The Visual Computer
https://doi.org/10.1007/s00371-021-02266-4

ORIGINAL ARTICLE

A multi-stream CNN for deep violence detection in video sequences using handcrafted features

Seyed Mehdi Mohtavipour[1] · Mahmoud Saeidi[2] · Abouzar Arabsorkhi[2]

Accepted: 18 July 2021

Abstract
Intelligent video surveillance systems have been used recently for automatic monitoring of human interactions. Although they play a significant role in reducing security concerns, there are many challenges for distinguishing between normal and abnormal behaviors such as crowded environments and camera viewpoint. In this paper, we propose a novel deep violence detection framework based on the specific features derived from handcrafted methods. These features are related to appearance, speed of movement, and representative image and fed to a convolutional neural network (CNN) as spatial, temporal, and spatiotemporal streams. The spatial stream trained the network with each frame in the video to learn environment patterns. The temporal stream contained three consecutive frames to learn motion patterns of violent behavior with a modified differential magnitude of optical flow. Moreover, in spatio-temporal stream, we introduced a discriminative feature with a novel differential motion energy image to represent violent actions more interpretable. This approach covers different aspects of violent behavior by fusing the results of these streams. The proposed CNN network is trained with violence-labeled and normal-labeled frames of 3 Hockey, Movie, and ViF datasets which comprised both crowded and uncrowded situations. The experimental results showed that the proposed deep violence detection approach outperformed state-of-the-art works in terms of accuracy and processing time.

Keywords Computer vision · Deep learning · Convolutional neural network · Human behavior analysis

1 Introduction

In recent years, due to security concerns and low-cost of Closed-Circuit TeleVision (CCTV) cameras, the use of video surveillance systems in public or private places has been expanded. These cameras are utilized to monitor indoor and outdoor environments such as homes and offices to analyze events. In traditional ways, frames of a camera are stored in a system to investigate occurred abnormal events in advance. Also, in some cases, an operator is employed to monitor these cameras. As there are several CCTV cameras in different places, employing an operator to monitor them is too costly and human errors like fatigue and distraction would degrade the accuracy. But nowadays, many efforts have been made to create intelligent CCTV cameras in which behaviors have been analyzed automatically in the shortest possible time. These systems are one of the active areas in the machine vision and artificial intelligence communities [1, 2]. Researchers are focusing on the design of real-time smart systems to detect abnormal behaviors and prevent dangerous accidents. Human actions can be categorized into three different levels. The first level is gesture recognition including movements of a body part such as hand raising or punching [3]. The next level covers simple actions like running, walking, and sitting that lasts for a short duration [4]. The top level of human actions involves interactions between multiple targets composed of complex behaviors like violence [5].

Violence detection is a challenging topic because actions are different from one person to another and everyone performs them in his/her own way [6]. On the other hand, there

✉ Seyed Mehdi Mohtavipour
mehdi_mohtavipour@elec.iust.ac.ir

Mahmoud Saeidi
msaeidi40@itrc.ac.ir

Abouzar Arabsorkhi
abouzar_arab@itrc.ac.ir

[1] School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran

[2] Iran Telecommunication Research Center, Tehran, Iran

Published online: 26 July 2021

are some similarities between normal and violent behaviors such as hugging and handshaking which makes it difficult to distinguish between them. Moreover, changes in camera viewpoint lead to different visual appearances of actions [7]. For example, a similar action from a different viewpoint can be seen in different forms and this will increase the complexity of the recognition system. A good recognition method must be independent of these conditions and detect anomalies accurately.

The analysis of human action involves low- and high-level parts. Low-level analysis is based on machine vision techniques. In this step, different feature descriptors using texture [8–10], motion [11–15], and shapes [16–19] are extracted by handcrafted methods. For this kind of analysis, it is difficult to find a global feature to represent target behavior in every condition such as poor camera resolution [20], shadow of targets [21], noise in the image [22], partial or full target occlusion [23], and illumination changes [24]. On the other hand, high-level analysis is based on machine learning techniques such that models are trained to classify actions to different categories There are many classification approaches for categorizing normal and violent behavior such as support vector machine (SVM) [25], nearest neighbor search (NNS) [26], random forest (RF) [27], and deep learning [28–30].

Among machine learning techniques, deep learning networks with convolutional layers are widely used in the field of violence detection to learn different complex features of actions [30]. They are more adapted for analyzing images and video, and their convolutional layers apply various types of filters to extract specific features. After training the network, violent behavior in frames is detected with minimal processing time which makes it appropriate for online surveillance cameras. Although it is possible to find an acceptable solution at the lowest time, provided detection accuracy has remained a problem. To improve detection accuracy, it is required to do some preprocessing steps before training the network. This will help the network to learn different aspects of violent behavior and extract features more precisely.

In this paper, we focus on the impact of input features in a deep violence detection network. We show that simultaneous coverage of spatial, temporal, and spatiotemporal aspects for violence behavior with discriminative features would be very helpful in achieving an accurate deep network. We proposed discriminative features that are based on the appearance, speed of movement, and representative image of actions (by accumulating frames), and they have been derived in a differential mode to better emphasize the key points of violent behavior.

The remainder of this paper is organized as follows, in Sect. 2 related works of violence detection methods are described and compared to each other, in Sect. 3 the proposed framework including feature extraction and CNN architecture is introduced, in Sect. 4 the experimental results and discussion on datasets are presented, and finally, Sect. 5 shows the conclusion and future works of this paper.

2 Related works

Most Abnormal behavior recognition approaches are based on action representation by handcrafted methods. This representation is categorized into two global and local parts.

2.1 Global handcrafted representation

The model of this group describes the appearance or motion of the entire human body. It models the global structure of a target, and the region of interest is encoded as a whole part. The global representation of human action is implemented by silhouette [31] or optical flow [32]. The first one uses a mask shape by background subtraction or frame differencing technique in order to obtain moving region information. In [33], the authors used silhouette to introduce five classes of violent action. These classes include pushing, punching, shaking, knocking, and physical contact. In this method, moving targets are extracted by frame differencing technique. For each frame, spatial features are derived using the silhouette information such as the size and center of gravity. As actions have been restricted to certain classes, violent behaviors that are not similar to the above classes will not be detected. The second group uses optical flow information to represent human actions. Optical flow is the distribution of the apparent velocities of the target and is an accurate feature to estimate the motion in each frame. In [34], optical flow context histogram (OFCH) feature is presented which is a combination of the histogram of optical flow (HOF) and histogram of orientation. Optical flow is calculated for each frame and points are distributed in the log-polar system based on the magnitude and orientation range. OFCH feature is calculated based on the distribution of these points. This method shows false alarms for actions with high optical flow magnitude like running. Another method based on optical flow is used in [35] in which frames are accumulated over time and the video is split into a fixed-dimensional non-overlapping spatiotemporal cube. For each cube, the histogram of optical flow orientation and magnitude and entropy (HOFME) is calculated. Also, authors in [36] used differential histogram of optical flow (DHOF) to consider its variations and showed some improvements in the accuracy of violence detection. The violent flow vector (ViF) descriptor is proposed in [37] to split frames into non-overlapping spatial cells. For each cell, variation in the optical flow magnitude per pixel is calculated and a fixed dimension histogram is obtained. The final vector is obtained by accumulating the histogram of cells in consecutive frames. This technique demonstrates good results in crowded or close-up situations, but in scenes which targets appear in a small region of the frame, the efficiency is not acceptable as most of the cells or cubes are nonviolent. Moreover, it was assumed that the length of frames is fixed for each action.

2.2 Local handcrafted representation

It describes the motion and appearance information as a collection of independent local regions of the video. In these methods, key points in the video are extracted using feature detector approaches like Harris 3D [38] or cuboid 3D [39] and then local descriptors capture motion or shape information in the neighborhoods of the detected points [40]. In [41], spatiotemporal interest points (STIP) are extracted using a Harris 3D detector. For each neighborhood of detected points, histogram of oriented gradient (HOG), histogram of optical flow (HOF), and some combination of them are obtained for action classification and the video is represented by using the bag of words (BOW) framework. This method showed high runtime and is used only in offline violence detection. To improve these disadvantages, Distribution of Magnitude and Orientation of Local Interest Frame (DiMOLIF) is proposed in [42] to extract STIPs with Harris 3D detector and modeling movements of neighborhood points.

2.3 Deep learning

These methods reduced the computation complexity and showed improvements in the learning of complex interactions. A violence classification is introduced in [43] using 2D convolutional neural network (CNN). This method used a representative image (RI) as input to discriminate normal and violent behavior. The RI is built by combining several consecutive frames such that heavy weights are assigned to regions of a frame with more motion information. The remained regions such as still background received light weights. After that, the CNN network is trained by the normal and violent dataset of RIs. A pre-trained CNN network named MobileNet is used for violence detection in [44]. To use this network, the input frames have been divided into shots. Among these shots, one of them is selected as key shot, and based on it, the network is fine-tuned by the transfer learning method. In [45] three-dimensional convolutional network (3D ConvNets) is introduced to detect violence in videos. The network included 15 layers and in each convolutional layer, 3D kernels with a size of $3 \times 3 \times 3$, and stride of 1 were used. This method used spatial and temporal information of each volume but could not achieve higher accuracy. A combination of two CNN networks (MobileNet and 3D CNN) is introduced in [46] for violence detection. In this method, moving targets are detected using MobileNet network. Afterward, a volume of 16 frames including moving targets is considered as input of the 3D CNN network. The authors claimed more accuracy in this method. Another model based on 2D CNN and support vector machine (SVM) is proposed in [47]. The utilized architecture for the 2D CNN in this model is a bi-channel network to extract appearance and temporal features. After feature extraction, two linear SVMs have been used for violence classification. A FightNet architecture has been introduced in [48] by modifying the temporal segment network (TSN) model. The TSN model is a pre-trained bi-channel network that can receive both spatial and temporal features. In FightNet a new acceleration input is added to the TSN model. Frames, optical flow field, and acceleration were the inputs of the FightNet network. A network combination model for violence detection is proposed in [49] which is called ConvLSTM. This architecture used the CNN network to extract spatial features as the input of LSTM (Long Short-Term Memory) network for feature analyzing and action classification. Also, target trajectory is used in [50] to define three-dimensional volumes for data preparations. After that, 2D CNN and SVM networks are utilized for feature extraction and data classification.

Table 1 summarizes the mentioned methods for violence detection. It has been concluded that silhouette-based methods are very sensitive to camera viewpoint but provides useful information about actions in different conditions of color, texture, and light. Optical flow-based methods provide more information but to represent actions, they need estimations for reducing heavy computations.

In local representation-based methods, there is no need for background subtraction or tracking algorithm but spatiotemporal dependency is not modeled in them. Finally, deep learning-based methods demonstrate acceptable accuracy with little processing time. However, they require large datasets or discriminative sample data as well as strong GPUs in the phase of network weights training.

3 Proposed violence detection framework

It was described in the previous section that there are many challenges in violence detection including appearance variations, movement speed variations and, etc. To focus more on these conditions, it is appropriate to train a deep neural network based on the specific features of violent behavior. In this section, a three-stream deep network has been proposed with three spatial, temporal, and spatiotemporal inputs. Spatial stream analyses the appearance in video sequences by focusing on a grayscale image. The temporal stream considers the speed of movement for moving targets by using a modified optical flow approach. In the spatiotemporal stream, we form the shape of actions by building a differential motion energy image (DMEI). This image easily describes the occurring violent actions among different

Table 1 Summary advantages and disadvantages of violence detection methods in surveillance cameras

Method	Feature types		Advantages and disadvantages
Handcrafted	Global features	Silhouette based	+ Nonsensitive to color, texture, and light changes – Sensitive to viewpoints
		Optical flow based	+ Accurate information about the movement and velocity of the targets – Need estimation to find solutions
	Local features		+ No need for background subtraction or tracking algorithm – Sensitive to noise – Heavy computational – Not modeling the spatiotemporal relationship between regions
Deep learning			+ Automatically learn the video features + Learn complex features + Fast detection – Need powerful GPU for training phase – Need large dataset for training phase –Discriminative sample data

persons. Figure 1 shows the block diagram of the proposed deep violence detection framework with handcrafted features. A sliding window technique is used in the proposed framework to obtain frame sequence. This frame sequence is fed as the input of the CNN network in three streams. In each frame sequence, one, three, and six frames are selected for the input of spatial, temporal, and spatiotemporal streams, respectively. The plus sign at the end of this framework is the concatenation function.

3.1 Spatial stream

Each video consists of frame sequences and each frame is made of important information about humans and their environment. To learn and extract spatial features such as the shape of body and appearance, frames should be processed directly in the CNN network. For example, fighting with physical contact or using weapons to shoot somebody includes a unique shape and appearance in the environment. Frame sequences can be represented by the following equation:

$$S(x,y) = \left\{ S_1(x,y), S_2(x,y), \ldots, S_T(x,y) \right\} \tag{1}$$

In this equation, (x, y) is the position of pixels in the frame, and T is the sequence length. As the sequence length in this paper is computed six frames, we selected the middle frame ($S_4(x, y)$) for the spatial input. The size of each frame is equal to $(H \times W \times 3)$, where H is the height, W is the width and 3 is the number of image channels in the frame.

Greyscale image is calculated by Eq. (2).

$$I_4(x,y) = 0.2989 \times R_4(x,y) + 0.5870 \times G_4(x,y) + 0.1140 \times B_4(x,y) \tag{2}$$

In this equation, $R_4(x, y)$ is the image of red channel, $G_4(x, y)$ is the image of green channel, $B_4(x, y)$ is the image of blue channel, and $I_4(x, y)$ represents the gray-level image. Coefficients of this equation are obtained based on the wavelength ratio of red, blue, and green color [51]. The gray-level image is the input of the CNN network in the spatial stream to extract appearance features from the frames of video. The spatial stream in the CNN network comprises six convolutional and two fully connected layers. There are two

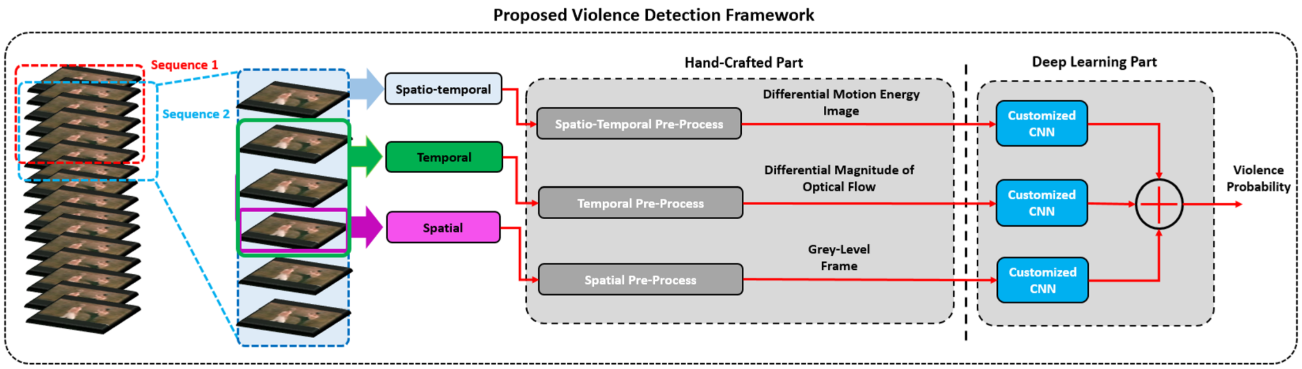


Fig. 1 Proposed deep violence detection framework using handcrafted and deep learning parts

advantages for using convolutional layers in the learning of local patterns of the image as follows:

(1) Learning patterns are translation-invariant in the CNN network. For example, after learning the action of fighting between two persons on the left side of the image, the network is able to detect any similar actions in any other position of the image.
(2) The CNN network follows a hierarchal procedure to better understand the complicated actions. In the first layer, small local patterns such as edges and corners are extracted. In the next depth layers, larger patterns such as contour or body shape of humans are considered for network training.

3.2 Temporal stream

In addition to spatial features, there are some motion information in the frames of video. In violent actions, movements of hands, legs, and other parts of the body are fast and happen suddenly. This is a distinguishable feature between normal and violent behavior with emphasis on the motion. One of the methods for estimating motion in consecutive frames is optical flow. With this method, complete information about the speed and direction of each moving target is obtained. Considering u and v as optical flow vectors in directions of x and y, the optical flow constraint equation is obtained by brightness constancy assumption as follows [52]:

$$I_x u + I_y v + I_t = 0 \tag{3}$$

There are three constant values in the optical flow constraint equation denoted by I_x, I_y, and I_t as image derivatives in the direction of x, y, and t. Several (u, v) pairs satisfy Eq. (3), but to derive a unique solution, Horn Schunk in [52] proposed a smoothness constraint. Using this new constraint, it is assumed that neighborhood pixels contain similar speeds. Equation (4) shows this constraint in the energy function of optical flow.

$$E = \iint [\left(I_x u + I_y v + I_t\right)^2 + \alpha^2(\|\nabla u\|^2 + \|\nabla v\|^2)]\mathrm{d}x\mathrm{d}y \tag{4}$$

The energy function E consists of two parts, the first part is the optical flow equation and the second part is smoothing constraint which is the gradient of optical flow and α in this equation is the smoothness constant. To satisfy the smoothness constraint, the above energy function should be as low as possible. A recursive equation has been proposed in [53] to obtain the minimum value in Eq. (4) as follows:

$$u^{n+1} = \overline{u}^n - \frac{I_x\left[I_x\overline{u}^n + I_y\overline{v}^n + I_t\right]}{\left(\alpha^2 + I_x^2 + I_y^2\right)} \tag{5}$$

$$v^{n+1} = \overline{v}^n - \frac{I_y\left[I_x\overline{u}^n + I_y\overline{v}^n + I_t\right]}{\left(\alpha^2 + I_x^2 + I_y^2\right)} \tag{6}$$

where n is the number of iterations and 100 iterations would be enough for this phase. $\overline{u}$ and $\overline{v}$ are the average optical flow around each pixel denoted by i, j in the image and obtained with Eq. (7) and (8).

$$\begin{aligned}\overline{u} \approx &\frac{1}{6}\left(u^n_{i-1,j} + u^n_{i+1,j} + u^n_{i,j-1} + u^n_{i,j+1}\right)\\ &+ \frac{1}{12}\left(u^n_{i-1,j-1} + u^n_{i+1,j-1} + u^n_{i-1,j+1} + u^n_{i+1,j+1}\right)\end{aligned} \tag{7}$$

$$\begin{aligned}\overline{v} \approx &\frac{1}{6}\left(v^n_{i-1,j} + v^n_{i+1,j} + v^n_{i,j-1} + v^n_{i,j+1}\right)\\ &+ \frac{1}{12}\left(v^n_{i-1,j-1} + v^n_{i+1,j-1} + v^n_{i-1,j+1} + v^n_{i+1,j+1}\right)\end{aligned} \tag{8}$$

The magnitude of optical flow (MOF) which represents movements of each pixel between two consecutive frames is obtained as follows:

$$MOF = \sqrt{u^2 + v^2} \tag{9}$$

MOF shows the pixel velocity, so it is a valuable feature in detecting violent behavior. However, velocity variations are more discriminative than absolute velocity. In this paper, we utilize the differential magnitude of optical flow (DMOF) as follows:

$$DMOF = |MOF_{(t+1,t)}(x, y) - MOF_{(t,t-1)}(x, y)| \tag{10}$$

In this equation, $t, t-1, t+1$ are three consecutive frames in the video sequence and x, y are the pixel positions.

3.3 Spatiotemporal stream

To accumulate spatial and temporal information, we used motion energy image (MEI) [54] for the input of the third stream of the CNN network. MEI is a binary image that represents moving objects in consecutive video frames. In this image, foreground objects in the sequence are processed and information are gathered in a three-dimensional way including spatial information in the frames and temporal information in the sequence. This image comprises valuable information about the occurring actions in the video sequence. To build an MEI, Moving objects are extracted by the frame differencing method. The output of this method is a binary image $D_t(x, y)$ including white regions of moving objects and black regions of stationary objects and background. To construct $D_t(x,y)$, it is required to compute the difference image between $t, t-1$ consecutive frames. The accumulation of $D_t(x,y)$ in consecutive frames is the MEI image and calculated as follows:

$$MEI_t(x,y) = \bigcup_{i=0}^{\tau-1} D_{t-i}(x,y) \tag{11}$$

In this equation, i is the index of the frame and τ is the number of frames for the accumulation. As the length of the video sequence is selected six frames, then τ is equal to 6. *MEI* is a high-level feature that demonstrates a global description of actions. As violent behavior occurs with very fast actions, this feature may not be able to construct an appropriate descriptor. This problem arises from the fact that there are many overlapping regions in the summation image of foreground objects. To overcome this problem, in this paper we extend the *MEI* feature into the differential mode for better representation of violent behavior. *DMEI* feature can be computed as follows:

$$DMEI_t(x,y) = \bigcup_{\substack{i=0 \\ i=i+2}}^{\tau-1} (D_{t-i}(x,y) - D_{t-(i+1)}(x,y)) \tag{12}$$

In this equation $D_{t-i}(x,y)$ and $D_{t-(i+1)}(x,y)$ are the foreground binary image for frame number $(t-i, t-i-1)$ and $(t-i-1, t-i-2)$, respectively. Figure 2 shows examples of *MEI* and *DMEI* extraction in six consecutive fight frames. As can be seen in *MEI* extraction, the white area includes all moving pixels and it cannot describe the action behavior well, but in *DMEI* extraction, only very fast pixels have been detected and it can produce a discriminative feature for violent behavior. Also, in DMEI feature, unimportant regions with low speed have been removed from the feature extraction procedure and it can be seen in Fig. 2 that they only contain small white regions.

3.4 Network architecture

As mentioned in previous sections, there are some discriminative information within the video frames that can be utilized for violence detection. Appearance, movements, and spatiotemporal information almost cover requirements to train a deep neural network. In this paper, a CNN network with three streams has been used to learn and identify violent action. Figure 3 shows the proposed architecture with selected configurations.

In the convolutional layers, the number of filters and size of them have been defined as $num \times size \times size$. The nonlinear activation function in each convolutional layer is *ReLu* function and computed as follows:

$$ReLu(x) = \max(0, x) \tag{13}$$

After each convolutional layer, there is one *maxpooling* layer to prevent over-fitting, reduce the spatial size of the representation and therefore, reduce the number of parameters in the network. The configuration of *maxpooling* layer which is defined as $size \times stride$ is equal to 2×2. At the end of convolutional layers, there are two fully connected layers with 64 neurons. Also, between these fully connected layers, there is one *dropout* layer to improve the generality of training. The elimination rate of this *dropout* layer is selected 0.5. Finally, *Softmax* layer at the end of network architecture produces output probability to classify video sequences. This layer classifies the output of the network into two normal and violent classes.

In the training phase of the CNN network, predicted outputs are compared with actual outputs to build a *Loss function*. This is performed for every sample of the dataset to

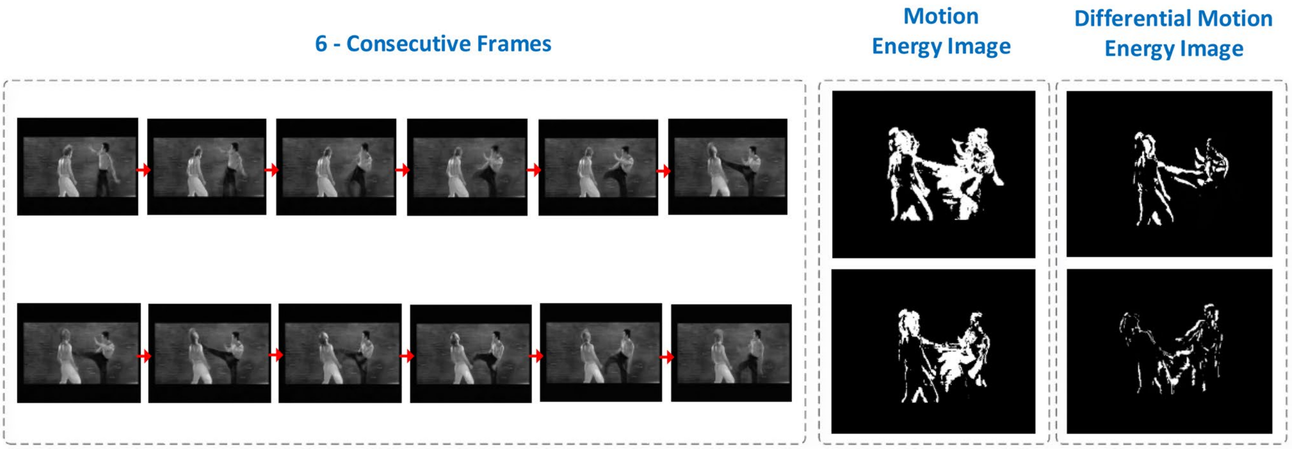


Fig. 2 MEI and DMEI image extraction in six consecutive frames

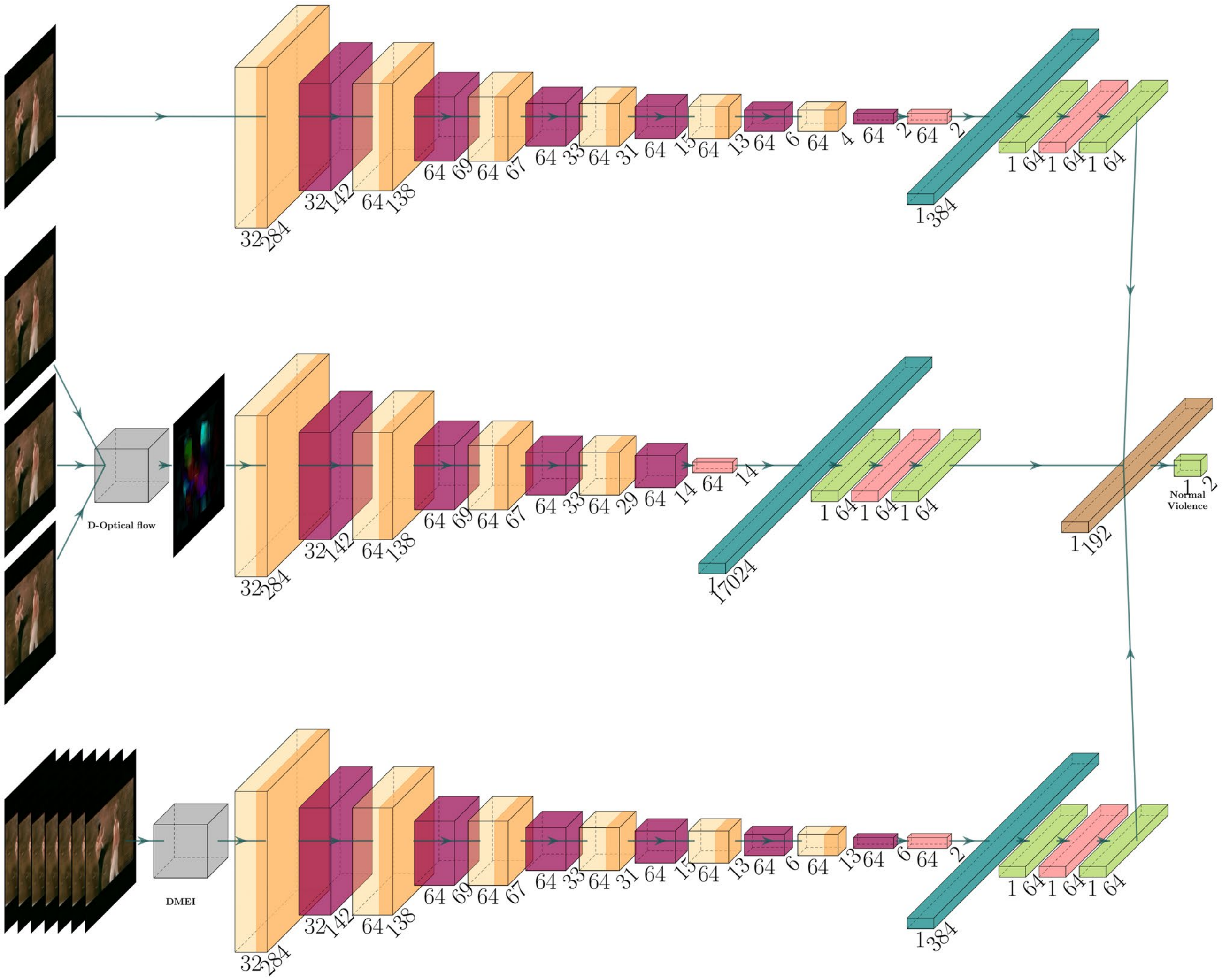


Fig. 3 Proposed convolutional neural network architecture with input of frame, DMOF, and DMEI (orange box: Convolution, purple box: MaxPool, pink box: Dropout, teal box: Flatten, green box: Fully connected, brown box: Concatenate)

find the best values for the weights. Also, to better converge into the optimal point, new weights (W_{update}) are proportional to the Weights of Previous Sample (*WPS*). Altogether, both *Loss function* and *WPS* are two major factors in the training phase and it is mentioned in Eq. (14).

$$W_{update} = f(Lossfunction, WPS) \tag{14}$$

It was described in previous works [55] to rewrite Eq. (14) in the subtraction form and use gradient descent as the objective function.

$$W_{update} = WPS - \eta \times gradient(Lossfunction, WPS) \tag{15}$$

In this equation, η is the learning rate of the training phase. There are many optimization algorithms for estimating the gradient descent objective function such as Nesterov Accelerated Gradient [56], Adam [57], and Adamax [57]. In this paper, we used Adamax optimization algorithm for weights updating as it is more stable and more straightforward for implementation. Weights updating process will be carried out for all data samples and several iterations. Each iteration of this process is called *Epoch*. Before the training process, the whole data are divided into train, validation, and test sets. The train set is used to update the weight of the network and the performance of the network at the end of each *Epoch* is checked by the validation set. Final evaluation will be performed by the test set. In this paper, 70%, 15%, and 15% of data are assigned to train, validation, and test sets, respectively. The minimum number of samples in the considered datasets is 6000 frames and 15% of them would be enough for a small-variance evaluation.

To better understand how the network works, it is better to visualize the output of convolutional layers. Each convolutional layer encodes independent features into the new feature maps. As can be seen in Fig. 3, these feature maps are three-dimensional cubes with width, height, and depth. Each depth channel of this cube is a two-dimensional image and it can be visualized. In Fig. 4, we chose randomly one of the depth channels in layers 1, 3, and 5 to illustrate the effects of convolutional filters on the network input. As can be seen, features in the first layers are low-level and most of the input data have been preserved. But in the last layers, features are high-level and less visually interpretable. Feature maps of the last layers are capable of classification and indicate discriminative information about fight and normal behavior classes.

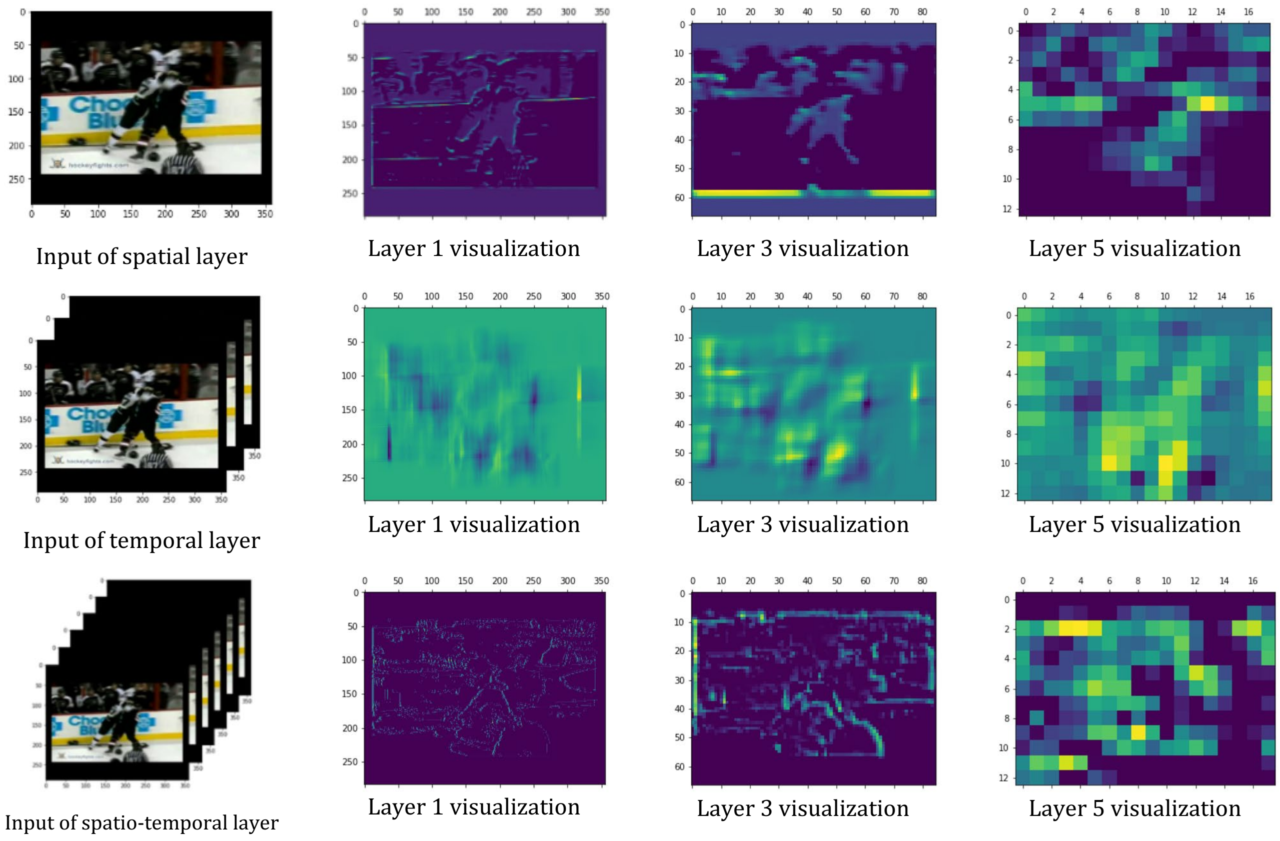
4 Experimental results

In this section, we evaluate the proposed deep network with state-of-the-art works in the field of violence detection. The experimental framework is built in a GoogleColab environment using TensorFlow 2.0 toolbox to create CNN architecture and extract deep features. To generalize the evaluation, we tested the proposed approach on both crowded and uncrowded datasets.

4.1 Datasets

We used Hockey [58] and Movie [58] datasets for uncrowded and violent flow (ViF) [37] dataset for crowded type. Figure 5 shows some examples of these datasets. The first row in this figure belongs to the Movie dataset, the second row belongs to the Hockey dataset, and the third row belongs to the ViF dataset. Also, the left-side frames show the fight class and the right-side frames show the normal class in dataset. Hockey dataset included 1000 clips of National Hockey League (NHL) games. Each clip consists of approximately 40 frames. This dataset is classified into fight and normal actions and each of them included 500 clips. The camera viewpoint is not fixed in this dataset, and there are many background motions that make it a challenging data. The frame rate and resolution of this dataset are 25 Fps and 288 × 360.

Movie dataset is composed of various normal and violent actions. Violent and normal classes have been collected from action movies and public action recognition datasets,



Input of spatial layer

Layer 1 visualization

Layer 3 visualization

Layer 5 visualization

Input of temporal layer

Layer 1 visualization

Layer 3 visualization

Layer 5 visualization

Input of spatio-temporal layer

Layer 1 visualization

Layer 3 visualization

Layer 5 visualization

Fig. 4 Feature maps of first, third, and fifth layer for spatial, temporal, and spatiotemporal streams

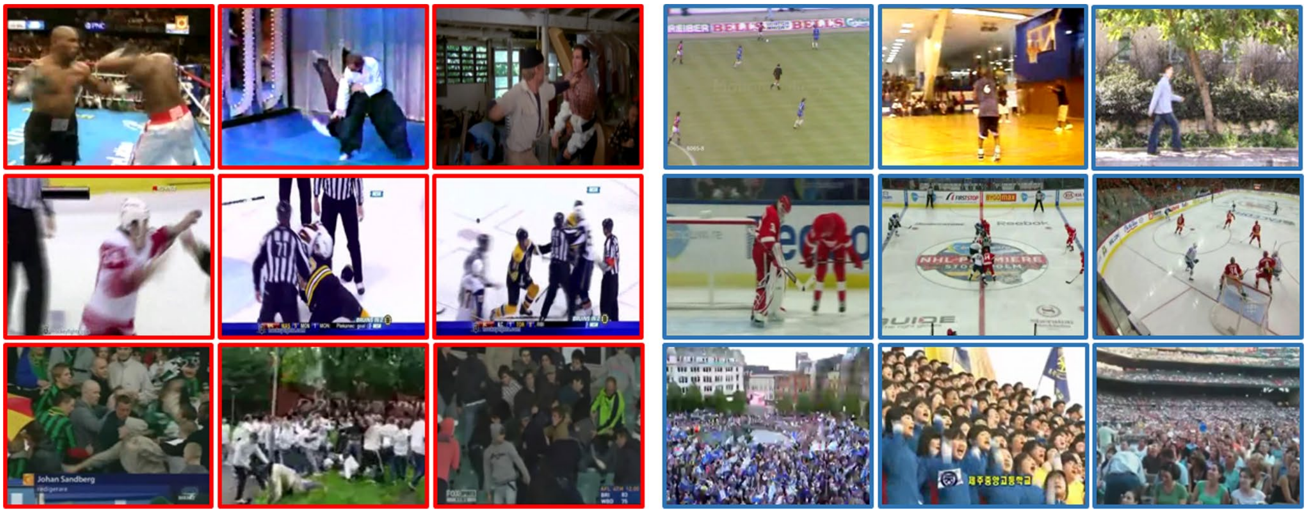


Fig. 5 Sample frames from Movie, Hockey, and ViF datasets

respectively. This dataset is generally included two- or three-person scenes, and also, there is no motion camera in them. Walking, running, and speaking are some examples of normal class actions and punching, kicking, and wrestling are some examples of violent class actions. It contains 100 clips for both classes and the frame rate and resolution are 25 Fps and 288 × 360, respectively. ViF dataset has been built based on the real-world videos from stadiums and streets. There are many crowded scenes in this dataset and it was reported in the state-of-the-art methods that it is difficult to achieve high accuracy. There are 123 clips for each fight and normal classes, and video resolution and frame rate of this dataset are 320 × 240 and 25 Fps, respectively. Table 2 shows the details of these datasets.

4.2 Evaluation metrics

To better demonstrate the performance of the proposed approach, different evaluation metrics have been utilized. They are Accuracy (ACC), Precision, Recall, and $F1$ which are defined in the following. ACC is the metric for corrected predictions of the network. This has been defined as the number of corrected predictions in both violence and normal classes out of total samples.

$$ACC = \frac{TP + TN}{TotalSamples} \tag{16}$$

TP, TN are true positive and true negative which are defined as corrected predictions in the fight and normal classes, respectively. Recall is another metric that considers only predictions in fight class. It is defined as the number of correctly classified fight samples divided by all the fight samples.

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

FN is false negative which means fight sample has been incorrectly classified to normal class. On the other hand, Precision metric is defined as the number of correctly classified fight samples divided by all predicted fight samples.

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

where FP is false positive which means normal sample has been incorrectly classified to fight class. $F1$ score is the combination of Recall and Precision metrics and is computed by the harmonic mean to simultaneously evaluate FN, FP.

$$F1 = 2\left(\frac{Precision * Recall}{Precision + Recall}\right) \tag{19}$$

Table 2 Detailed descriptions of used datasets

Dataset	Frame rate	Resolution	Clips	Frames	Actions
Hockey	25 Fps	360 × 288	1000	41,000	Violence/normal
Movies	25 Fps	360 × 250	200	6000	Violence/normal
Violent flow (ViF)	25 Fps	320 × 240	246	18,500	Violence/normal

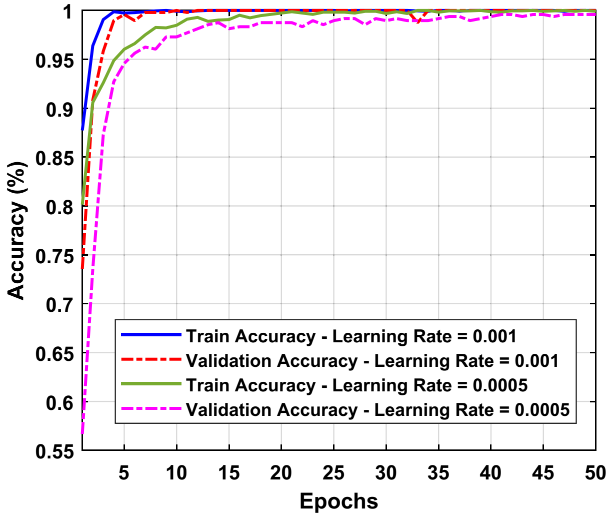


Fig. 6 Accuracy convergence for Movie dataset

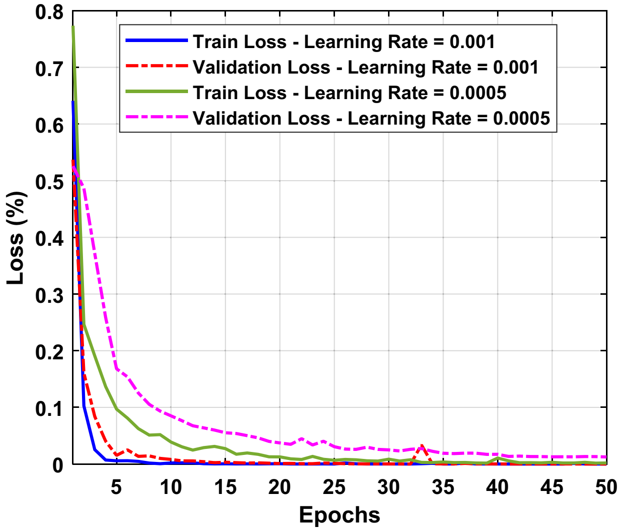


Fig. 7 Loss convergence for Movie dataset

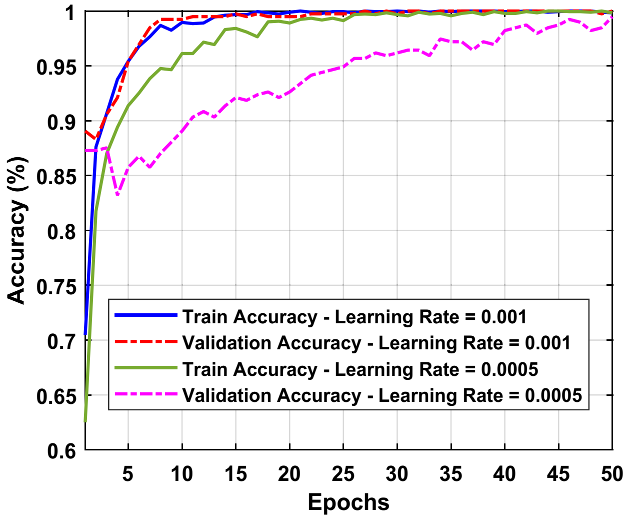


Fig. 8 Accuracy convergence for Hockey dataset

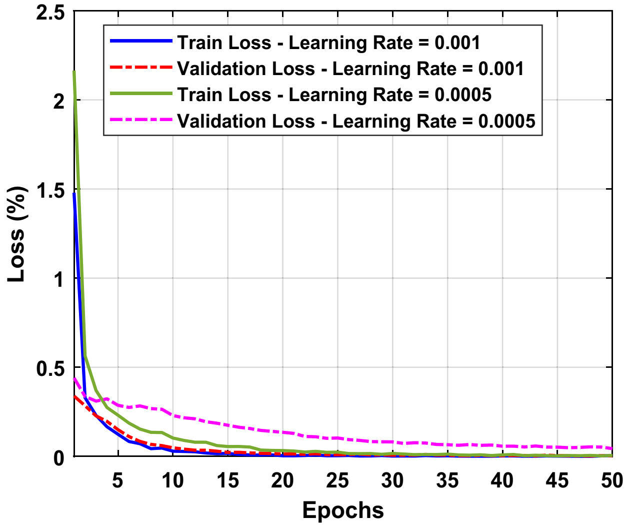


Fig. 9 Loss convergence for Hockey dataset

4.3 Accuracy and Loss Curves

As described in previous sections, weights of the network will be updated in every epoch. In each epoch, the Loss function indicates the amount of error between actual and predicted outputs. By increasing epochs, network output should be converged to the optimal point. In this network, the evaluation metrics are the accuracy and Loss of the fight and normal classes. Figures 6, 7, 8, 9, 10, and 11 show the accuracy and Loss curves for several epochs of training in all datasets. These curves are plotted for two different learning rates. As can be seen, in Hockey and Movie datasets both train and validation curves reached the maximum accuracy after about 10 epochs. Also, it is obvious that selecting 50 epochs is enough for training with no over-fitting. The results have been illustrated for 0.0005 and 0.001 learning rates. The learning rate shows the weight updating speed in the training phase, and it is required to select an appropriate value for it. For the first learning rate, the accuracy curve converges to the maximum value slowly, but after increasing it, convergence speed is fast. Also, the Loss value reached the zero point and it can be seen that based on the predicted and actual outputs, weights of the network have been updated well. The experiment showed that 0.001 learning rate is appropriate for the entire evaluation section. Hockey and Movie datasets included scenes with a few persons (two or three

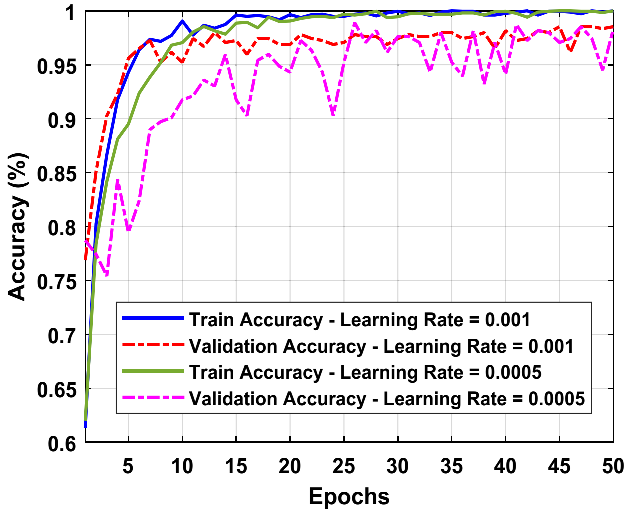


Fig. 10 Accuracy convergence for ViF dataset

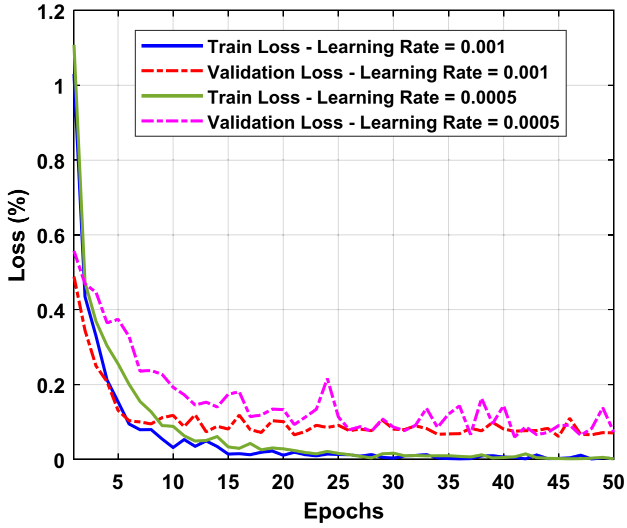


Fig. 11 Loss convergence for ViF dataset

persons), so it is necessary to evaluate the proposed approach in crowded datasets, as well. For this reason, Figs. 10 and 11 show the accuracy and Loss curves for ViF dataset. As this dataset included more complicated scenes, accuracy value is reduced slightly, nevertheless, it is acceptable and outperforms previous works. It is more obvious in this dataset that a small learning rate is not provided a stable training and more epochs are needed to reach the optimal point.

4.4 Classification results

This section demonstrates the performance of the proposed architecture including ACC, Recall, Precision, and $F1$ score results for all Hockey, ViF, and Movie datasets. These results are reported in Table 3. In Hockey and Movie datasets, Recall, Precision, and $F1$ score results reached their maximum values and ACC obtained 100%. It can be seen that all fight and normal frames have been detected correctly. In ViF dataset, there are only 18 false detected frames due to the complicated behaviors in the streets and stadiums. The superiority of the proposed framework is demonstrated in ViF results, and the achievements show that it is not dependent on the environments or camera viewpoint, and also, it can be applied to different crowded or uncrowded scenes.

4.5 Evaluation on combinational dataset

To add generality in the training phase of the proposed architecture, it is appropriate to include all variations of input data such as crowded and uncrowded regions, streets, stadiums, films, and so on. For this purpose, we combine all of Hockey, Movie, and ViF data to build a combination dataset. We trained the network with predicted and actual outputs of this dataset to test its performance on various types of inputs. Figure 12 shows the curves of Accuracy and Loss in the training phase for 50 epochs. Final validated Accuracy and Loss for this dataset are obtained 98.35% and 0.047, respectively. As can be seen the performance for the combinational dataset is reduced slightly due to the different conditions in the training dataset.

4.6 Impact of DMEI sequence length on the Accuracy

We chose DMEI feature in the spatiotemporal stream to capture long-term temporal dependency in actions. To obtain this feature, it is required to process several consecutive frames. In this section, we evaluate the sequence length of this feature and its impact on the violence detection results. Figure 13 shows the accuracy curve for different values of sequence length in three considered datasets. It can be seen that the sequence length must be selected appropriately to better represent the

Table 3 Results of Recall, Precision, $F1$ Score, and ACC metrics for Hockey, Movie, and ViF datasets

Dataset	*TP*	*TN*	*FP*	*FN*	Recall	Precision	*F*1 score	ACC (%)
Hockey	2967	3183	0	0	1	1	1	100
Movie	440	460	0	0	1	1	1	100
Violent flow (ViF)	1353	1404	14	4	0.9970	0.9897	0.9933	99.35

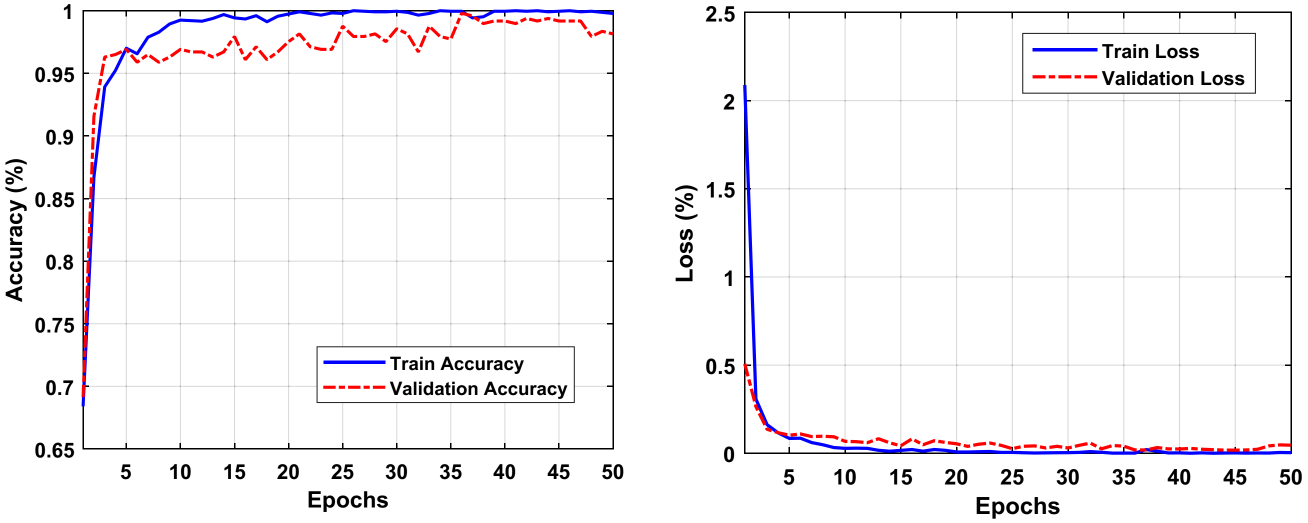


Fig. 12 Accuracy and Loss curves for combination dataset

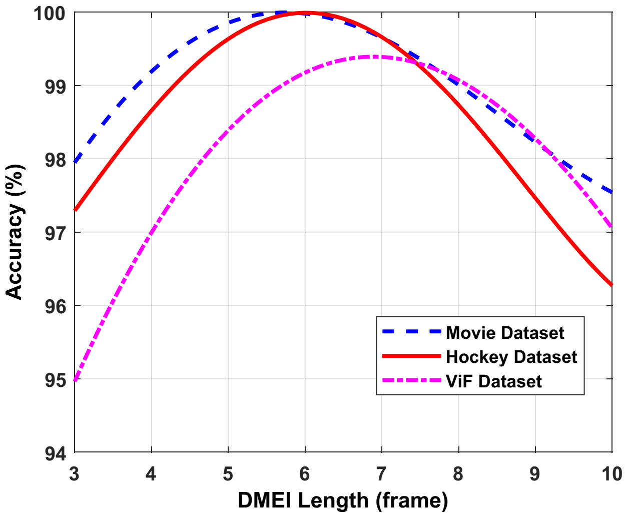


Fig. 13 Impact of DMEI sequence length on the violence detection accuracy

actions. If the sequence length is short, then it is not possible to describe the actions properly. On the other hand, if the sequence length is long, then the actions have been mixed together and there would be many overlapped regions in the final descriptor image. Finally, it can be concluded that selecting the sequence length in the range of 6–8 frames will result in the highest accuracy.

4.7 Discussion

Table 4 shows the comparison table in term of accuracy. We compared the proposed method with most of the violence detection works including handcrafted and deep learning approaches. The compared works achieved accuracy varies from 81.3 to 100% for different datasets. Interpreting these results, it can be concluded that ViF dataset is more challenging in comparison with Hockey and Movie datasets as it contains crowded environments and only some of the works could achieve accuracy higher than 95%. The violent flow [37] approach accuracy is lower than others because it is only covered the temporal aspect of violent behavior by accumulation of optical flow vectors. Also, methods in [59, 60] provided low accuracy as they only focused on improving the processing time. Zhou et al. [61] is a handcrafted method that modified the histograms of optical flow and oriented gradient and achieved better results.

Hough forest in [43] is a deep network approach that used spatiotemporal features as the input of 2D CNN and demonstrated high accuracy. [45, 49], and [50] are some deep network approaches but showed more false predictions in crowded scenes and it seems they require more discriminative features. The results in [46] showed the impact of spatiotemporal feature, and this approach could achieve an acceptable accuracy only by processing this aspect of violent behavior. Moreover, the results in [62] showed good accuracy in uncrowded scenes by developing the spatial feature in the attention region. A lower accuracy has been obtained in [63] with a multi-stream CNN, and this shows the importance of input feature extraction. The combination of CNN and LSTM which proposed in [64] showed the results around 99%; however, the computational time overhead is the bottleneck of this approach. Also, a similar work in [65] used the spatial and temporal features that have been obtained from a pre-trained VGG16 network and process them in an LSTM network.

Table 4 Accuracy comparison for violence detection works

Approach	ACC (%)			Average improvement compared to this work (%)
	Hockey dataset	Movie dataset	ViF dataset	
Hough forest and 2D CNN [43]	94.60	99	N/A	2.95
Spatial and temporal streams [62]	99.50	100	N/A	0.5
Spatial, temporal, and rhythm streams [63]	89.10	100	N/A	1.12
CNN-BiLSTM [64]	99.27	100	98.64	0.71
Multi-frame feature fusion [65]	98.80	100	97.10	1.93
Low-level features [61]	95.10	N/A	94.31	4.96
Spatiotemporal 3D CNN [46]	96	99.50	98	1.92
ConvLSTM [49]	97.10	100	94.75	2.47
C3D[45]	87.40	93.60	N/A	9.50
Violent flow [37]	82.90	N/A	81.30	17.69
Fast fight detection [59]	82.40	97.80	N/A	9.90
Fast violence detection [60]	90.10	98	N/A	5.70
Trajectory-pooled CNN [50]	98.60	N/A	92.50	4.20
This work	100	100	99.35	–

Evaluating the mentioned results, it can be seen that the proposed method achieved the highest accuracy among the mentioned works and provided average improvements about 0.5% to 17.69%. The average improvement is obtained based on the accuracy of three datasets which included both crowded and uncrowded scenes. It can be concluded that a trained network with specific features is the vital point for obtaining the maximum accuracy. This point can be seen in [62] as it is achieved high accuracy in uncrowded scenes only by covering spatial and temporal features. However, the utilized features in previous work cannot be useful in different situations. The proposed framework introduced novel preprocessing features for a customized deep network with the following aspects:

1- It considers differential values instead of absolute values
2- It covers long-term temporal dependency to properly describe the actions
3- It uses the sliding window technique to augment the processing data
4- It analyzes the sequence length to add temporal segmentation.

Table 5 Processing time of violence detection in different works

Method	Time (s)
Hough forest and 2D CNN [43]	0.46 (CPU)
Spatiotemporal 3D CNN [46]	0.11 (GPU)
convLSTM [49]	N/A
C3D [45]	0.41 (CPU)
ViF [37]	16.27 (CPU)
Spatial and temporal streams [62]	N/A
Spatial, temporal, and rhythm streams [63]	N/A
CNN-BiLSTM [64]	0.92 (GPU)
Multi-frame feature fusion [65]	N/A
Low-level features [61]	N/A
Fast fight detection [59]	0.95 (CPU)
Fast violence detection [60]	1.51 (CPU)
Trajectory-pooled CNN [50]	N/A
This work	0.49 (CPU)–0.14 (GPU)

Table 5 shows the time required to process one sample and determines whether it is violent behavior or not. This time is dependent on the implementation with CPU or GPU. We computed both and reported them separately. The processing time is composed of feature extraction time and classification time. In our method and by CPU configuration, we find that it takes 0.49 s to process one frame of a video sequence including 0.15 s for feature extraction and 0.34 s for data classification. By GPU configuration, these values are changed into 0.14 s for total processing time, 0.05 s for feature extraction, and 0.09 s for data classification.

Although the proposed architecture provides the outstanding results, there are some limitations in this work. One of them is unseen environments in the input data. As the network has been trained with labeled samples in a supervised procedure, unseen data of new environments will degrade its performance. The other one is the processing power of the host computer. To implement the proposed architecture for high frame-per-second surveillance camera systems, it is required to use hardware with a powerful GPU.

5 Conclusion

This paper introduced a novel multi-stream CNN for deep violence detection that covered most aspects of abnormal behavior between persons. We focused on the appearance, speed of movement, and representative image of actions in each stream to facilitate the interpretation of behaviors. The proposed network architecture included two handcrafted and deep learning parts which are used for feature extraction and data classification, respectively. DMOF and DMEI were two novel discriminative features that trained well the CNN network to predict all input frames of datasets. The experimental results showed that violence detection accuracy is obtained approximately 100% for both crowded and uncrowded environments. After the training phase of the network, processing time for evaluation of one sample is obtained about 0.14 s which is suitable for online violence detection in surveillance camera systems. Altogether, this approach is accurate and fast enough for being used in video sequences at the speed of about 7 Fps. In future work, we intend to propose a pre-trained network based on this approach that is capable of detecting violent behavior in every environment and condition with acceptable accuracy.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Jafri, R., Ali, S.A., Arabnia, H.R., Fatima, S.: Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. Vis. Comput. **30**, 1197–1222 (2014)
2. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. **29**, 983–1009 (2013)
3. Mitra, S., Acharya, T.: Gesture recognition: a survey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **37**, 311–324 (2007)
4. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1510–1517 (2017)
5. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Advances in Neural Information Processing Systems, pp. 64–72 (2016)
6. Tripathi, R.K., Jalal, A.S., Agrawal, S.C.: Suspicious human activity recognition: a review. Artif. Intell. Rev. **50**, 283–339 (2018)
7. Hao, T., Wu, D., Wang, Q., Sun, J.S.: Multi-view representation learning for multi-view action recognition. J. Vis. Commun. Image Represent. **48**, 453–460 (2017)
8. Zhang, Y., Dong, L., Li, S., Li, J.: Abnormal crowd behavior detection using interest points. In: International Symposium on Broadband Multimedia Systems and Broadcasting, pp. 1–4 (2014)
9. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. **36**, 18–32 (2013)
10. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1975–1981 (2010)
11. Zhang, T., Jia, W., Yang, B., Yang, J., He, X., Zheng, Z.: MoWLD: a robust motion image descriptor for violence detection. Multimed. Tools Appl. **76**, 1419–1438 (2017)
12. Berlin, S.J., John, M.: Spiking neural network based on joint entropy of optical flow features for human action recognition. Vis. Comput. 1–15 (2020).
13. Zhu, S., Hu, J., Shi, Z.: Local abnormal behavior detection based on optical flow and spatio-temporal gradient. Multimed. Tools Appl. **75**, 9445–9459 (2016)
14. Gnanavel, V.K., Srinivasan, A.: Abnormal event detection in crowded video scenes. In: Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (Ficta), pp. 441–448 (2015).
15. Mu, C., Xie, J., Yan, W., Liu, T., Li, P.: A fast recognition algorithm for suspicious behavior in high definition videos. Multimed. Syst. **22**, 275–285 (2016)
16. Nguyen, V.D., Le, M.T., Do, A.D., Duong, H.H., Thai, T.D., Tran, D.H.: An efficient camera-based surveillance for fall detection of elderly people. In: IEEE Conference on Industrial Electronics and Applications, pp. 994–997 (2014)
17. Aslan, M., Sengur, A., Xiao, Y., Wang, H., Ince, M.C., Ma, X.: Shape feature encoding via fisher vector for efficient fall detection in depth-videos. Appl. Soft Comput. **37**, 1023–1028 (2015)
18. Vishwakarma, D.K., Dhiman, C.: A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel. Vis. Comput. **35**, 1595–1613 (2019)
19. Wang, J., Xu, Z.: Crowd Anomaly Detection for Automated Video Surveillance (2015)
20. Ryoo, M.S., Rothrock, B., Fleming, C., Yang, H.J.: Privacy-preserving human activity recognition from extreme low resolution. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
21. Saravanakumar, S., Vadivel, A., Ahmed, C.S.: Multiple human object tracking using background subtraction and shadow removal techniques. In: International Conference on Signal and Image Processing, pp. 79–84 (2010)
22. Mendez, C.G.M., Mendez, S.H., Solis, A.L., Figueroa, H.V.R., Hernandez, A.M.: The effects of using a noise filter and feature selection in action recognition: an empirical study. In: International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE), pp. 43–48 (2017)
23. Dapogny, A., Bailly, K., Dubuisson, S.: Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. Int. J. Comput. Vis. **126**, 255–271 (2018)
24. Stratou, G., Ghosh, A., Debevec, P., Morency, L.P.: Effect of illumination on automatic expression recognition: a novel 3D relightable facial database. In: Face and Gesture, pp. 611–618 (2011)
25. Nazir, S., Yousaf, M.H., Nebel, J.C., Velastin, S.A.: A bag of expression framework for improved human action recognition. Pattern Recogn. Lett. **103**, 39–45 (2018)

26. Shen, M., Jiang, X., Sun, T.: Anomaly detection based on nearest neighbor search with locality-sensitive B-tree. Neurocomputing **289**, 55–67 (2018)
27. Yu, G., Goussies, N.A., Yuan, J., Liu, Z.: Fast action detection via discriminative random forest voting and top-k subvolume search. IEEE Trans. Multimed. **13**, 507–517 (2011)
28. Ehsan, T.Z., Mohtavipour, S.M.: Vi-Net: a deep violent flow network for violence detection in video sequences. In: 11th International Conference on Information and Knowledge Technology (IKT), pp. 88–92 (2020).
29. Berlin, S.J., John, M. (2020) Particle swarm optimization with deep learning for human action recognition. Multimed. Tools Appl. 1–23 (2020)
30. Wang, L., Huynh, D.Q., Koniusz, P.: A comparative review of recent kinect-based action recognition algorithms. IEEE Trans. Image Process. **29**, 15–28 (2019)
31. Jalal, A., Kamal, S., Azurdia-Meza, C.A.: Depth maps-based human segmentation and action recognition using full-body plus body color cues via recognizer engine. J. Electr. Eng. Technol. **14**, 455–461 (2019)
32. Sevilla-Lara, L., Liao, Y., Güney, F., Jampani, V., Geiger, A., Black, M.J.: On the integration of optical flow and action recognition. In: German Conference on Pattern Recognition, pp. 281–297 (2018)
33. Zin, T.T., Kurohane, J.: Visual analysis framework for two-person interaction. In: IEEE 4th Global Conference on Consumer Electronics (GCCE), pp. 519–520 (2015)
34. Chen, Y., Zhang, L., Lin, B., Xu, Y., Ren, X.: Fighting detection based on optical flow context histogram. In: Second International Conference on Innovations in Bio-inspired Computing and Applications, pp. 95–98 (2011).
35. Colque, R.V.H.M., Caetano, C., de Andrade, M.T.L., Schwartz, W.R.: Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Trans. Circuits Syst. Video Technol. **27**, 673–682 (2016)
36. Ehsan, T.Z., Nahvi, M.: Violence detection in indoor surveillance cameras using motion trajectory and differential histogram of optical flow. In: 8th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 153–158 (2018).
37. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: Real-time detection of violent crowd behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6 (2012).
38. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**, 107–123 (2005)
39. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005).
40. Dawn, D.D., Shaikh, S.H.: A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis. Comput. **32**, 289–306 (2016)
41. De Souza, F.D., Chavez, G.C., do Valle Jr, E.A., Araújo, A.D.A.: Violence detection in video using spatio-temporal features. In: 23rd SIBGRAPI Conference on Graphics, Patterns and Images, pp. 224–230 (2010).
42. Mabrouk, A.B., Zagrouba, E.: Spatio-temporal feature using optical flow based distribution for violence detection. Pattern Recogn. Lett. **92**, 62–67 (2017)
43. Serrano, I., Deniz, O., Espinosa-Aranda, J.L., Bueno, G.: Fight recognition in video using hough forests and 2D convolutional neural network. IEEE Trans. Image Process. **27**, 4787–4797 (2018)
44. Khan, S.U., Haq, I.U., Rho, S., Baik, S.W., Lee, M.Y.: Cover the violence: a novel deep-learning-based approach towards violence-detection in movies. Appl. Sci. **9**, 4963–4976 (2019)
45. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
46. Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W.: Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors **19**, 2472–2486 (2019)
47. Xia, Q., Zhang, P., Wang, J., Tian, M., Fei, C.: Real time violence detection based on deep spatio-temporal features. In: Chinese Conference on Biometric Recognition, pp. 157–165 (2018)
48. Zhou, P., Ding, Q., Luo, H., Hou, X.: Violent interaction detection in video based on deep learning. J. Phys. Conf. Ser. 844 (2017)
49. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2017).
50. Meng, Z., Yuan, J., Li, Z. (2017) Trajectory-pooled deep convolutional networks for violence detection in videos. In: International Conference on Computer Vision Systems, pp. 437–447 (2017).
51. Poynton, C.: Digital video and HD: Algorithms and Interfaces. Elsevier (2012).
52. Meinhardt-Llopis, E., Pérez, J.S., Kondermann, D.: Horn-schunck optical flow with a multi-scale strategy. Image Process. Online **3**, 151–172 (2013)
53. Horn, B.K., Schunck, B.G.: Determining optical flow. Tech. Appl. Image Underst. **281**, 319–331 (1981)
54. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 257–267 (2001)
55. François, C.: Deep Learning with Python. Manning Publications Company (2017)
56. Su, W., Boyd, S., Candes, E.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In: Advances in Neural Information Processing Systems, pp. 2510–2518 (2014).
57. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
58. Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R.: Violence detection in video using computer vision techniques. In: International Conference on Computer Analysis of Images and Patterns, pp. 332–339 (2011)
59. Serrano, G.I., Deniz, S.O., Bueno, G.G., Kim, T.K.: Fast fight detection. PLoS One, 10, e0120448 (2015)
60. Deniz, O., Serrano, I., Bueno, G., Kim, T.K.: Fast violence detection in video. In: International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 478–485 (2014)
61. Zhou, P., Ding, Q., Luo, H., Hou, X.: Violence detection in surveillance video using low-level features. PLoS One 13, e0203668 (2018)
62. Li, H., Wang, J., Han, J., Zhang, J., Yang, Y., Zhao, Y.: A novel multi-stream method for violent interaction detection using deep learning. Measurement Control **53**, 796–806 (2020)
63. Carneiro, S.A., da Silva, G.P., Guimaraes, S.J.F., Pedrini, H.: Fight detection in video sequences based on multi-stream convolutional neural networks. In: IEEE SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 8–15 (2019).

64. Halder, R., Chatterjee, R.: CNN-BiLSTM model for violence detection in smart surveillance. SN Comput. Sci. **1**, 1–9 (2020)
65. Asad, M., Yang, J., He, J., Shamsolmoali, P., He, X.: Multi-frame feature-fusion-based model for violence detection. Vis. Comput. 1–17 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mehdi Mohtavipour received the B.S. and M.Sc. in electrical engineering from University of Guilan and Iran University of Science and Technology (IUST) in 2012 and 2014, respectively. He is now a Ph.D. candidate at Iran University of Science and Technology. His research interests include intelligent transportation systems, reconfigurable computing, machine learning, and computer vision. Several papers in his research fields have been published so far.

Mahmoud Saeidi received his B.Sc. in Electrical Engineering from K. N. Toosi University, Tehran, Iran, in 2000, his M.Sc. degree in Electrical Engineering from Amirkabir University, Tehran, Iran, in 2003, and his Ph.D. degree in Electrical Engineering from K. N. Toosi University, Tehran, Iran, in 2020. He is now a faculty member in Information and Communications Technology Research Institute. His current research interests include Deep Learning and Pedestrian Detection.

Abouzar Arabsorkhi yreceived his Ph.D. Degree from the University of Tehran in the field of Information Systems Management. He is a faculty member and the head of the Network and System Security Assessment Unit at the Information and Communications Technology Research Institute. Over the past few years, he has been involved in Security Management and Planning, Security Architecture, Risk Management, Security assessment and Prototype Certification, and the Design and Implementation of Specialized Security Labs. The Internet Security of Objects is one of his main research interests. During the past 10 years, he has been teaching in the field of Information Systems and E-Commerce Security.