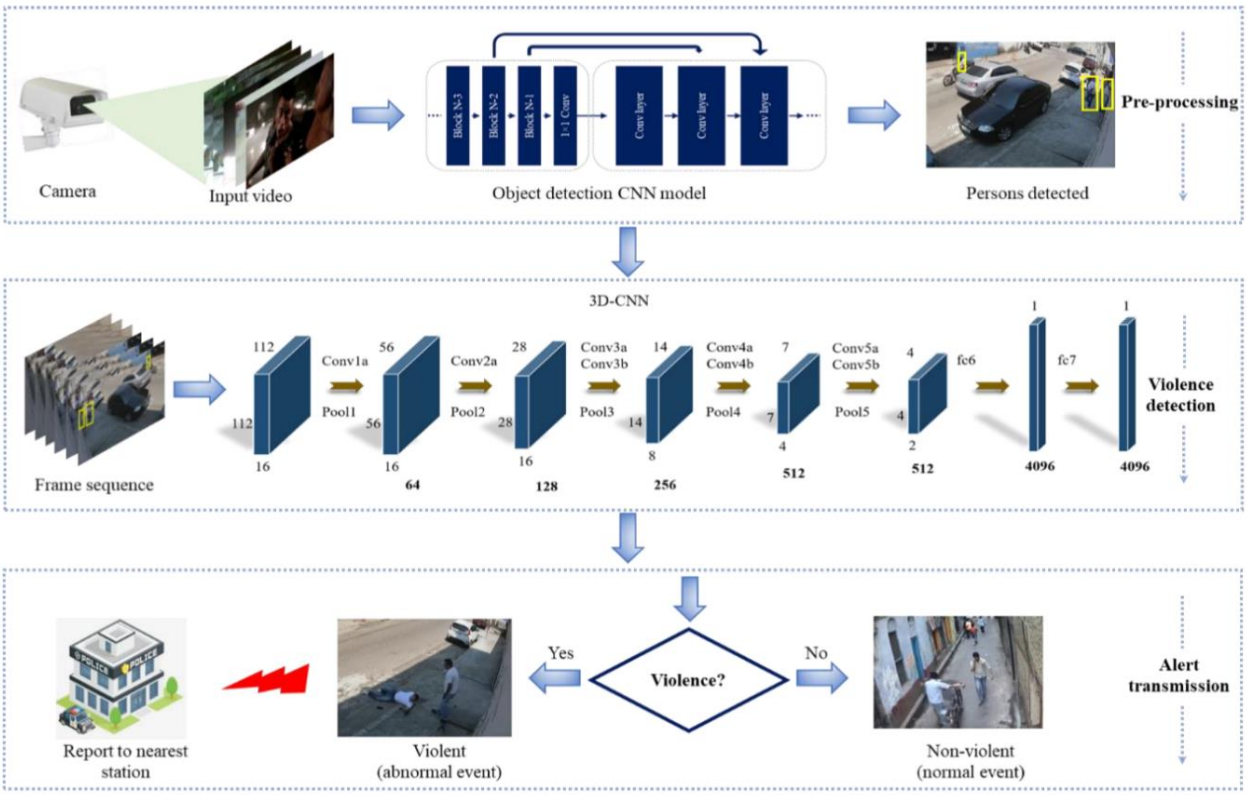

Fig. 1. The presented method for recognition of violent behavior in video frames.

Fig. 2. An instance of the detection of the persons existing in a video frame by using the MobileNet-SSD model.

B. Creation of The 3D CNN Model and Learning Phase

A 3D convolution model is implemented correctly to extract the appropriate spatiotemporal features, and it preserves better temporal information than the Pooling and the 3D convolution operations. There is only spatial information in 2D CNNs, while a 3D CNN can have total temporal information about the input video trail. Some existing methods use the two-dimensional ConvNets. This method is used to exploit the spatial relevance in the film (which simultaneously has temporal relevance). For example, in [14, 15], the 2D CNN model processes several frames. Also, it provides the relevance of total temporal features cumulatively. The 3D convolution model works by convolving a 3D mask in a designed cube (by using assembling the connected frames). In order to capture the motion information, the convolutionally produced feature maps are linked to multiple connected frames on the previous layer. Therefore, the obtained value at the x. y. z location in the map of the feature q on the layer p, which has a bias equal to t_{pq}, is defined by the following relationship:

$$N_{pq}^{xyz} = \tanh\left(t_{pq} + \sum_k \sum_{a=0}^{A_p-1} \sum_{b=0}^{B_p-1} \sum_{c=0}^{C_p-1} \omega_{pqk}^{abc} N_{(p-1)k}^{(x+a)(y+b)(z+c)}\right) \quad (1)$$

Where Cp is the size of the three-dimensional mask with a time dimension and ω_{pqk}^{abc} is equal to $(a.b.c)$th value of mask value connected to kth feature mapping on the previous layer. Only the 3D convolutional mask can extract one type of feature because the kernel weights are repeated throughout the cube. Fig. 3 shows the 3D CNN feature maps, which consist of two layers, $conv3a$, and $conv5a$. The presented sample input in Fig. 3 is obtained from the violence class in the dataset.

The average volume of the training data and test data is calculated before starting the training. The proposed network model is well-tuned to take these trails as input. In the Softmax layer, the final divination is calculated as it belongs to the violent category or the non-violent category.

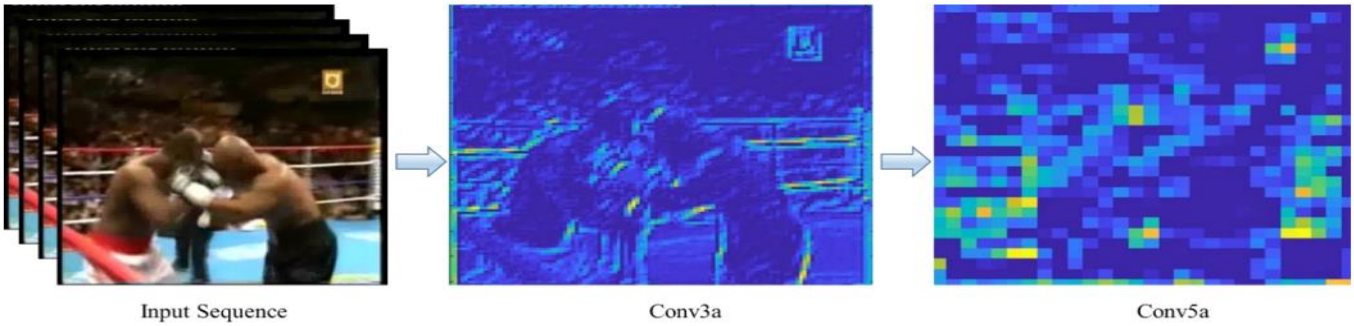

Fig. 3. The 3D CNN feature maps for two layers $conv3a$ and $conv5a$.

C. Preparation of Data

To detect the violence, we use a violence dataset that includes a certain number of the stunted film with various times. Each film dataset includes two categories: the violent category and the non-violent category. The entire dataset is separated into a trail of 16 frames with 8-frames overlay among every two consecutive films before starting the training. Then, after obtaining the frames, the total available data is divided into the training set and the test set. A considered ratio for the training set and the test set is equal to 80% and 20%, respectively. When the training set and the test set are taken, a list of files containing paths of the training $L_{tr} = \{S_1.S_{17} \ldots .S_{N\,}\}$ and the test $L_{te} = \{S_1.S_{17} \ldots .S_N\ \}$is generated.

D. Architecture of Proposed Three-Dimensional Convolution

By using the presented 3D CNN model in [16-19], we present our 3D CNN model. The proposed network consists of eight convolutions: five pooling layers, two fully-connected layers, and one Softmax layer. Each convolution layer has a 3×3×3 kernel by one stride, as well as all pooling layers have a maximum kernel size of 2×2×2. Of course, the first pooling layer has a kernel size of 2×2×1 with 2-strides. This exception is to preserve temporal-based data. In the first layer, the number of the considered filters for each convolution, second layer, and third layer, respectively, is 64, 128, and 256. Two fully-connected layers ($fc6$, $fc7$) contain 4096 neurons; also Softmax layer contains N outputs which depend on the number of classes on a dataset. In this paper and in the used dataset, the number of the outputs is equal to two owing to which we have two categories: the violent scenes and the non-violent. Detailed architecture is shown in Fig. 4.

The proposed architecture catches a trail of 16 frames for input. The input dimensions are equal to 128×171, but we have used the random cuts with a size equal to 3×16×112×112 to avoid the problem of overtraining and the problem of not achieving efficient learning. Then, the frames trail is followed by the convolution operations and the 3D pooling operations. The network works as a public feature extractor when the training is done. The various features are trained on the various layer of this network hierarchy. Finally, the exit class is doped as violent or non-violent.

E. Optimization of Proposed Model

The optimization of this model is a process that is applied to produce an optimal design model based on the prioritized limitations. Meanwhile, the power, efficiency, and reliability of this model are maximized. With these methods, we have used the open-source tool of OPENVINO which was created by Intel. This tool develops the workflow through hardware with the maximization of the hardware performance. It runs on the hardware of Intel, and also it takes the prior-trained modules like ONNX, Caffe, TensorFlow, and MXNet as the input, then it transforms them to IR with the use of the model optimizer. Simultaneously and accurately, the model optimizer is applied to make possible a transmission among the training and deployment layers to tune the defined model for optimal implementation in the final model. The stream and the trend of the platform optimization are brief: the model training, the model optimizer, the output (IR), as well as the final platform.

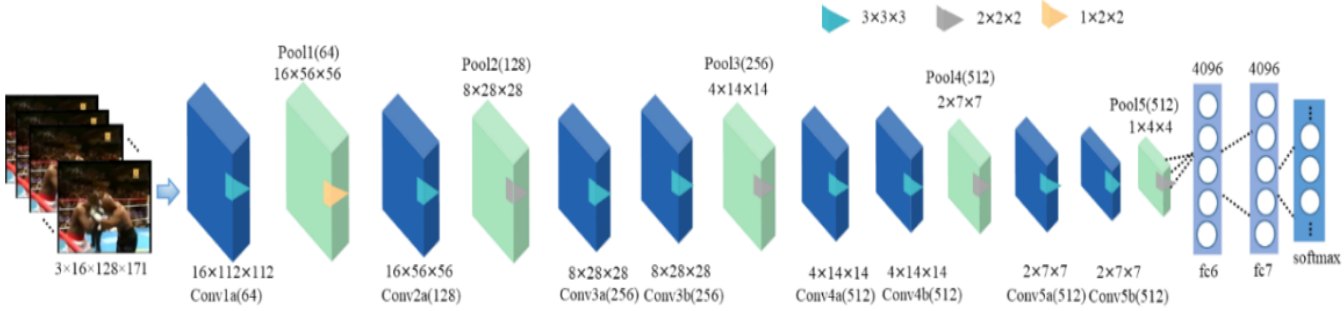

Fig. 4. The structure of the proposed three-dimensional convolution network.

III. Tests and Review of Results

In this part of the article, we present the implementation details of the proposed algorithm, and also we show the performed tests on the dataset and the obtained results. The Python programming language has been used for the implementation of these tests. The presented method is implemented on a computer with a 3.0 GHz Intel(R) Core(TM) i7 CPU and 8G RAM. The convolutional neural network is implemented in GPU, and the graphics card used in this method is NVIDIA GEFORCE 840M.

A. Used Dataset

In this part, the dataset used for the evaluation of our method is explained. In the field of detection of violent behavior, there are different datasets, but in this article, we use two datasets of, Violence in Movies [5] and Hockey Fight [5], which are used in most of the articles in this field.

The dataset of Violence in Movies was nominated by Nievas et al. [5]. This dataset contains 200 videos which include person-to-person combat videos. These scenes are taken from the action movies. Also, in this dataset, the non-combat films are exploited from the datasets of the available act recognition. This dataset discovers the various locations with a mean resolution equal to 360×250 pixels, and each film is restricted to 50 frames. The first person in this dataset, which in the sequence is placed, has little or no camera movement. Similar to the previous dataset, the dataset of Hockey Fight was nominated by Nievas et al. [5] and consisted of 1000 stunted films obtained from the National Hockey League (NHL). Five hundred clips in this dataset are tagged as violent, as well as 500 clips are tagged as non-violent. Each film contains 50 frames which have a resolution equal to 360×288 pixels. The examples of the frames in these two datasets are mentioned in Fig. 5. The first row is related to the dataset of Violence in Movies, and the second row is related to the dataset of Hockey Fight.

B. Evaluation Results of Proposed Method

In this part, details of the obtained results from the performed experiments on the introduced datasets are presented. Table I displays the results of the performed tests in the dataset of Violence in Movies. As it is known, the highest obtained accuracy is equal to 99.9%, which has a loss equal to 1.67×10^{-7}. This result is obtained at the maximum iterations equal to 5000 with a base learning rate equal to 1×10^{-5}. The important point is that violence detection is easier in the dataset of Violence in Movies than violence detection in the dataset of Hockey Fight. The reason is that there are more people in the clips.

Also, Table II displays the results details of the performed tests on the dataset of Hockey Fight where the highest obtained accuracy is equal to 96% with a loss equal to 5.77×10^{-4}. This result is obtained at the maximum iterations of 5000 as well as the learning rate equal to 0.0001.

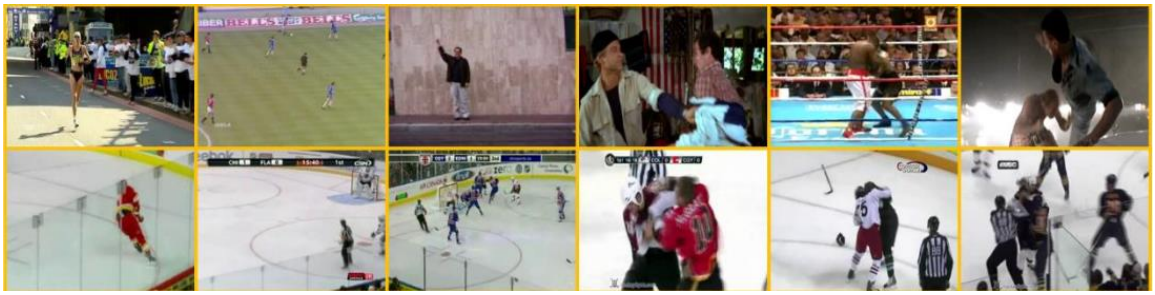

Fig. 5. Examples of frames in a used dataset.

TABLE I. The Results of our Method on the Dataset of Violence in Movies

Learning Rate	Number of Iterations	Loss	Accuracy
0.001	1000	0	99.4%
	3000	0	
	5000	1.21×10^{-2}	
1×10^{-5}	1000	1.99×10^{-3}	99.9%
	3000	5.4×10^{-4}	
	5000	1.67×10^{-7}	

TABLE II. Results of the Proposed Method on the Dataset of Hockey Fight

Learning Rate	Number of Iterations	Loss	Accuracy
0.001	1000	1.49×10^{-2}	94.9%
	3000	0	
	5000	1.85×10^{-2}	
0.0001	1000	1.79×10^{-3}	96%
	3000	2.27×10^{-3}	
	5000	5.77×10^{-4}	

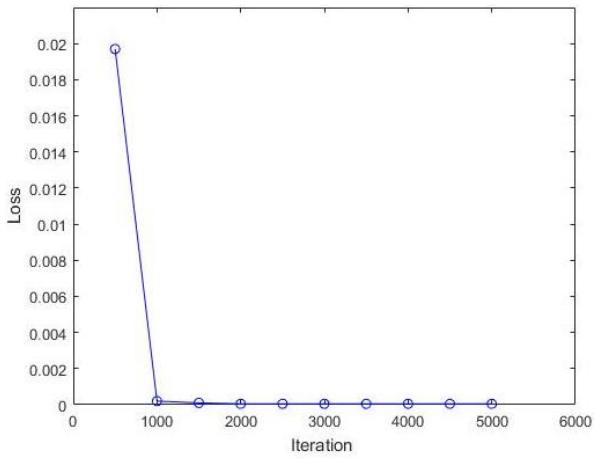

Fig. 6. The trend of the obtained loss value in the dataset of Hockey Fight.

In the experiments, it was found that a learning rate has a significant trace in the amount of the loss. In Fig. 6, the process of changing the loss value is specified. This trend is demonstrated with the change in several learning rates and the iterations equal to 0.001 in the dataset of Hockey Fight. In the 500th iteration of this process, the gained loss value is equal to 1.97×10^{-2}. This value reduces at the same time as the number of iterations. In maximum iteration equal to 5000, this value is equal to 2.32×10^{-7} while the test conditions have not changed, and only the learning rate has changed to 0.0001.

The value of the loss in the dataset of Violence in Movies is very high in the early stages. Then, with increasing in the iteration, the value of the loss is decreased. In this way, the value of the obtained loss in the iteration of 5000 is equal to 5.4×10^{-4}. The trend of the loss reduction in the dataset of Violence in Movies is shown in Fig. 7.

Also, we have appraised the performance of our method by checking the precision, recall, and comparison between datasets with the presentation of AUC in Table III. This table shows the performance of our platform for two datasets. The relationship between the precision and recall is as follows:

$$\text{Precision} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Positive\ (FP)} \quad (2)$$

$$\text{Recall} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)+False\ Negative\ (FN)} \quad (3)$$

Moreover, the taken confusion matrix is displayed in Table IV. The values of recall as well as precision values for two datasets, respectively, are between $Xmin$, $Ymin$ and $Xmax$, $Ymax$ where X displays precision and Y displays recall for two datasets.

C. Comparison of Presented Method with Similar Methods

One of the important parts of the presentation of a new network is the provision of a complete report of the efficiency of the designed network and the correctness and accuracy of the network performance in different conditions. For this purpose, to demonstrate the better performance and accuracy of the designed network in comparison to similarly designed networks that have been presented and implemented by other researchers, the obtained results from the performance of this network have been compared with other networks. The similarity of the test conditions is related to the same dataset and the same quality evaluation parameters. Therefore, in this part, we contrast the results of two datasets with existing methods. A comparative evaluation with the existing platforms is displayed in Table V. We present the results of the presented method in [20] in the first row. It uses Oriented Violent Flows for motion enlargement. Also, it uses AdaBoost as a feature. Another comparison method that uses Hough forests with two-dimensional CNN for violence detection is presented in [21], and its results are listed in the second row. In addition, we have contrasted the results of our method with the proposed method in [22] that uses motion bubbles and random forests for rapid violence detection. Its results are presented in the third row. Also, in [23], two descriptors are applied in order to identify unusual activities. They applied a simple histogram of the directional tracks together with a dense optical stream in order to detect the abnormal behavior in the terminal result. The fourth row shows the results related to this method. In [24], the writers applied a sliding window method. Also, they use the method of the improved Fisher's vector for violence detection. The results of this method are also displayed in the fifth row.

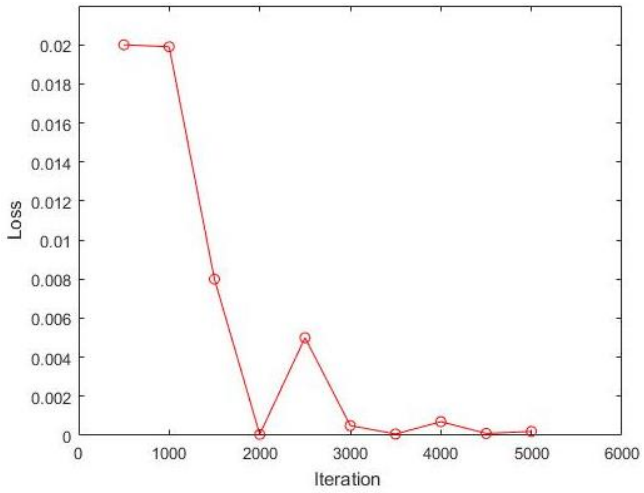

Fig. 7. The trend of the obtained loss value in the dataset of Violence in Movies.

TABLE III. PRECISION, RECALL AND AUC FOR OUR METHOD IN TWO DATASETS

Dataset	Values				Precision	Recall	AUC
	TP	TN	FP	FN			
Hockey Fight	262	230	11	9	0.9957	0.9667	0.97
Violence in Movies	50	57	0	0	1.0	1.0	0.997

TABLE IV. CONFUSION MATRIX

The classes in the dataset	1Hockey Fight		1Violence in Movies	
	Violent	Non-Violent	Violent	Non-Violent
Violent	262.0	11.0	50.0	0
Non-Violent	9.0	230.0	0.0	57

TABLE V. RESULTS OF COMPARISON OF OUR METHOD WITH THE SIMILAR METHODS

Methods	Achieved Accuracies (%)	
	Violence in Movies	Hockey Fight
The method presented in [21]	-	87.5
The method presented in [22]	99	-
The method presented in [23]	96.9	-
The method presented in [24]	98.5	83.1
The method presented in [25]	99.5	93.7
Our proposed method	99.9	96

IV. CONCLUSIONS AND SUGGESTIONS

In this article, the main goal is to identify the violent behaviors in the video frames. Therefore, a three-step method for detecting the violence in the video frames is presented. In this method, first, the people in the video frames are identified. The identification of the people is done by using CNN, which makes the undesirable frames to be ignored and the overhead of the proposed system is reduced. In the second step, a sequence of the frames containing the detected individuals is injected into a trained 3D CNN model, which is performed on two standard datasets. In this dataset, the spatio-temporal features are used and sent to Softmax for the final predictions. Finally, this paper uses the toolbox of OPENVINO to perform the optimization operations to increase the performance. The results of conducted experiments on different datasets show the excellent performance of our proposed platform. These results show that our method is the best suited for detecting the violence in the video surveillance and has better accuracy than several other techniques. For future research, it is suggested that the researchers can ensure that our proposed system can be implemented on the resource-constrained devices. Also, the presented dataset is limited to few violent scenes that can be tried to complete this dataset. On the other hand, some violence is verbal and it is possible to prepare a dataset with this feature in the future works and evaluate the proposed method. In addition, the researchers can propose the edge intelligence to detect the violence in the Internet of Things by using the smart devices for the quick responses.

REFERENCES

[1] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," 2015, pp. 896-904.

[2] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," Expert Systems with Applications, vol. 91, pp. 480-491, 2018.

[3] M. S. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies, "Robot-centric activity prediction from first-person videos: What will they do to me?," 2015, pp. 295-302.

[4] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo, "Robot-centric activity recognition from first-person rgb-d videos," 2015: IEEE, pp. 357-364.

[5] H. H. Park, G. S. Oh, and S. Y. Paek, "Measuring the crime displacement and diffusion of benefit effects of open-street CCTV in South Korea," International Journal of Law, Crime and Justice, vol. 40, no. 3, pp. 179-191, 2012.

[6] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," 2011: Springer, pp. 332-339.

[7] W. Yu, K. Yang, Y. Bai, T. Xiao, H. Yao, and Y. Rui, "Visualizing and comparing AlexNet and VGG using deconvolutional layers," 2016.

[8] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3d deep convolutional descriptors," 2016, vol. 1, p. 3.

[9] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," 2019, pp. 1227-1236.

[10] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," 2016: Springer, pp. 20-36.

[11] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," 2017, pp. 5783-5792.

[12] M. Rezaee, Y. Zhang, R. Mishra, F. Tong, and H. Tong, "Using a vgg-16 network for individual tree species detection with an object-based approach," 2018: IEEE, pp. 1-7.

[13] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Advances in neural information processing systems, vol. 27, 2014.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," 2014, pp. 1725-1732.

[16] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," 2016, pp. 1049-1058.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015, pp. 4489-4497.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Deep end2end voxel2voxel prediction," 2016, pp. 17-24.

[19] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," Pattern Recognition Letters, vol. 130, pp. 370-375, 2020.

[20] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," Image and vision computing, vol. 48, pp. 37-41, 2016.

[21] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using hough forests and 2D convolutional neural

network," IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 4787-4797, 2018.

[22] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, "Fast fight detection," PloS one, vol. 10, no. 4, p. e0120448, 2015.

[23] H. Rabiee, H. Mousavi, M. Nabi, and M. Ravanbakhsh, "Detection and localization of crowd behavior using a novel tracklet-based model," International Journal of Machine Learning and Cybernetics, vol. 9, pp. 1999-2010, 2018.

[24] P. Bilinski and F. Bremond, "Human violence recognition and detection in surveillance videos," 2016: IEEE, pp. 30-36.

[25] Bilinski, P.; Bremond, F. Human violence recognition and detection in surveillance videos. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; pp. 30–36.