Vi-Net: A Deep Violent Flow Network for Violence Detection in Video Sequences

Tahereh Zarrat Ehsan
University of Guilan
School of Electrical Engineering
Rasht, Iran
Tahere.zarrat@msc.guilan.ac.ir

Seyed Mehdi Mohtavipour
Iran University of Science and Technology
School of Electrical Engineering
Tehran, Iran
mehdi_mohtavipour@elec.iust.ac.ir

***Abstract*— Video surveillance cameras are widely used due to security concerns. Analyzing these large amounts of videos by a human operator is a difficult and time-consuming job. To overcome this problem, automatic violence detection in video sequences has become an active research area of computer vision in recent years. Early methods focused on hand-engineering approaches to construct hand-crafted features, but they are not discriminative enough for complex actions like violence. To extract complex behavioral features automatically, it is required to apply deep networks. In this paper, we proposed a novel Vi-Net architecture based on the deep Convolutional Neural Network (CNN) to detect actions with abnormal velocity. Motion patterns of targets in the video are estimated by optical flow vectors to train the Vi-Net network. As violent behavior comprises fast movements, these vectors are useful for the extraction of distinctive features. We performed several experiments on Hockey, Crowd, and Movies datasets and results showed that the proposed architecture achieved higher accuracy in comparison with the state-of-the-art methods.**

Keywords—deep learning, computer vision, convolutional neural network, action detection, violence detection

I. INTRODUCTION

Violence detection in video surveillance systems is motivated by the increasing concern for people's safety. In recent years many research has been conducted in this field [1-4], but the problem is still open. Proposed methods for violence detection can be grouped into hand-crafted and deep learning methods. Hand-crafted features are extracted directly from the frames using various analyses such as optical flow, acceleration, appearance, human pose [5-6]. These features are computed manually by human engineers and contain discriminative information for violence detection. Violent Flows (ViF) [7] method computes a descriptor based on the target speed in consecutive frames. In violent action, speed is faster than normal and by analyzing their optical flow magnitudes over time, violent action is detected. Also, another method called the Histogram of Optical Flow Orientation and Magnitude and Entropy (HOFME) [8] used optical flow distribution as a violent feature. In this method, video is split into a non-overlapping spatio-temporal cube and for each cube, orientation and magnitude distribution are computed. In Differential Histogram of Optical Flow (DHOF) [9] descriptor, a volume of the targets bounding box in consecutive frames is computed. For violence detection. Acceleration is considered as another metric for violence detection and it can demonstrate many shapes of movement of the human body [3].

Deep Neural Networks (DNN) are machine learning techniques inspired by learning of the human brain and have been used widely in many research fields such as graph processing [10], network communication [11], and intelligent transportation [12]. A DNN consists of one input layer, multiple hidden layers, and one output layer. This network is made up of several units called neurons. Each neuron receives input from the preceding layers and feeds its output to neurons in the next layer. These hidden layers are built upon each other hierarchically. The deeper the number of layers, the more complex features are extracted. These features are not handcrafted and are obtained from data using a learning procedure. These features may not have real-world interpretation but they are useful for classification. A deep network for anomaly detection is proposed in [13]. Motion features are extracted from hidden layers of the network and by using an SVM classifier anomaly is revealed. In [14] a pre-trained MobileNet network is used for violence detection. This architecture is based on the 3D CNN layers and construction of bounding box volumes for targets in consecutive frames. Authors in [15] utilized a pre-trained CNN for feature extraction and SVM for feature classification. Representative Image (RI) is another approach proposed in [16] to combine target shape in several consecutive frames. Each RI image contains cumulative information of appearance and motion and considered as input in deep CNN architecture. In [17], key frames which have more non-zero pixel intensity are selected as an input of a MobileNet network with ImageNet and fine-tuned violence datasets. To apply longer-term temporal dynamics in detection architecture, two streams 2D CNN with Long Short-Term Memory (LSTM) is proposed in [18]. Also, a combination of 3D CNN and LSTM is proposed in [19] but the computational complexity of this network is reported high. There are many challenges in violence detection such as low resolution of frames in surveillance systems, changes in camera viewpoint that create different shapes for each action, and the way of fighting. In order to deal with these challenges, CNN-based architectures are a promising approach. The main advantage of this kind of deep network is high-level discriminative feature extraction.

In this paper, we focused on the Vif descriptor which is based on the optical flow vectors. In violent action, Movements of body, hand, and leg are fast. But in normal action, people move slower and there is no sudden changes in the motion. Optical flow vectors demonstrate the statues of pixels in consecutive frames which make them appropriate for the estimation of motion patterns. Feature extraction based on the

ViF descriptor with handcrafted methods didn't provide acceptable results, and we intended to use this descriptor together with CNN architecture to improve the violent behavior accuracy.

The rest of this paper is organized as follows: In section II, the ViF descriptor has been explained. In section III, Vi-Net architecture with designed parameters has been introduced. Section IV shows the experimental results on real-world datasets. Finally, section V is the conclusion of this paper.

II. ViF Descriptor

To obtain ViF descriptor, frames are split into non-overlapping spatial cells. For each pixel of the cell, the magnitude of optical flow is calculated. The distribution of magnitude in each cell shows is related to pixel displacement in consecutive frames. For optical flow estimation, dense *farneback* method is used [20] and it consist of several basic steps. At first, it uses a polynomial expansion assumption for motion estimation. The neighborhood of each pixel is approximated by a polynomial sentence as follows:

$$f_1(X) = X^T A_1 X + {b_1}^T X + c_1 \quad (1)$$

X is two-dimensional 1×2 vector containing coordinates of x, y, A_1 is a symmetric 2×2 matrix, b_1 is a 2×1 vector, c is a scalar and f is the neighborhood area for each pixel. It is assumed that this neighborhood is shifted by $d = (u, v)$ which is the optical flow vector:

$$f_2(X) = f_1(X - d) = (X - d)^T A_1 (X - d) + b_1^T (X - d) + c_1 \quad (2)$$

$$f_2(X) = (X)^T A_1 (X) + (b_1 - 2A_1 d)^T (X) + d^T A d - b_1^T d + c_1 \quad (3)$$

The current frame polynomial expansion is given by equation (4).

$$f_2(X) = (X)^T A_2 (X) + b_2^T (X) + c_2 \quad (4)$$

The polynomial coefficients of equation 2 and 3 are equal according to the assumption that pixel brightness remains constant between two consecutive frames.

$$A_1 = A_2 \ , \ b_2 = b_1 - 2A_1 d \ , \quad c_2 = d^T A d - b_1^T d + c_1 \quad (4)$$

By solving the above equation, the optical flow vector is calculated:

$$d = \left(\sum w \left(\frac{A_1 + A_2}{2} \right)^T \left(\frac{A_1 + A_2}{2} \right) \right)^{-1} \sum w \left(\frac{A_1 + A_2}{2} \right)^T \left(\frac{b_2 - b_1}{2} \right) \quad (5)$$

w is weight function of the neighborhood points. For each pixel magnitude of optical flow is calculated as follows:

$$M = \sqrt{u^2 + v^2} \quad (6)$$

u and v are optical flow vectors in the direction of x- and y-axis, respectively. Finally, ViF descriptor is the subtraction of optical flow vectors and is given by the following equation:

$$b(x, y, t) = |M(x, y, t) - M(x, y, t - 1)| \quad (7)$$

t represents the number of video frame. A basic idea is comparing the value of equation (7) with a pre-defined threshold in two consecutive frames to detect violent behavior. Although it may result in good accuracy, this cannot be a global and comprehensive approach for violence detection. Therefore, the combination of ViF descriptor and CNN architecture is the key to the interpretation of complex and abnormal behaviors.

III. Proposed Vi-Net Architecture

Convolutional Neural Networks (CNN) is a deep learning method which widely used for image and video classification tasks and discussed in previous sections. CNN is composed of two stages: feature extraction and classification. A combination of convolutional and pooling layers act as a feature extraction function and fully connected layers act as a classification function. In the hidden layer, there are many convolutional, pooling, and fully connected layers. The core building block of each CNN network is the convolutional layer. In this layer, the input image convolves with a set of learnable filters known as kernels to produce a new compressed image called a feature map. The results of each convolutional layer pass through one activation function to add non-linearity in the procedure of the training phase. Feature maps describe the features of the input image. To compress and decrease the spatial size of the results in each layer, there is one pooling layer at the end of each convolutional layer. Reducing the parameters in the network plays a key role in decreasing the computational cost for both feature extraction and classification. There are several functions for building a pooling layer such as average pooling, max pooling. The most common function in the pooling layer is max-pooling. In the max pooling layer, a kernel will be applied to a region in the image and then, the maximum value within it will be selected. Therefore, in the next layer, this value represents the compressed region. After processing the input image by several convolutional and pooling layers, this data should be merged together to construct new interpretable data which finally show the probability of each class. For this purpose, fully connected layers have been used. This kind of layer takes the final feature map of each filter and by flattening the data, computes the output score for each class. In a fully connected layer, each neuron in the previous layer is connected to all neurons of the next layer. As violence detection comprises two violent and normal behavior classes, there are two output neurons in the output of the Vi-Net network.

The architecture of Vi-Net is shown in Figure 1. As can be seen, optical flow is extracted for each frame. By subtracting optical flow vectors for two consecutive frames, the motion pattern is computed. This motion pattern is fed as input to the CNN network. In this Figure, green cuboids demonstrate convolution operations and the ReLU activation function. After each green cuboid, max pooling is applied to the output of the convolutional layer. After the last pooling layer, the matrix is flattened to a vector. Two fully connected layers are applied to the final result and are shown with an orange vector in this Figure. The final layer gives the probability of each class by using a softmax function. The parameters of Vi-Net architecture have been shown in Table1.

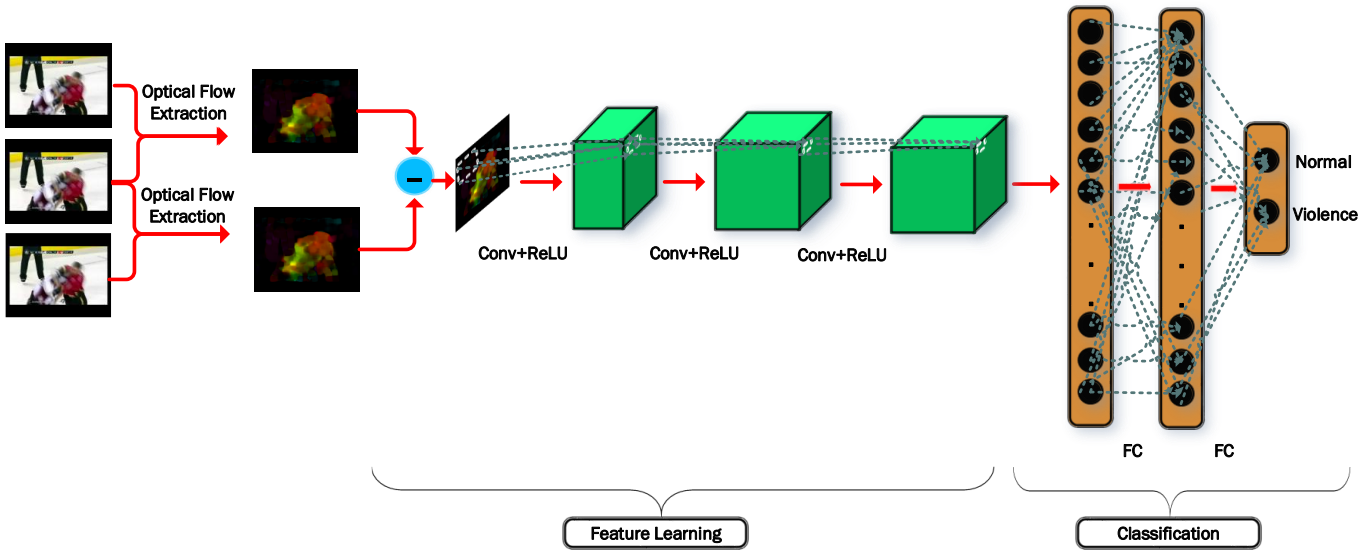


Fig. 1. Proposed Vi-net architecture for violence detection in videos

Table. 1. Selected hyper parameters of the proposed Vi-Net

Hyper parameters	Input Layer	Hidden Layer 1		Hidden Layer 2		Hidden Layer 3		Hidden Layer 4	Output Layer
		Conv	Max Pooling	Conv	Max Pooling	Conv	Max Pooling	Fully Connected	
Filters	-	32	-	64	-	64	-	-	-
Kernel Size	-	5×5	2×2	5×5	2×2	3×3	2×2	-	-
Stride Size	-	1	2	1	2	1	2	-	-
Neurons +Dropout Rate	-	-	-	-	-	-	-	64 + 0.5	2 *class*
Map Size	288×360	284×356	142×178	138×174	69×87	67×85	33×42	-	
Activation Function	-	*ReLU*	-	*ReLU*	-	*ReLU*	-	*ReLU*	*SoftMax*

In Table 1, selected hyper-parameters for each layer have been given. Layers 1 to 3 consist of a convolutional and pooling layer. The output of each convolution layer is fed to a ReLU activation function. ReLU passes values which are greater than zero and otherwise output is zero. Feature map of the convolutional layer is calculated as follows:

$$H_n^{l+1} = ReLU(w_n^l * H^l), ReLU(x) = \max(0, x) \quad (8)$$

H_n^{l+1} is feature maps for $l+1$th layer, w_n^l is a set of filters and $ReLU$ is the activation function. The main purpose of the convolutional layer in our Vi-Net is to detect features which are related to violent action. Starting layers extract low-level features whereas the output of final layers including complex features which are not visually interpretable. In Figure 2, the feature maps extracted from the first, second, and third convolutional layers are shown for normal and violent data. As can be seen, violence feature maps are clearly different from normal data. After each convolutional layer, a max pooling layer is used. It takes feature maps as input and down-sample it. It was shown in Table 1 that the size of feature maps reduced after each pooling layer. After fully connected layer, there is one dropout layer. It determines the elimination rate such that some neurons are randomly dropped from the network during training. The dropout layer prevents the over-fitting in training and improves the generalization. The activation function of the final layer is softmax which computes the probability of each normal and violent class. To train the weights of the network, a loss function is defined. The loss function is the difference between true and predicted labels and should be as minimized as possible. A gradient descent optimization algorithm has been used to update the weights and it is based on the calculation of derivatives of the loss function with respect to weights of each layer.

IV. EXPERIMENTAL RESULT

To evaluate the proposed Vi-Net architecture, experiments have been carried out on Hockey [6], Crowd [7], and Movies [6] dataset. The Tensorflow library is used to perform the experiments on a computer with GeForce-840M GPU. The Hockey dataset contains 1000 video clips of Hockey games. In this dataset, 500 clips are related to violent actions and 500 clips are related to the normal actions. Each clip has 50 frames with a resolution of 288×360 pixels. In the Movie dataset, there are 200 clips of actions in a wide range of categories. Half of the clips are related to violent actions and the other half are related to normal actions. Each clip has 50 frames with a resolution of 250×360 pixels. Finally, the Crowd dataset contains 246 videos taken online from YouTube. 123 clips are

related to violent actions and 123 clips are related to normal actions. The resolution of clips is 240×320 and the length of videos is variable from 50 to 150 frames. For evaluation, The accuracy and loss criterion is considered. Loss is the difference between the model predictions and true labels. Accuracy is model performance in classifying each class and is calculated as follows:

$$ACC = \frac{TP + TN}{total} * 100 \qquad (9)$$

TP is True Positive and corresponds to the number of violent samples that are correctly classified. TN is True Negative and corresponds to the number of normal samples that are correctly classified. $total$ is the number of all samples. For each dataset, accuracy and loss curves have been obtained for training and validation data. These curves have been shown in Figure 3. As can be seen, the proposed method learned all datasets well and achieved high accuracy for training and validation data. Also, the proposed method is compared with other state-of-the-art methods and the results are shown in Table 2. The accuracy of the ViF method is around 82% in the Crowd dataset and 83% in the Hockey dataset. The result in the Movies dataset is not reported. OViF is another hand-crafted method that added orientation to the ViF method. This method obtained 78% accuracy in the Crowd dataset and 87% accuracy in the Hockey dataset. This shows that these methods are not able to detect complex violent behavior. STIP method is another handcrafted method that achieved 89.5% accuracy in Movie dataset. Comparing handcrafted methods, fast violence detection which used the acceleration metric obtained higher accuracy and it is 98.9% for Movie dataset and 90.1% for the Hockey dataset. The accuracy situation in deep learning methods is better than handcrafted methods. HFCNN used a Hough Forest for feature extraction with a CNN network for data classification. This method obtained 99% accuracy on the Movies dataset and 94.6% accuracy on the Hockey dataset. Since HFCNN used RI images as input for feature extraction, it is not possible to use it in crowded environments. The KeyF method used key frames of video to extract high informative data and utilized it as an input of CNN This method has 99.5% accuracy in the Movies dataset and 87% accuracy in the Hockey dataset. The proposed Vi-Net architecture obtained 96% accuracy in the Crowd dataset, 98% accuracy in the Hockey dataset, and 99% accuracy in the Movies dataset. This method outperformed previous works, as considered optical flow vectors for the input of the CNN network. This input included valuable information of motion patterns. Moreover, it can be used in environments with any condition.

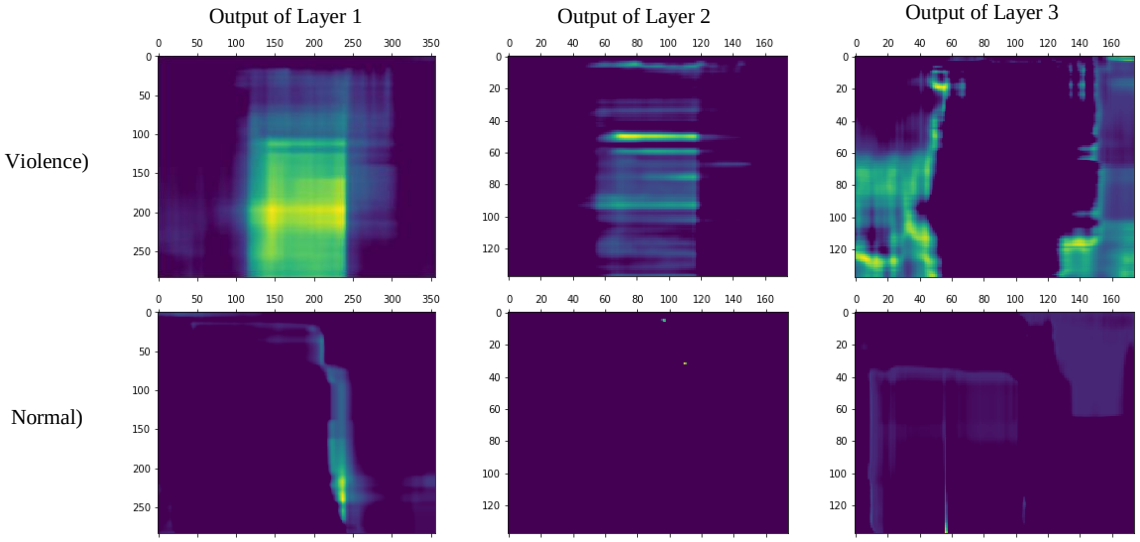
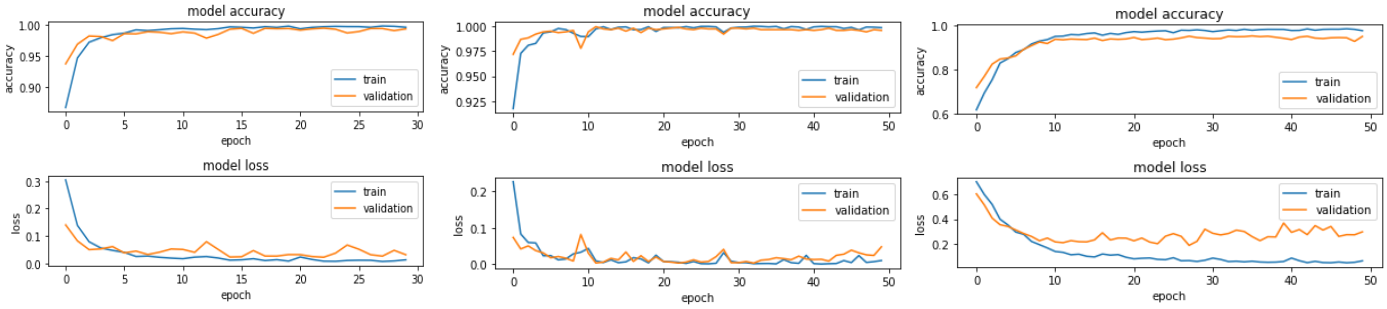


Fig. 2. Feature maps of convolutional layers for normal and violent actions



a) Hockey dataset b) Movies dataset c) Crowd dataset

Fig. 3. Accuracy and loss diagrams for Hockey, Movies, and Crowd datasets

Table. 2. Comparison of Vi-Net with other state-of-the-art methods

Methods	Datasets Accuracies (%)		
	Movies	**Crowd**	**Hockey**
ViF [7]	-	82	83
OViF [4]	-	78	87
Fast violence method [3]	98.9	-	90.1
STIP+BoW [6]	89.5	-	-
HFCNN [16]	99	-	94.6
KeyF [17]	99.5	-	87
Vi-Net (This work)	99	94	98

V. CONCLUSION

In this paper, a violence detection architecture based on convolutional layers and optical flow vectors was proposed. In violent action, motion patterns in consecutive frames are fast and they contain valuable information for behavior classification. Analyzing motion patterns with optical flow vectors in handcrafted approaches was a challenging problem, however, with convolutional layers, discriminative features were obtained automatically through a trainable network. The proposed architecture filtered the optical flow vectors and produced a specific feature map for each kind of violent behavior. Extensive experiments in several environmental conditions approved that the proposed architecture was able to detect violent behavior with a high percentage of accuracy in comparison with previous works.

REFERENCES

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

[2] De Souza, F.D., Chavez, G.C., do Valle Jr, E.A. and Araújo, A.D.A, "Violence detection in video using spatio-temporal features", *23rd SIBGRAPI Conference on Graphics, Patterns and Images* pp. 224-230, *2010.*

[3] Deniz, O., Serrano, I., Bueno, G. and Kim, T.K., "Fast violence detection in video", *international conference on computer vision theory and applications (VISAPP)* , Vol. 2, pp. 478-485, 2014.

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

[5] Gracia, I.S., Suarez, O.D., Garcia, G.B. and Kim, T.K., "Fast fight detection". *PloS one*, Vol. 10, p.e0120448, 2015.

[6] Nievas, E.B., Suarez, O.D., García, G.B. and Sukthankar, R., "Violence detection in video using computer vision techniques", *International conference on Computer analysis of images and patterns*, pp. 332-339, 2011.

[7] Hassner, T., Itcher, Y. and Kliper-Gross, O., "Violent flows: Real-time detection of violent crowd behavior", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-6, 2012.

[8] Colque, R.V.H.M., Caetano, C., de Andrade, M.T.L. and Schwartz, W.R., "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 27, pp.673-682, 2016.

[9] Ehsan, T.Z. and Nahvi, M., "Violence detection in indoor surveillance cameras using motion trajectory and differential histogram of optical flow", *8th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 153-158, 2018.

[10] Mohtavipour, S.M. and Shahhoseini, H.S., "A link-elimination partitioning approach for application graph mapping in reconfigurable computing systems" *The Journal of Supercomputing*, Vol. 76, pp.726-754, 2020.

[11] Mohtavipour, S.M., Mollajafari, M. and Naseri, A., "A novel packet exchanging strategy for preventing HoL-blocking in fat-trees" *Cluster Computing*, Vol. 23, pp.461-482, 2020.

[12] Mohtavipour, S.M. and Mollajafari, M., "An analytically derived reference signal to guarantee safety and comfort in adaptive cruise control systems" *Journal of Intelligent Transportation Systems*, Vol. 25, pp.1-20, 2021.

[13] Tran, H.T. and Hogg, D., "Anomaly detection using a convolutional winner-take-all autoencoder", In *Proceedings of the British Machine Vision Conference*, 2017.

[14] Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U. and Baik, S.W., "Violence detection using spatiotemporal features with 3D convolutional neural network", *Sensors*, Vol. 19, pp.2472, 2019.

[15] Smeureanu, S., Ionescu, R.T., Popescu, M. and Alexe, B., "Deep appearance features for abnormal behavior detection in video" , *International Conference on Image Analysis and Processing*, pp. 779-789 , 2017.

[16] Serrano, I., Deniz, O., Espinosa-Aranda, J.L. and Bueno, G., "Fight recognition in video using hough forests and 2D convolutional neural network", *IEEE Transactions on Image Processing*, Vol. 27, pp.4787-4797 , 2018.

[17] Khan, S.U., Haq, I.U., Rho, S., Baik, S.W. and Lee, M.Y., "Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies", *Applied Sciences*, Vol.9, p.4963, 2019.

[18] Dai, Q., Zhao, R.W., Wu, Z., Wang, X., Gu, Z., Wu, W. and Jiang, Y.G., "Detecting Violent Scenes and Affective Impact in Movies with Deep Learning", *MediaEval*, 2015.

[19] Peixoto, B., Lavi, B., Martin, J.P.P., Avila, S., Dias, Z. and Rocha, A., "Toward subjective violence detection in videos", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8276-8280, 2019.

[20] Farnebäck, G., "Two-frame motion estimation based on polynomial expansion", *Scandinavian conference on Image analysis*, pp. 363-370, 2003.