2022 12th Iranian/Second International Conference on Machine Vision and Image Processing (MVIP)
Department of Electrical and Computer Engineering, Shahid Chamran University of Ahvaz, Iran, 23 & 24 February 2022

DABA-Net: Deep Acceleration-Based AutoEncoder Network for Violence Detection in Surveillance Cameras

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

Manoochehr Nahvi
University of Guilan
School of Electrical Engineering
Rasht, Iran
nahvi@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*— Violent crime is one of the main reasons for death and mental disorder among adults worldwide. It increases the emotional distress of families and communities, such as depression, anxiety, and post-traumatic stress disorder. Automatic violence detection in surveillance cameras is an important research area to prevent physical and mental harm. Previous human behavior classifiers are based on learning both normal and violent patterns to categorize new unknown samples. There are few large datasets with various violent actions, so they could not provide sufficient generality in unseen situations. This paper introduces a novel unsupervised network based on motion acceleration patterns to derive and abstract discriminative features from input samples. This network is constructed from an AutoEncoder architecture, and it is required only to use normal samples in the training phase. The classification has been performed using a one-class classifier to specify violent and normal actions. Obtained results on Hockey and Movie datasets showed that the proposed network achieved outstanding accuracy and generality compared to the state-of-the-art violence detection methods.**

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

I. INTRODUCTION

In recent years, violent crimes such as extortion, harassment, and aggravated assault have been increased in different countries. Death due to violent crime is one of the threatening and significant issues in the world. Serious works such as behavior monitoring are needed to detect suspicious actions as fast as possible to reduce violence. For this purpose, Closed-Circuit TeleVision (CCTV) cameras received more attention to analyzing human activities. An automatic surveillance system based on computer vision techniques is a promising solution for checking cameras' videos without any need for human effort.

Violence detection approaches are categorized into handcrafted and deep learning techniques. Handcrafted techniques are based on complex mathematical equations to extract useful features from videos. Various methods such as gradient [1], motion [2], appearance [3], and trajectory [4] are used to represent violence and normal actions. After extracting features, they are learned with machine learning approaches such as random forest or nearest neighbor to classify each sample. In [5], an acceleration descriptor is introduced to represent movement patterns of human actions. Random transform is applied on the power spectrum of the successive frames to calculate acceleration. Support Vector Machine (SVM) and Adaboost are utilized for the classification part. A combination of appearance and motion information is proposed in [6] to include both spatial and temporal features. Scale Invariant Feature Transform (SIFT) technique is applied with several Gaussian filters on each image to extract key points in each frame. These points are described by motion information to consider the temporal aspect of violent actions. At the end of this violence detection method, SVM is used for the classification. Another technique based on keypoint detection is presented in [7]. In this work, Spatio-Temporal Interest Point (STIP) is extracted in a volume of frames. This technique is the improved version of 2D Harris corner detection and finds high variation points in spatial and temporal domains. Detected points are described with Histogram of Optical Flow (HOF). An optical flow-based method has been introduced in [8] to obtain the magnitude of optical vectors and split them into non-overlapping blocks. The descriptor of this method is constructed by subtracting and concatenating obtained blocks in consecutive frames. A trajectory-based approach is introduced in [4] to detect moving targets and track them with a Kalman filter. By analyzing the motion around the neighborhood points of trajectory, the violent activity has been detected. Another feature for representing violent action is gradient. A three-dimensional Histogram of Oriented Gradients (3DHOG) is introduced in [1] to obtain spatio-temporal information in videos. This descriptor is described by the Bag of Words (BoW) technique to extract a feature vector for each video sequence. Different classifications methods such as SVM or Kernel Extreme Learning Machine (KELM) are used for final evaluation. In [9], Optical flow information is computed, and the Canny operator is applied to the motion information to extract motion regions in each frame. They are described by Local Histogram of Optical Flow (LHOF) and Local Histogram of Oriented Gradients (LHOG) to obtain meaningful information. In [10], moving pixels are identified by differencing consecutive frames, and violence descriptor is built with attributes such as distance, shape, and perimeter. A violence detection approach in schools is introduced in [11] that detects objects with the nearest neighbor technique and extracts features such as height, width, centroid, aspect ratio, and area from the objects bounding box. Several morphological operations have been utilized to improve detection accuracy. Handcrafted techniques require complex

978-1-6654-1216-2/22/$31.00 ©2022 IEEE

computations to obtain valuable features from videos. Achieving high accuracy is difficult with these techniques due to low discrimination. High-level methods such as deep learning are needed to extract accurate features to analyze complex behaviors.

A deep neural network (DNN) is a machine learning technique that learns complex features from raw data. This technique has gained popularity in many fields of computer vision, such as object detection, action recognition, video understanding, and anomaly detection [12, 13]. DNN constitutes one input layer, some hidden layers, and one output layer. Several features of low-to-high levels are identified in the hidden layers. Feature extraction can be performed only by comparing input samples and their corresponding labels. With this capability, DNN can exploit accurate representation without the need for complex hand engineering. Convolutional Neural Network (CNN) is a famous architecture proposed for image analysis. It consists of convolutional layers for high-level feature extraction. Many filters are applied to the input image to learn local features in each layer. A supervised CNN network for violence detection is proposed in [14]. Frame sequences are given to the network, and 2D filters are applied to them. To improve the accuracy and high level interpretation, pre-processing techniques are considered in this work. A 3-Dimensional CNN (3DCNN) is presented in [15] to simultaneously consider spatial and temporal situations. After obtaining features with 3D filters, SVM is used for data classification. Some pre-trained networks are utilized for feature extraction, as well. For example, MobileNet is a deep CNN network with 28 layers trained on the Imagenet large-scale dataset with 1000 classes. Authors [16] used the MobileNet network for feature extraction and classified samples with a 3DCNN. Another method based on the VGG network is presented in [17] with a 2D BiGRU-CNN architecture. Gated Recurrent Unit (GRU) and CNN extract temporal and spatial features, respectively. A deep multi-network based on pre-trained AlexNet and GoogleNet is proposed in [18]. Authors in [19] used Long Short-Term Memory (LSTM) to focus on specific temporal conditions. LSTM is composed of gates that remember previous frame information. A combination of VGG and LSTM networks is introduced in [20]. With attention technique, high informative features are selected to learn with the LSTM network. As LSTM has millions of parameters, it is a complex network for training. A multi-stream 3DCNN network composed of spatial and temporal parts is proposed in [21]. In the spatial and temporal stream, frames and optical flow information have been computed, and the results have been concatenated for classification. Another multi-stream network is proposed in [22] to construct the volume of frames and apply it to handcraft and deep learning techniques. A combinational method for violence detection is proposed in [23]. In each frame, Key points are extracted with a Fast detector. These points in consecutive frames are concatenated to build a representative image. This image has been classified with a 2D CNN network. Also, authors in [24] used differential motion information as input of CNN network to obtain more discriminative features for the classification. This method did not consider temporal dependence in video sequences, and therefore some short-time sudden actions might be misinterpreted as violence. In the proposed approach, we consider long-term motion dependence among video frames to better discriminate between actions. Also, because the network [24] is supervised, it learns some specific violent actions, and in case of new violent actions, the performance will be decreased. Therefore, the generality of the detection model in every situation is a vital task that was not analyzed in most of previous works.

Many challenges in violence detection include inter-class similarities, intra-class variations, poor resolution, and camera viewpoint. Traditional methods that extract features manually cannot find discriminative patterns to overcome all of these challenges. With deep layers of DNN, it can now be possible to improve the features for action representation. As DNNs require large-scale datasets for training, limited input data will reduce the generality of classification. Methods of violence detection that have been published so far are supervised, and in the training phase, both normal and violence samples must be labeled manually. In this paper, a unique unsupervised network using normal behavior modeling is proposed. We utilize a combination of the handcrafted and deep network to model the normal behavior distribution. By comparing the distribution pattern between normal and violent behaviors, input samples will be classified. Our method includes three parts, handcrafted motion feature extraction, AutoEncoder network training with normal samples, and building behavior distribution from latent space of AutoEncoder to apply a one-class classifier.

The rest of this paper is arranged as follows: in section II, our handcrafted motion extraction is described, in section III, the proposed DABA network is presented, in section IV, the results of experiments on Hockey and Movie datasets are shown, and the conclusion is the last part of this paper.

II. PROPOSED MOTION FEATURE EXTRACTION

Motion patterns of violence and normal actions are entirely different. For example, in a fight situation, legs and hands will move suddenly and rapidly. But, when people perform normal activities like talking to each other or sitting, they show a slower pace than violent actions. Therefore, in the preprocessing phase of this paper, motion information is used for describing activities and building motion features. For motion estimation, $TV-L^1$ [25] method is used. This method improved Horn-Shunk [26] optical flow estimation to provide better accuracy. The brightness constancy assumption is considered as follows to obtain optical flow vectors:

$$\frac{d}{dt}I(x(t),y(t),t) = 0 \qquad (1)$$

$I(x,y,t)$ is video sequence, x,y are pixel locations and t denotes frame number. Equation (1) states that the pixel colour is constant along the video frames. By applying the chain rules on this equation, the following equation will be obtained:

$$\nabla I.u + \frac{\partial I}{\partial t} = 0 \qquad (2)$$

$u = (u_1, u_2)$ is the optical flow vector that demonstrates the displacement in x and y directions, respectively. This equation

is linear with two variables and for solving it, a constraint term should be added. Horn-Shunk used smooth constraint term as follows:

$$E_{HS}(u) = \int \left(\nabla I.u + \frac{\partial}{\partial t} I\right)^2 + \alpha \left(|\nabla_{u_1}|^2 + |\nabla_{u_2}|^2\right) \quad (3)$$

E_{HS} is energy function, the first term is optical flow constraint and the second term is the regularization term for obtaining smooth displacement fields. This function must be minimized to find a displacement vector between two consecutive frames. Authors in [25] computed Taylor expansions of equation (2) and obtained the following equation:

$$\rho(u) = \nabla I_1(x + u^0).(u - u^0) + I_1(x + u^0) - I_0(x) = 0 \quad (4)$$

ρ is the Taylor expansions of optical flow constraint equation. By using this equation, the energy function can be rewritten as follows:

$$E(u) = \int |\nabla_{u_1}| + |\nabla_{u_2}| + \gamma|\rho(u)| \quad (5)$$

Several numerical schemes were used to minimize the above equation and find the optical vectors. Finally, the magnitude of optical flow is calculated as follows:

$$M = \sqrt{{u_1}^2 + {u_2}^2} \quad (6)$$

Our handcrafted feature is based on acceleration feature in consecutive frames. For this purpose, the difference of magnitude is computed.

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

$M(x, y, t)$ is the magnitude in frame t which is subtracted from the magnitude $M(x, y, t + 1)$ in frame $t + 1$. To consider temporal information in our descriptor, the sliding window technique is utilized to build video segments. The video sequence is considered as follows:

$$I(x, y, t) = \{I(x, y, 1), I(x, y, 2), \dots, I(x, y, 2), \dots, I(x, y, T)\} \quad (8)$$

I is the gray level image, x and y denote pixels location and T is video length. Each video is split into several video segments with a length of L frames. Then, acceleration values of each segment are accumulated to obtain the final descriptor:

$$AL_S = \sum_{t=0}^{L-1} AL(x, y, t) \quad (9)$$

AL_S is the summation of acceleration in a video segment. This feature describes motion variations of actions and it can discriminate between normal and violent behaviours. In the experimental results section, we demonstrate that the best value for the video length is 7 frames.

III. Deep Acceleration-Based AutoEncoder (DABA) Network

In the previous section, we obtained an acceleration feature to process in the DABA network. It describes acceleration patterns of normal and violent actions in a video segment. In the DABA network, an AutoEncoder is utilized to abstract the feature and learn high-level patterns. Input samples will be converted to a compact form in a lower dimension space. This will separate the noisy information from the beneficial information. An AutoEncoder consists of an encoder that compacts the input to a lower dimension latent space and a decoder that maps the latent space to a regeneration of the input. So, the encoder part is helpful for feature learning and dimensionality reduction, and it is possible to extract valuable representation from the input samples. To train the AutoEncoder, normal examples are given to the network, and they should be reconstructed accurately at the output. A Loss function is determined based on the similarity of the inputs and outputs to update the network weights properly. After training the AutoEncoder part, the latent space of the hidden layer is given to a one-class classifier to discriminate between normal and violent samples. Figure 1 displays the details of the proposed violence detection framework, including feature extraction part and feature classification part.

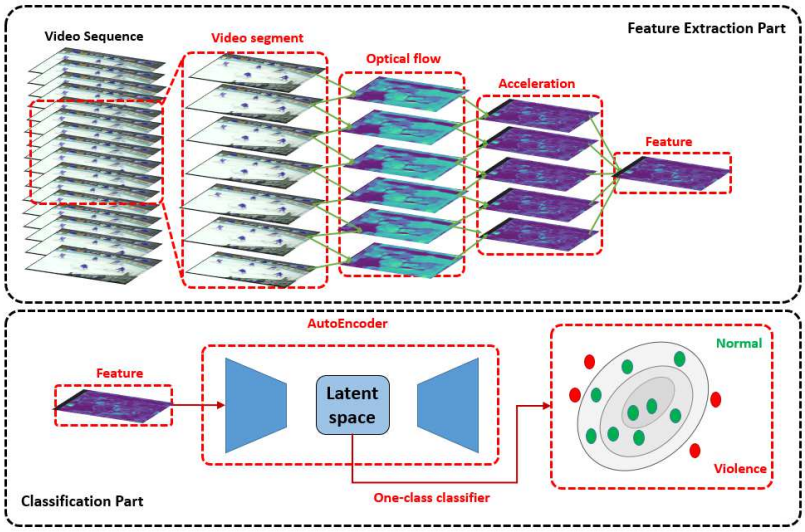

Fig. 1. Proposed violence detection framework including feature extraction and feature classification

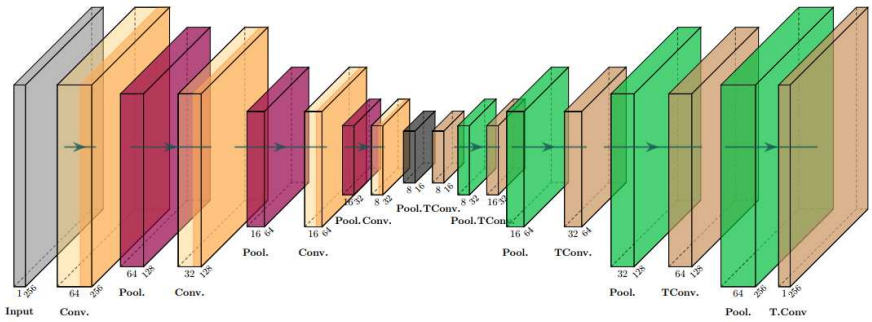

Fig. 2. The architecture of AutoEncoder with parameters of convolutional, maxpooling, transposed convolutional and upsampling layers

Figure 2 shows the architecture of the AutoEncoder part, such that the encoder is composed of convolutional and max-pooling layers, and the decoder is composed of transposed convolutional and upsampling layers. In the convolutional layer, input is convolved with kernels and passed through the activation function to produce a feature map. Weights of kernels are updated in the learning phase. After each convolutional layer, a max-pooling layer is used to diminish the dimension of the feature map and the number of network parameters. There is no weight in the max-pooling layer, and its objective is to prevent overfitting and computational cost. Transposed convolutional layers perform an inverse convolution operation. It applies kernels on the feature map to increase the dimension of the image. Upsampling is also the opposite of max-pooling and has no weights. It repeats the columns and rows of a feature map to increase its dimensionality. All parameters of the proposed AutoEncoder have been shown in figure 2. For obtaining the weights of the network, a loss function is specified as follows:

$$loss(AL, \widetilde{AL}) = MSE = \frac{1}{m}\sum_{i=1}^{m}\left(AL_i(x,y) - \widetilde{AL}_i(x,y)\right)^2 \quad (10)$$

m is the total number of segments, AL is the original feature, $\widetilde{AL}$ is the reconstructed feature and (x, y) denotes the pixel locations. In the loss function, Mean Square Error (MSE) is utilized to compare the original feature with the reconstructed feature and calculate the error of reconstruction. Loss function should be minimized to estimate the optimal weights of the network. For this purpose, Adadelta optimizer that is a gradient descent variation is applied. This optimizer measures the gradient of the loss function with respect to the network's weights. Updating the weights follows the backpropagation rule from the last hidden layer to the first hidden layer.

After the training phase, the latent space of AutoEncoder is used to extract abstracted information of input samples. We employed the Local Outlier Factor (LOF) [27] as a one-class classifier to determine the output label to detect anomalous samples. Because of the considerable difference between normal and violence distributions, LOF is appropriate for discriminating between classes. It identifies outliers by computing the density of the samples and comparing it with the density of normal trained samples. If they have similar densities, the test sample is selected as normal. Otherwise, it will be considered violence.

IV. Experimental Result

In this section, we compared the proposed violence detection approach with state-of-the-art methods. The simulation framework for train and test is performed in the GoogleColab environment using Keras and Tensorflow libraries. Two Hockey and Movies [7] datasets are considered for the evaluation. The hockey dataset contains more than 40000 frames of normal and violent actions. This dataset includes the actions of National Hockey League (NHL) players in several games. There are 500 clips of normal and 500 clips of violent actions in this dataset. 90% of normal clips have been used for training, and the remaining clips, including both normal and violent actions, are selected for testing. Videos are taken at a rate of 25 frames per second with a resolution of 360×288. This dataset is challenging because the camera is moving, and the background motions are combined with the movement of the targets. The Movies dataset is comprised of 200 video clips of violence and normal actions in different environments. Violence clips are gathered from various action movies, and normal clips are taken from available action recognition databases. The frame rate and resolution of these videos are different for each clip. The camera is still in most clips. This dataset is more similar to videos of real-world surveillance systems placed in public and private environments.

To train the AutoEncoder part, only normal samples are required to learn how to abstract and compact the image of normal motions and reconstruct it from a low dimensional space. Normal datasets are divided into training and validation samples. Training samples are used for updating the network weights, and validation samples are used to check the network performance on new unseen samples. For each dataset, training and validation loss have been displayed in figure 3. As can be seen, training and validation loss is decreased and converged after several epochs. The loss curve in the Movie dataset shows lower values and fluctuations because the camera viewpoint is fixed, and it is easy to identify motions. As the camera viewpoint in the Hockey dataset is moving, it is more challenging to train the network accurately.

We performed another experiment to choose the length of the video segment which the results are shown in figure 4. If the length is selected small, it is impossible to consider the

temporal aspect of actions and degrade the accuracy. Also, when the length is large, the motions of violence and normal samples will not be discriminative. As it can be seen, the optimal value for the video segment is 7.

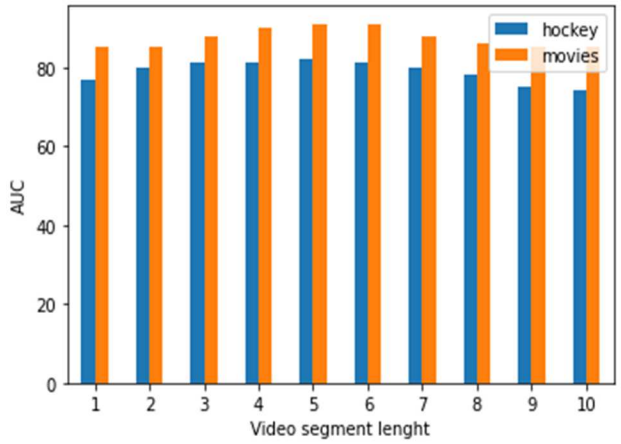

Fig. 4. AUC values for different video segment lengths in Hockey and Movies dataset

Table 1 presents the comparison results in terms of Area Under ROC Curve (AUC). This metric computes the average of accuracy values in different thresholds. LMP [2] is based on dictionary learning and provided 84% AUC for the Hockey dataset and 92% AUC for the Movies dataset. AMDN [3] used motion and appearance of normal and violence classes to classify each sample and obtained 89% AUC for the Hockey dataset. Another supervised method proposed in [5] which used acceleration attributes obtained 95% AUC in the hockey dataset and 74% AUC in the Movies dataset. Authors in [10] used blob analysis for violence detection and obtained 76% AUC in the Movies dataset and 97% AUC in the Hockey dataset. A convolutional neural network proposed in [15] obtained 87% in the Hockey dataset and 93% in the Movies dataset. Authors in [16] utilized pre-trained MobileNet model to extract features and obtained 96% and 99% for the Hockey and Movies dataset, respectively. In [23] a combination of Handcrafted and deep based techniques are used to obtain 94% in the Hockey dataset and 99% in the Movies dataset. Another convolutional neural network proposed in [24] that obtained 98% in the Hockey dataset and 99% in the Movies dataset. Finally, our proposed network provided 82% AUC for the Hockey dataset and 92% AUC for the Movies dataset. It should be noted that the proposed network has been trained only on normal samples in an unsupervised manner. The implementation of this framework is more straightforward, and also, the accuracy is comparable and acceptable in comparison with supervised work. As normal samples are more available than violent samples, this network can be trained more effectively compared to the supervised techniques.

Table. 1. AUC comparison between violence detection works

Supervision type	Approach	AUC (%)	
		Hockey dataset	Movies dataset
Supervised	LMP [2]	84	92
	AMDN [3]	89	N/A
	Fast-vi + SVM [5]	85	90
	Fast-fi + SVM [10]	87	72
	C3D [15]	87	93
	3D CNN [16]	96	99
	HF[23]	94	99
	Vi-Net[24]	98	99
Unsupervised	This work	82	92

To analyze the network generality, we performed another experiment to measure the performance of our network on new unseen environments. The proposed network is trained on the Hockey dataset and tested on the Movies dataset. As it can be seen, our method provides 39% improvement compared to 3D CNN [16]. Also, by training the network on the Movies dataset, and testing on the Hockey dataset, we obtained 10% improvement compared to 3D CNN [16]. Therefore, the generality of our method in new environments is superior to the state-of-the-art approach.

Table 2. Generality experiment result for new unseen data

Training dataset	Testing dataset	AUC (%)		
		Our method	3D CNN[16]	Improvement (%)
Hockey	Movies	88	49	+39
Movies	Hockey	73	63	+10

I. CONCLUSION

In this paper, an unsupervised violence detection network was introduced. The architecture of this network was made of AutoEncoder that made it possible to perform the training phase only with normal samples. Also, to emphasize discriminative attributes of violent behavior, a novel motion acceleration feature was proposed. This feature was obtained analytically based on the differential of optical flow vectors. Moreover, the classification was carried out by using the latent space of the trained AutoEncoder and evaluating the similarity between the distribution of normal and violent actions. This approach was more straightforward in implementation as it just trained on normal samples. Experimental results on datasets showed that the obtained accuracy was comparable with previous supervised works, and up to 92% accuracy was achieved.
In future works, we intend to add more discriminative features in the network and build a multi-stream unsupervised architecture to improve the violence detection accuracy.

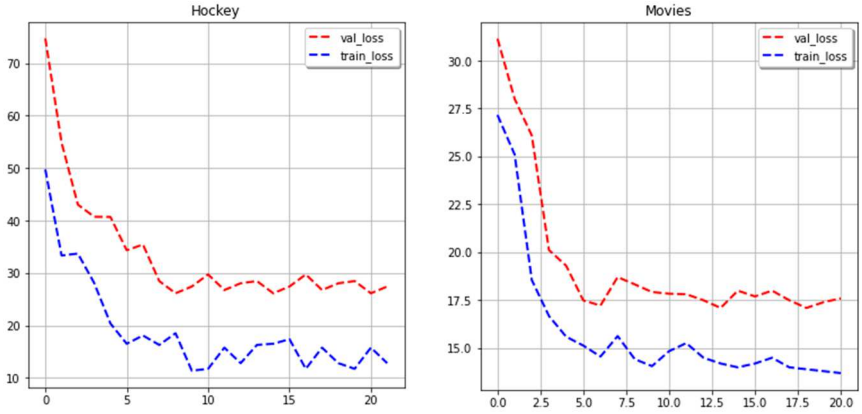

Fig. 3. Training and validation loss curves for Hockey and Movies datasets

REFERENCES

[1] Yu, J., Song, W., Zhou, G. and Hou, J.J., 2019. Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation. Multimedia Tools and Applications, 78(7), pp.8497-8512.

[2] Ward R, Guha T (2012) Learning sparse representations for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34: 1576–1588.

[3] Xu, D., Yan, Y., Ricci, E., Sebe, N.: Detecting anomalous events in videos by learning deep representations of appearance and motion.Comput. Vis. Image Underst. 156, 117–127 (2017)

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

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

[6] Senst, T., Eiselein, V., Kuhn, A. and Sikora, T., 2017. Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation. IEEE transactions on information forensics and security, 12(12), pp.2945-2956.

[7] Nievas, E.B., Suarez, O.D., García, G.B. and Sukthankar, R., 2011, August. Violence detection in video using computer vision techniques. In International conference on Computer analysis of images and patterns (pp. 332-339). Springer, Berlin, Heidelberg.

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

[9] Zhou, P., Ding, Q., Luo, H. and Hou, X., 2018. Violence detection in surveillance video using low-level features. PLoS one, 13(10), p.e0203668.

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

[11] Ye, L., Wang, L., Ferdinando, H., Seppänen, T. and Alasaarela, E., 2020. A Video-Based DT–SVM School Violence Detecting Algorithm. Sensors, 20(7), p.2018.

[12] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[13] Farha, Y.A. and Gall, J., 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3575-3584).

[14] Mohtavipour, S.M., Saeidi, M. and Arabsorkhi, A., "A multi-stream CNN for deep violence detection in video sequences using handcrafted features" The Visual Computer, pp.1-16, 2021.

[15] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[16] 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, 19(11), p.2472, 2019.

[17] Mumtaz, A., Bux Sargano, A. and Habib, Z., 2020. Fast Learning Through Deep Multi-Net CNN Model For Violence Recognition In Video Surveillance. The Computer Journal.

[18] Sumon, S.A., Goni, R., Hashem, N.B. and Rahman, R.M., 2020. Violence detection by pretrained modules with different deep learning approaches. Vietnam Journal of Computer Science, 7(01), pp.19-40.

[19] Traoré, A. and Akhloufi, M.A., 2020, June. 2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-to-End Violence Detection in Videos. In International Conference on Image Analysis and Recognition (pp. 152-160). Springer, Cham.

[20] Halder, R. and Chatterjee, R., 2020. CNN-BiLSTM Model for Violence Detection in Smart Surveillance. SN Computer science, 1(4), pp.1-9.

[21] Li, C., Zhu, L., Zhu, D., Chen, J., Pan, Z., Li, X. and Wang, B., 2018, December. End-to-end multiplayer violence detection based on deep 3D CNN. In Proceedings of the 2018 VII International Conference on Network, Communication and Computing (pp. 227-230).

[22] Li, H., Wang, J., Han, J., Zhang, J., Yang, Y. and Zhao, Y., 2020. A novel multi-stream method for violent interaction detection using deep learning. Measurement and Control, 53(5-6), pp.796-806.

[23] 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.

[24] Ehsan, T.Z. and Mohtavipour, S.M., 2020, December. Vi-Net: A Deep Violent Flow Network for Violence Detection in Video Sequences. In 2020 11th International Conference on Information and Knowledge Technology (IKT) (pp. 88-92). IEEE.

[25] Zach, C., Pock, T. and Bischof, H., 2007, September. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium* (pp. 214-223). Springer, Berlin, Heidelberg.

[26] Berthold K. P. Horn and Brian G. Schunck. \Determining optical flow": a retrospective. Artificial Intelligence, 17:185203, 1981

[27] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000, May. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).