A novel multi-stream method for violent interaction detection using deep learning

Hongchang Li*, Jing Wang, Jianjun Han, Jinmin Zhang, Yushan Yang and Yue Zhao

Abstract
Violent interaction detection is a hot topic in computer vision. However, the recent research works on violent interaction detection mainly focus on the traditional hand-craft features, and do not make full use of the research results of deep learning in computer vision. In this paper, we propose a new robust violent interaction detection framework based on multi-stream deep learning in surveillance scene. The proposed approach enhances the recognition performance of violent action in video by fusing three different streams: attention-based spatial RGB stream, temporal stream, and local spatial stream. The attention-based spatial RGB stream learns the spatial attention regions of persons that have high probability to be action region through soft-attention mechanism. The temporal stream employs optical flow as input to extract temporal features. The local spatial stream learns spatial local features using block images as input. Experimental results demonstrate the effectiveness and reliability of the proposed method on three violent interactive datasets: hockey fights, movies, violent interaction. We also verify the proposed method on our own elevator surveillance video dataset and the performance of the proposed method is satisfied.

Keywords
Violent interaction detection, hand-craft features, deep learning

Date received: 26 November 2019; accepted: 27 December 2019

Introduction
In recent years, automatic detection and recognition of activity in real video has become increasingly important for human–computer interaction, video surveillance, and content-based video retrieval applications. Violent interaction detection is a special sub-problem in the field of action recognition. With the development of computer vision, the research on violent interaction detection (VID) has made some big progress. However, due to the low quality of surveillance video, occlusion and large changes in motion scale and illumination, VID is still a challenging task.

The VID is to identify whether a violent action has occurred in a video, such as fighting. The artificially trimmed video of the general related dataset is a fragment containing only action and is labeled as "violent action" and "non-violent action." In early time, trajectory information and orientation information of a person’s limbs are presented by Datta et al. to detect violent behavior. And some researchers attempted violent behavior detection methods on audio features. However, it is difficult to include high-quality sound in current datasets and in actual surveillance videos, so we mainly identify visual violence in video based on visual features. In fact, the previous VID method has low detection performance and high false alarm rate. If the accuracy of features extracted to represent violent action is satisfactory enough, the performance will be better. The accuracy of the spatio-temporal hand-craft features, such as motion scale invariant feature transform (MoSIFT), spatio-temporal interest points (STIP), and violent flow (ViF), in-deed increased to an extent, and the violent detection method based on these features performs better. Methods based on optical flow are also employed for detection task. However, it still cannot meet the requirement of real-time detection in different scenarios.

Recently, with the development of deep learning, action recognition has achieved rapid development. At the same time, there are many methods based on deep learning method. The authors in [10] employ the method based on deep learning to recognize violent behavior. However, the method is not applicable to the real surveillance video, because the dataset is not large enough and the training is not enough. Recently, a new method proposed by Liu et al. in [11] uses a multi-stream deep learning framework to recognize violent behavior. This method is accurate and reliable, but it is still not applicable to real surveillance video. Therefore, we propose a new robust violent interaction detection framework based on multi-stream deep learning in surveillance scene. The proposed approach enhances the recognition performance of violent action in video by fusing three different streams: attention-based spatial RGB stream, temporal stream, and local spatial stream. The attention-based spatial RGB stream learns the spatial attention regions of persons that have high probability to be action region through soft-attention mechanism. The temporal stream employs optical flow as input to extract temporal features. The local spatial stream learns spatial local features using block images as input. Experimental results demonstrate the effectiveness and reliability of the proposed method on three violent interactive datasets: hockey fights, movies, violent interaction. We also verify the proposed method on our own elevator surveillance video dataset and the performance of the proposed method is satisfied.

*Corresponding author:
Hongchang Li, Xi’an Special Equipment Inspection Institute, Xi’an 710068, China.
Email: lihongchang2018@163.com
learning in the field of VID. Meng et al.\textsuperscript{11} obtained better detection performance by fused integration trajectory and deep convolutional neural network (CNN) feature. But it did not consider the temporal feature. Ding et al.\textsuperscript{12} uses three-dimensional (3D) ConvNet to model the spatial–temporal feature, but the 3D network has large amounts of parameters and cannot use deeper networks. Multi-stream deep CNNs\textsuperscript{13} combine spatial stream, temporal stream, and acceleration stream to achieve good detection performance. However, this method only considers global spatial–temporal information, ignoring the impact of local information.

In this paper, we designed a multi-stream fusion model to address the above issues. Different from multi-stream method proposed by Dong et al.,\textsuperscript{13} in our method, the multi-stream deep learning framework mainly includes attention-based spatial RGB stream, temporal stream (by optical flow), and local spatial stream (by block-based spatial RGB). For the attention-based spatial RGB stream, we used typical soft-attention to learn attention region on the video frame; the temporal stream is similar to the optical flow mode in the temporal segment networks (TSN)\textsuperscript{14} method, which is for action recognition and the local spatial stream divide the frames into different blocks and applies CNN to learn different regional features. Finally, the results from different stream are fused to obtain the final detection results. In addition, most of the previous methods apply a variety of modes fusions, however, it is still confused which mode has the greatest impact on action recognition or VID. In this paper, we compare VID results of different mode fusion, trying to explore the impact of different mode fusion on VID and its real reasons.

In summary, the main contributions of this paper are in the following aspects: (1) A new multi-stream deep learning approach in the VID is proposed to identify whether violence action is included in the video. (2) The effects of multi-mode on the results of VID are discussed and the optimal multi-mode fusion method for VID tasks is presented (Figure 1).

The remainder of this paper is organized as follows. Section “Related works” describes the related works. Section “Multi-stream deep learning algorithm,” details the proposed multi-stream deep learning approach. Section “Experiments and analysis” reports the implementation details, the experimental results and analysis compared with the baseline and other VID methods and activity recognition methods. At the end of the paper, we conclude the proposed algorithm.

### Related works

#### Action recognition

Action recognition is a fundamental research topic in the field of video analysis and has been extensively researched in the past few years. In the past few years, action recognition algorithms mainly employ hand-crafted features, such as scale invariant feature transform (SIFT),\textsuperscript{15} histogram of oriented gradient (HOG),\textsuperscript{16} improved dense trajectories (iDT),\textsuperscript{17} and the classifier is support vector machine (SVM) or others to obtain the final prediction results. Among all the hand-crafted features, iDT achieves better performance. In recent years, deep learning methods have greatly improved the performance of action recognition. Wang et al.\textsuperscript{18} proposed the test-driven development (TDD) algorithm. By replacing the traditional features in the iDT algorithm with the features based on the deep CNN, two normalization methods were designed: spatio-temporal normalization and channel normalization. The two-stream method\textsuperscript{19,20} applies a single RGB image and flow images as input to CNN and fuse features to achieve simultaneously modeling of spatio-temporal features. TSN method\textsuperscript{14} proposes multi-segments method on the basis of two-stream to enhance the expressive ability of the model. Y Zhu et al.\textsuperscript{21} by designing the optical flow network to replace the pre-extraction of the optical flow in the video, the optical flow network is added to the optical flow branch of the two-stream model, thereby greatly accelerating the recognition speed based on achieving higher recognition accuracy. Wang et al.\textsuperscript{22} designed a non-local block to learn the global information of the original input data.

Based on C3D method,\textsuperscript{23} action recognition is directly performed by modeling spatio-temporal features of the input frame sequence and the speed of recognition is greatly improved. D Tran et al.\textsuperscript{24} compares the network structure commonly used in the action recognition task, and proves that the performance of 3D convolution is better than the two-dimensional (2D)
convolution performance. And since this year’s one-dimensional (1D) convolution in deep learning is widely used in channel changes, the authors proposed the R(2 + 1)D structure. The R(2 + 1)D structure decomposition of the 3D convolution into a 2D spatial convolution and a 1D temporal convolution, thereby increasing the number of nonlinear layers of the network, making the network easier to optimize, and good performance has been achieved on the action recognition dataset. Y Zhao et al.\textsuperscript{25} believes that although the decomposition of 3D convolution into 2D spatial convolution and 1D temporal convolution can achieve good recognition performance. However, there is an implicit assumption in 1D convolution: the characteristics of different time steps are aligned, and this assumption can facilitate the 1D convolution to aggregate the features of the same position. This assumption is very disadvantageous for the action recognition task. The author believes that motion in video is a crucial clue and is therefore subject to traditional trajectory-based methods: DT,\textsuperscript{2,26} iDT,\textsuperscript{1,17} TDD.\textsuperscript{18} This article presents TrajectoryNet. TrajectoryNet can fully consider the deformation of the motion subject, as well as the content changes caused by the continuation of the motion in the video, allowing the visual features to be aggregated along the path of motion. This method has achieved significant performance improvement on two large-scale datasets by directly replacing the 1D temporal convolutional layer in Separable-3D (S3D)\textsuperscript{27} with the trajectory convolutional layer. Efficient convolutional network for online video understanding (ECO) method\textsuperscript{28} proposes two stages for recognition: extracting 2D feature of the input frame sequence and stacking the corresponding channels of all 2D feature maps as the input of 3D CNN, and it achieves better performance under the premise of faster speed.

VID

Similar to action recognition, VID is also divided into methods using hand-crafted features and methods based on deep learning. In the method based on hand-crafted features, the iDT\textsuperscript{17} feature is a more common one. Serrano Gracia et al.\textsuperscript{29} proposed a new feature extracted from the motion trajectory between successive frames to recognition violent action. Similar to two-stream algorithm in action recognition, EY Fu and colleagues\textsuperscript{30,31} apply optical flow to obtain motion statistics. Y Gao et al.\textsuperscript{32} proposed a new feature extraction method for the dynamic amplitude change information of the VID task: oriented violent flows (OViF). At the same time, feature fusion and multi-classifiers achieved good detection performance. Wang and Schmid\textsuperscript{33} demonstrates that the Gray Level Co-Occurrence Matrix (GLCM) feature can be used to determine the density of the population, as evidenced by the fact that\textsuperscript{34} uses violent crow texture to detect violent action in the population of the input video. Zhou et al.\textsuperscript{25} segments the motion regions as the input video according to the distribution of the optical flow field, and employs Bag of words (BoW)\textsuperscript{36} to extract two low-level features of the spatial and temporal features of the segmented motion regions. The features are encoded to eliminate redundant information in the feature, and finally SVM is used to classify the final detection result. This method achieves good detection performance on Hockey dataset and the other two related datasets. With the development of hardware platforms and deep learning techniques, feature extraction methods based on deep CNN are becoming more and more popular. Inspired by TSN,\textsuperscript{14} acceleration field\textsuperscript{37} is designed, and it is merged with the original two modes of TSN. Similarly, Xia et al.\textsuperscript{38} design a new frame difference modality to complement the original RGB modal and finally fuse the results. Since this algorithm only uses RGB frames, it can improve recognition speed while ensuring recognition performance. In addition, most of the videos of related datasets are usually cropped from movie clips or web videos, and the video of the real scene is relatively small. Therefore, to cope with this problem,\textsuperscript{39} takes advantage of the inherent commonality between humans and animals and the ensemble learning mechanism to propose a new cross-species learning method, which has low computational cost features, thus effectively solving the human fight video limited issue.

Multi-stream deep learning algorithm

We propose a new VID model based on multi-stream deep learning. Our model includes the streams of attention-based spatial RGB stream, temporal stream, and local spatial stream. Finally, the results of the three streams are fused to obtain the final recognition result (Figure 2).

Attention-based spatial RGB stream

The details of this stream are shown in Figure 3. This stream first divides each trimmed video into several segments, and then randomly samples one frame from each segment, so that for each input video we can get several frames, and the spatial features of each frame are extracted using batch normalized (BN)-Inception network. Based on the extracted spatial features, we use the attention mechanism to make the stream pay more attention to the main body of the action, improving the performance of the stream.

Let $S$ denote a video. We divide video $S$ into $M$ segments with the same length, $S = \{S_1, S_2, \ldots, S_M\}$, and $S_i (i = 1, 2, \ldots, M)$ denotes the $i$th segment of video $S$. For each video $S_i$, a frame is randomly sampled, and finally $M$ frames are obtained as the input of BN-Inception network. A feature map $h_j$ of $7 \times 7$ size for each frame is extracted by BN-Inception.

After extracting feature maps for each image, soft attention is employed to capture a finer representation of the feature, as illustrated in Figure 3.
mechanism can generate weights $\alpha_i$ by applying multi-layer perceptron on the previous hidden layer $h_{t-1}$

$$\alpha_i = \frac{\exp(e_{ti})}{\sum_{k=1}^{M} \exp(e_{tk})} \quad (1)$$

where $e_{ti} = f_{att}(b_i, h_{t-1})$, $b_i$ is a series of feature representations in the current hidden layer, and $i = 1, 2, \ldots, M$.

Attention can be seen as a series of weights generated from feature maps. The region of the specific violent action on the original feature map will be learned to be closer to one. Attention mechanisms focus on the specific region of the action, and ignore the interference of the surrounding noise on the prediction to achieve the purpose of improving the prediction performance.

Accordingly, we employ long short-term memory (LSTM) network and soft attention to learn the regions, such as the human body region that makes the action. And it improves the quality of feature representation extracted by RGB frames. Sigmoid is employed as the activation function to predict the label of each frame. The prediction result of the final attention stream is obtained by combining the prediction results of LSTM cells.

**Local spatial stream**

Local information has been successful applied in person Re-ID, which also takes one person as the object of study by the model. They divide the image into blocks to get the local information of human body, and align the output results to get the final detection.
results. Inspired by this, we propose to extract human local information using block-RGB stream in VID.

Similar to attention-based spatial RGB stream, the video is also divided into segments with the same length. For each segment, the frame corresponding to attention-based spatial RGB stream is divided from top to bottom into three sub-blocks. The inputs of this stream contain a RGB frame and its three sub-blocks. The details of this stream are shown in Figure 4.

BN-Inception is used to extract the corresponding local feature representation of the three divided sub-blocks. Although each local feature representation only contains a part of the overall information, it is not affected by the context. For this reason, the divisions can learn stronger local feature representation and the proposed method enhances the local feature representation. However, it does not interact with other information in the frame to determine the specific action in the video. For the sake of interaction of different divided parts, we contact the local information obtained from these three sub-blocks, and set an fully connected (FC) layer for local information interaction. Using the block method proposed above, three sub-blocks of RGB frame are input into the block-based RGB stream, and we can get 1D features to represent them. Finally, we interact the global feature representation extracted by BN-inception from the original size frame with the fused local information representation, and obtain better information representation to improve the performance.

**Temporal stream**

The details of this stream are shown in Figure 5. Optical flow is calculated from two adjacent frames in the input video.

For video data, not only is it necessary to model spatial feature representations, but temporal features are also an important feature representation for describing the type of activity occurring in the video. Temporal features can model the temporal association of actions performed by active subjects in a video to establish a contextual relationship between frames. Therefore, our model uses attention-based spatial RGB stream, local spatial stream to establish the feature representation in each frame of the video, and uses the temporal stream to establish the temporal relationship between frames. The difference in feature representation is large, so it can be fused to each other to obtain better recognition performance similar to the flow module in TSN.14

For a given video $S$, we first employ TVL-1 algorithm\textsuperscript{44} in OpenCV\textsuperscript{45} to extract optical flow of the
video. For each pixel in optical flow graph, it can be represented by x and y axes of optical flow

\[ |V_{i,j,t}| = \sqrt{(V_{i,j,t,x})^2 + (V_{i,j,t,y})^2} \]  

(2)

where \( t \) represents the \( t \)th frame of the input video, and \( i \) and \( j \), respectively, represents the positions of the corresponding pixels in the optical flow graph. For each optical flow extracted from the video, we save both its x-axis image and y-axis image for convenience. The optical flow method describes motion of a moving target, a surface or an edge in a video, and it approximates motion information of the object in the video that cannot be directly obtained from the sequence of frame.

Similar to attention-based spatial RGB stream, in temporal stream, the video is also divided into \( M \) segments with the same length. For each segment, the optical flow is extracted on the frame corresponding to the attention-based spatial RGB stream. And each frame corresponds to two images: x-axis image and y-axis image, which represent the optical flow graph.

**Fusion**

For different streams, the model can be used to model spatial feature representations and temporal feature representations simultaneously. For each of our three streams, the probability of the category to which each input video belongs can be obtained from each stream. Therefore, for each video, the probability of three different videos belonging to the category can be obtained. To make these three streams fusion for better recognition performance, we use the last fusion method to fuse the class probabilities of the three stream outputs. So, for an input video, we have the following formula to get its prediction

\[ C_{video} = \text{maximum} \left( \sum_{c=1}^{3} \lambda_i C_{pred} \right) \]  

(3)

where \( C_{video} \) is the final prediction of the input video, \( C_{pred} \) is the class probability obtained by each stream, and \( \lambda_i \) is the weight of the category probability of the \( i \)th stream at the time of fusion.

**Loss function**

From the above three streams, the stream of attention-based spatial RGB, local spatial and temporal, we have obtained information from three modes as input to the BN-Inception network, and after the BN-Inception network, features are extracted directly to the classification layer to get the final results. Each stream applies a cross-entropy loss function and stochastic gradient descent (SGD) optimizer. This loss function is expressed as

\[ L = - \sum_{c=1}^{C} y_{i,c} \log p_{i,c} \]  

(4)

where \( C \) is the number of action classes, \( y_{i,c} \) is the true label of sample \( i \), and \( p_{i,c} \) is the probability that sample \( i \) is the category \( c \).

**Experiments and analysis**

**Datasets**

We compared our proposed algorithm with other methods on datasets that related to four violent action detections of hockey fight,\(^6\) movies,\(^6\) VID\(^5\) and our own elevator surveillance video (EsV) dataset.

**Hockey fight dataset.** The hockey fight dataset contains 1000 video clips extracted from hockey games. Among them, violence occurred in 500 videos, and another 500 is videos of no violence. Each video is about 1–2 s duration. We convert the videos into frames at 25 frames per second (FPS), and each video contains about 40 frames.

**Movies dataset.** The movies dataset contains 200 trimmed video clips. There are 100 video clips containing violent action extracted from action movies and 100 video clips extracted from other publicly available action recognition datasets that do not contain violent action. The duration of each video clip is around 2 s. We convert the videos into frames at 25 FPS, and each video contains about 40 frames.

**VID dataset.** The VID dataset is mainly extracted from the four datasets of hockey fights, movies, UCF101,\(^46\) and HMDB51.\(^47\) The video corresponding to the three categories kick, hit, punches extracted from the HMDB51 dataset, and 206 video clips are obtained. The samples corresponding to two categories, punches and Sumo Wrestling, are extracted from UCF101. A total of 271 video clips are obtained; the two datasets, hockey fight and movies, are combined. As a result, VID contains a total of 2314 videos, of which 1077 video clips contain violent action and 1237 video clips that do not contain violent action. The number of videos in VID dataset is much larger than the hockey fight and movies datasets. The larger amount of data can help the deep learning model to fit better, achieving better performance and generalization capabilities.

**EsV dataset.** Our own dataset adds 25 real elevator surveillance video clips to the VID. These new videos come from clips in our elevator surveillance videos, which are more authentic than the original video in the dataset. Examples of adding videos are shown in Figure 6. It shows two actions happening in the elevator: fighting and falling, and the former is containing violence in this paper and the other is not.
Implementation details

All models use BN-inception as the base model, our BN-inception is initialized using the model trained on ImageNet. We employ SGD method to optimize the network, and set the same values for the following parameter: weight decay 5e-4, momentum 0.9, batch size 8, learning rate 0.001. For the sake of increasing efficiency, we set the values of some parameters for different datasets. Because the training data distribution in the four datasets is different and the video content is different, the training difficulty of different datasets is different. For each of the four datasets, each of our models trained different epochs to get the best results, and it is summarized in Table 1.

Multimodal analysis

In this paper, we analyze the role of different modalities in VID tasks. It mainly includes optical flow mode (corresponding to temporal stream), warped flow mode (corresponding to warped flow in TSN),

Table 2. Exploration of different streams on dataset VID.

| Stream | Accuracy (%) |
|--------|--------------|
| RGB    | 97.89        |
| Clip   | 98.15        |
| Optical flow | 94.87 |
| Warped flow | 97.19 |
| RGB with attention | 98.15 |
| RGB & clip | 98.15 |
| RGB & Optical flow | 98.32 |
| RGB & Warped flow | 98.18 |
| Optical flow & RGB with attention | 98.35 |
| Optical flow & clip | 98.18 |
| Warped flow & RGB with attention | 98.35 |
| Warped flow & clip | 98.24 |
| Warped flow & RGB with attention & clip | 98.38 |
| RGB with attention & optical flow & clip | 98.58 |

| Stream | Accuracy (%) |
|--------|--------------|
| RGB    | 97.89        |
| Clip   | 98.15        |
| Optical flow | 94.87 |
| Warped flow | 97.19 |
| RGB with attention | 98.15 |
| RGB & clip | 98.15 |
| RGB & Optical flow | 98.32 |
| RGB & Warped flow | 98.18 |
| Optical flow & RGB with attention | 98.35 |
| Optical flow & clip | 98.18 |
| Warped flow & RGB with attention | 98.35 |
| Warped flow & clip | 98.24 |
| Warped flow & RGB with attention & clip | 98.38 |
| RGB with attention & optical flow & clip | 98.58 |

Symbol "&" represents the fusion of some modes. VID: violent interaction.

Table 1. The number of epochs trained for each stream in each of the four datasets.

| Stream                  | Hockey fight | Movies | VD | EsV |
|-------------------------|--------------|--------|----|-----|
| Attention-based spatial stream | 1000         | 60     | 400 | 400 |
| Block-based stream      | 55           | 20     | 30  | 100 |
| Optical flow stream     | 210          | 50     | 50  | 50  |

VID: violent interaction detection; EsV: elevator surveillance video.

spatial feature and the modality of the modeling temporal feature. As can be seen from Table 2, we can achieve good performance with a single RGB with attention stream. When fusing two modes, attention-based spatial RGB mode achieves the best performance when fusion with optical flow or warped flow. Attention-based spatial RGB mode employs attention mechanism and can obtain a better representation of the spatial features, so that the temporal feature representations of the modes associated with the flow can complement each other for better detection performance. Clip mode cannot be combined with optical flow like RGB with attention stream to get a particularly good performance boost. The general reason we guess this phenomenon is that the clip mode does not learn more spatial information than the optical flow difference compared with the RGB with attention stream. However, the clip mode is better able to learn local information. At the same time, for three-modal fusion, the optimal mode can be obtained using optical flow, attention-based spatial RGB, and block-based RGB. Although warped flow mode can achieve better performance than optical flow mode in single mode, it does not perform as well as optical flow mode in fusion.

Results and analysis

For hockey fight, as shown in Table 3, our model adds new modalities and multi-modal fusion. Not only does
the performance of multi-modal fusion exceeds TSN\textsuperscript{14} and FightNet,\textsuperscript{37} but the performance of each individual modality also exceeds these two.

The single-mode performance of the model is shown in Table 3. Through the last fusion, the local information of the attention-based spatial RGB stream and the local spatial stream can be better complemented by the global information of the optical flow stream. We think this is the main source of recognition performance improvement. Because of the number of videos in the movies dataset, our model is able to achieve 100% prediction accuracy on all modalities.

As shown in Table 4, the modalities of our model show good performance on large datasets such as VID. Not only the model, after last fusion, exceeds the FightNet, but also exceeds TSN. Our model has shown the best detection performance on the general three datasets, which can be verified that our proposed multi-stream fusion model has good detection performance and good generalization ability.

In Table 5, we list the results of each modality and various modal fusions on our dataset. Since the local features of the input video sequence are better learned, and the local spatial stream itself has a certain global feature learning capability, this stream can simultaneously capture local features as well as global features, so local spatial stream get the best recognition performance in a single model. Although the warped flow stream can obtain better results than the optical flow stream when the two modes are merged with each other, the optical flow stream can be better merged with more streams. Therefore, the new three-modal fusion model proposed by us has 98.71% performance higher than the 98.13% of the TSN model.

It can be seen from the experimental results of the above four datasets that the proposed multi-stream model has a larger performance improvement on the four datasets than the other VID models. We think the possible reasons are: first, our attention based with RGB spatial and local spatial stream effectively model the spatial features of the input video, so that a single stream can achieve better detection performance. At the same time, to verify the effectiveness of the attention mechanism we used, we visualized some of the attention results in the test video on the hockey dataset and VID dataset. As can be seen from Figure 7, the whiter regions in the captured visualization indicate that our model has a higher degree of interest in this region of the image. By comparing the original image and the attention image in Figure 7, the following conclusions can be obtained: the attention mechanism we use can effectively capture the body of the behavior object in the video to help the model improve detection performance. Second, the temporal stream we use can effectively model the temporal information in the video data to perform the VID task based on the temporal information. Third, for the spatial features and temporal features modeled by the three streams, we used last fusion to make our model the spatio-temporal features at the same time, thus improving the performance of the detection.

### Conclusion

In this paper, we propose a new multi-stream deep learning approach for VID. The attention-based spatial RGB stream learns the attention regions through attention mechanism, improving the performance of the global spatial information. The local spatial stream can better learn the local features by simply dividing the frame. The temporal stream achieves a temporal performance.
feature representation of the video by employing optical flow method. Through the fusion of the above three streams, we have reached state-of-the-art on hockey fight, movies, VID, and our EsV dataset. Through analyzing the influence of different modalities on detection performance, we find the optimal combination between different modalities. The proposed algorithm is not only limited in the application of VID, it can also be applied to other applications of activity recognition.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**ORCID iD**

Hongchang Li https://orcid.org/0000-0002-5865-4715

**References**

1. Datta A, Shah M, Da N, et al. Person-on-person violence detection in video data. *Int Conf Pattern Recogn* 2002; 1: 433–438.
2. Nam J, Alghoniemy M and Tewfik AH. Audio-visual content-based violent scene characterization. *Int Conf Image Pr* 1998; 1: 353–357.
3. Cristani M, Bicego M and Murino V. Audio-visual event recognition in surveillance video sequences. *IEEE T Multimedia* 2007; 9(2): 257–267.

4. Lin J and Wang W. Weakly-supervised violence detection in movies and audio video based co-training. In: Muneeesawang P, Wu F, Kumazawa I, et al. (eds) *Advances in Multimedia Information Processing—PCM 2009* (PCM 2009. Lecture Notes in Computer Science), vol 5879. Berlin: Springer, 2009, pp. 930–935.

5. Nievas EB, Suarez OD, Garcia GB, et al. Violence detection in video using computer vision techniques. In: *Proceedings of the international conference on computer analysis of images and patterns*, Seville, 29–31 August 2011, pp. 332–339. New York: Springer.

6. Nievas EB, Suarez OD, Garcia GB, et al. Violence detection in video using computer vision techniques. In: Reall P, Diaz-Permi D, Molina-Abril H, et al. (eds) *Computer Analysis of Images and Patterns* (CAIP 2011 Lecture Notes in Computer Science), vol 6855. Berlin: Springer, 2011, pp. 332–339.

7. Hassner T, Itcher Y and Kliper-Gross O. Violent flows: real-time detection of violent crowd behavior. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops*, Providence, RI, 16–21 June 2012. New York: IEEE.

8. De Souza FDM, Ch’avez GC, Do Valle EA, et al. Violence detection in video using spatio-temporal features. In: *Proceedings of the 23rd SIGRAPI conference on graphics, patterns and images*, Gramado, 30 August–3 September 2010, pp. 224–230. New York: IEEE.

9. Mabrouk AB and Zagrouba E. Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recogn Lett* 2017; 92(C): 62–67.

10. Zhang T, Yang Z, Jia W, et al. A new method for violence detection in surveillance scenes. *Multimedia Tool Appl* 2016; 75(12): 7327–7349.

11. Meng Z, Yuan J and Li Z. Trajectory-pooled deep convolutional networks for violence detection in videos. In: Liu M, Chen H and Vincze M. (eds) *Computer vision systems* (ICVS 2017. Lecture Notes in Computer Science), vol 10528. Cham: Springer, 2017, pp. 437–447.

12. Ding C, Fan S, Zha M, et al. Violence detection in video by using 3D convolutional neural networks. In: Bebis G., et al. (eds) *Advances in Visual Computing*, vol 10528. Cham: Springer, 2017, pp. 437–447.

13. Dong Z, Qin J and Wang Y. *Multi-stream deep networks for person to person violence detection in videos*. Singapore: Springer, 2016.

14. Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition, 2016. arXiv preprint, arXiv:1608.00595v1.

15. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 2004; 60(2): 91–110.

16. Dalal N and Triggs B. Histograms of oriented gradients for human detection. In: *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition* (CVPR’05), San Diego, CA, 20–25 June 2005, pp. 886–893. New York: IEEE.

17. Wang H and Schmid C. Action recognition with improved trajectories. In: *Proceedings of the 2013 IEEE international conference on computer vision*, ICCV ’13, Sydney, NSW, Australia, 1–8 December 2013, pp. 3551–3558. New York: IEEE.

18. Wang L, Yu Q and Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors, 2015. arXiv preprint, arXiv:1505.04868v1.

19. Simonyan K and Zisserman A. Two-stream convolutional networks for action recognition in videos. In: *Proceedings of the 27th international conference on neural information processing systems (NIPS’14)*, Montreal, QC, Canada, 8–13 December 2014, pp. 568–576. Cambridge, MA: MIT Press.

20. Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets, 2015, https://arxiv.org/abs/1507.02159.

21. Zhu Y, Lan Z, Newsam S, et al. Hidden two-stream convolutional networks for action recognition, 2017. arXiv preprint, arXiv:1704.00389v4.

22. Wang X, Girshick R, Gupta A, et al. Non-local neural networks, 2018. arXiv preprint, arXiv:1711.07971v3.

23. Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: *Proceedings of the 2015 IEEE international conference on computer vision (ICCV ’15)*, Venice, 22–29 October 2015, pp. 4489–4497. New York: IEEE.

24. Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition, 2017. arXiv preprint, arXiv:1711.11248v3.

25. Ziao Y, Xiong Y and Lin D. Trajectory convolution for action recognition. In: Bengio S, Wallach H, Larochelle H, et al. (eds) *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, 2018, pp. 2208–2219.

26. Wang H, Klasar A, Schmid C, et al. Action recognition by dense trajectories. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’11*, 2011, pp. 3169–3176. New York: IEEE.

27. Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning: speed-accuracy tradeoffs in video classification. In: *Proceedings of the 15th European conference*, Munich, 8–14 September 2018, Part XV, pp. 318–335.

28. Zollaghari M, Singh K and Brox T. ECO: efficient convolutional network for online video understanding, 2018. arXiv preprint, arXiv:1804.09066v2.

29. Gracia IS, Suarez OD, Garcia GB, et al. Fast fight detection. *PLoS ONE* 2015; 10: e0120448.

30. Fu EY, Leong HV, Ngai G, et al. Automatic fight detection based on motion analysis. In: *Proceedings of the 2015 IEEE international symposium on multimedia (ISM)*, Miami, FL, 14–16 December 2015, pp. 57–60. New York: IEEE.

31. Fu EY, Leong HV, Ngai G, et al. Automatic fight detection in surveillance videos. In: *Proceedings of the 14th international conference on advances in mobile computing and multi media*, Momm ’16, Singapore, 28–30 November 2016, pp. 225–234. New York: ACM.

32. Gao Y, Liu H, Sun X, et al. Violence detection using oriented violent flows. *Image Vision Comput* 2016; 48(C): 37–41.

33. Wang H and Schmid C. Action recognition with improved trajectories. In: *Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV ’13*, Santiago, Chile, 7–13 December 2013, pp. 3551–3558. New York: IEEE.

34. Lloyd K, Marshall D, Moore S, et al. Detecting violent crowds using temporal analysis of GLCM texture, 2016.
35. Zhou P, Ding Q, Luo H, et al. Violence detection in surveillance video using low-level features. PLoS ONE 2018; 13: e0203668.

36. Li FF and Perona P. A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). San Diego, CA, 20–25 June 2005, pp. 524–531. New York: IEEE.

37. Zhou P, Ding Q, Luo H, et al. Violent interaction detection in video based on deep learning. J Phys 2017; 844(1): 012044.

38. Xia Q, Zhang P, Wang J, et al. Real time violence detection based on deep spatio-temporal features. In: Zhou J., et al. (eds) Biometric recognition (CCBR 2018. Lecture Notes in Computer Science), vol. 10996. Cham: Springer, 2018, pp. 157–165.

39. Fu EY, Huang MX, Leong HV, et al. Cross-species learning: a low-cost approach to learning human fight from animal fight. In: Proceedings of the 26th ACM international conference on multimedia, MM ’18, 22–26 October 2018, pp. 320–327. New York: ACM.

40. Sak H, Senior AW and Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, 2014. arXiv preprint, arXiv:1402.1128v1.

41. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32Nd International Conference on Machine Learning, Volume 37 of Proceedings of Machine Learning Research (eds F Bach and D Blei), Lille, 7–9 July 2015, pp. 2048–2057.

42. Sun Y, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). New York: Springer, 2018.

43. Wei L, Zhang S, Yao H, et al. Glad: global-local-alignment descriptor for pedestrian retrieval, 2017. arXiv preprint, arXiv:1709.04329v1.

44. Zach C, Pock T and Bischof H. A duality based approach for realtime tv-l1 optical flow. In: Hamprecht FA, Schnörr C and Jähne B (eds) Pattern Recognition (DAGM 2007. Lecture Notes in Computer Science), vol 4713. Berlin: Springer, 2007, 214–223.

45. Bradski G. The OpenCV Library. Dr. Dobb’s Journal of Software Tools, 2000, https://www.drdobbs.com/open-source/the-opencv-library/184404319.

46. Soomro K, Zamir AR and Shah M. Ucf101: a dataset of 101 human actions classes from videos in the wild, 2012. arXiv preprint, arXiv:1212.0402v1.

47. Kuehne H, Jhuang H, Garrote E, et al. Hmdb: a large video database for human motion recognition. In: Proceedings of the 2011 International Conference on Computer Vision, ICCV ’11, Barcelona, 6–13 November 2011, pp. 2556–2563. New York: IEEE.

48. D’eniz-Su’arez O, Serrano I, García GB, et al. Fast violence detection in video. In: Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, 5–8 January 2014, pp. 478–485. New York: IEEE.