IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON RECENT ADVANCES IN VIDEO CODING AND SECURITY

Received February 11, 2019, accepted March 9, 2019, date of publication March 19, 2019, date of current version April 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2906275

A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks

WEI SONG[1,2], DONGLIANG ZHANG[1], XIAOBING ZHAO[1,2], JING YU[3], RUI ZHENG[1], AND ANTAI WANG[4]

[1]School of Information Engineering, Minzu University of China, Beijing 100081, China
[2]National Language Resource Monitoring and Research Center of Minority Languages, Minzu University of China, Beijing 100081, China
[3]School of Electronic Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[4]Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102, USA

Corresponding author: Wei Song (songwei@muc.edu.cn)

This work was supported in part by the National Science Foundation Project of China under Grant 61503424, Grant 61331013, and Grant 61701554, in part by the Promotion Plan for Young Teachers' Scientific Research Ability of the Minzu University of China, in part by the MUC 111 Project, in part by the First Class University and First Class Discipline of the Minzu University of China (Intelligent Computing and Network Security), and in part by the Youth Team Leadership Program.

ABSTRACT Violent video constitutes a threat to public security, and effective detection algorithms are in urgent need. In order to improve the detection accuracy of 3D convolutional neural networks (3D ConvNet), a novel violent video detection scheme based on the modified 3D ConvNet is proposed. In this paper, the preprocessing method of data is improved, and a new sampling method by using the key frame as dividing nodes is designed. Then, a random sampling method is adapted to produce the input frame sequence. With experimental evaluations on the crowd violence dataset, the results demonstrate the effectiveness of the proposed new sampling method. For three public violent detection datasets: hockey fight, movies, and crowd violence, individualized strategies are implemented to suit the varied clip length. For the short clips, the 3D ConvNet is constructed by using the uniform sampling method. For the longer clips, the new frame sampling strategy is adopted. The proposed scheme obtains competitive results: 99.62% on hockey fight, 99.97% on movies, and 94.3% on crowd violence. The experimental results show that our method is simple and effective.

INDEX TERMS Violent video detection, 3D ConvNet, key frame extraction.

I. INTRODUCTION

Video content makes up more and more proportion of the world's Internet traffic at present. Video service represented by short video and live streams becomes the new trends of the development of the Internet. However, Internet video content is filled with some violent videos, which are seriously harm the construction of the network ecology. Furthermore, monitoring sudden violence in time creates tremendous challenges for video surveillance. Thus, violent video detection is of vital importance.

Violent video detection generally refers to the detection of violence and violent scene in video, by using feature extraction algorithms to get the visual and auditory features, and then by virtue of linear classifier to classify it. To some extent, violent video detection can be regarded as the special case of video action recognition.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoqing Pan.

A video sequence is a series of still images shown in rapid succession to give the impression of continuous motion. Most frames are highly correlated with their neighbors. Considering the temporal dimension of the video, there was an increasing research focus on how to extract motion information among the adjacent frames. For traditional methods, spatio-temporal interest points (STIPs) [1] and improved Dense Trajectories (iDT) [2] are commonly used. For deep learning methods, corresponding improvement in two directions, two-stream structure based on 2D ConvNet and extended structure based on 3D ConvNet. Two-stream structure [3] uses temporal stream ConvNet to extract motion information, and applies spatial steam ConvNet to extract appearance information. Extended architecture based on 3D ConvNet takes consecutive frames as input; convolution and pooling operation are performed spatio-temporally. By using 3D convolution and 3D pooling, 3D ConvNet can capture appearance, as well as short-term motion. C3D [4] is a typical 3D ConvNet, successive clips of 16 frames are used as the input and the

frame sequence is uniformly sampled, which destroys the integrity of motion information in some degree. Meanwhile, it affects the overall performance as well.

In this paper, a modified 3D ConvNet framework is proposed to detect the violent video content. The new modified scheme improves the preprocessing method of 3D ConvNet. It attempts to cut video sequence into clips based on key frames. This method can reduce redundancy caused by uniform sampling to some extent, and it can decrease the destruction of motion integrity by the uniform sampling.

The paper is organized as follows. Section 2 presents the related work. Section 3 presents the modified 3D ConvNet framework and a new sampling method. Section 4 reports the performance of the proposed framework on public benchmarks. Finally, Section 5 concludes the paper and discusses possible future work.

II. RELATED WORK

In this section, related works of violent video detection are divided into two categories: traditional methods and deep learning methods.

A. TRADITIONAL METHODS

Traditional methods mainly based on the multimodal classification strategy. Methods based on audio features and methods based on multimodal audio-visual features are the two main ones.

Methods based on audio features: Pfeiffer *et al.* [5] proposed an algorithm for violence detection using audio tracks. Cheng *et al.* [6] proposed a hierarchical approach based on Gaussian mixture models and Hidden Markov models (HMM) [7] to recognize audio events. Giannakopoulos *et al.* [8] used six segment-level audio features for detecting violence in audio segments with an SVM classifier. Clarin *et al.* [9] designed an automated system consisting of 4 modules, and this method mainly used Kohonen's Self-Organizing Map to recognize skin and blood colors and motion intensity to detect violent actions involving blood.

Methods based on multimodal audio-visual features: Nam *et al.* [10] presented a scheme to recognize violent scenes in movies using audio-visual features, and this is one of the first proposals for violence recognition in video. Gong *et al.* [11] proposed a three-stage approach to detect violent scenes in movies, integrating low-level visual and auditory features and high-level audio effects related to violence. Lin and Wang [12] present a violent shot detection scheme, using a co-training method, combined a weakly-supervised audio classifier with a motion, explosion and blood video classifier. Giannakopoulos *et al.* [13] proposed a multi-modal violence detecting method in movies that combined audio features and video features using a k-Nearest Neighbor classifier.

Most of the above research focus on detecting violent content based on audio features and bloody color features. These features are very effective in the detection of violent content in movies. However, in the real world of video surveillance, audio and bloody scenes are rarely recorded. Therefore, a majority of later researches put emphasis on visual features. Recently, Datta *et al.* [14] used motion trajectory information and orientation information of a person's limbs to detect human violence in Video, such as fist fighting, kicking, hitting with objects, etc. Hassner *et al.* [15] designed a ViF (Violent Flow) descriptor to real-time detection of breaking violence in crowded scenes. Deniz *et al.* [16] proposed a method which uses extreme acceleration patterns as the main feature to the task of violence detection. Nievas *et al.* [17] used STIPs and MoSIFT [18] descriptors to assess the performance in the fight detection problem. Xu *et al.* [19] adopted sparse coding scheme to replace Bag-of-Words model to further improve the accurate of MoSIFT. Rota *et al.* [20] proposed a method to detect and localize dyadic human interactions in real videos. And this method used dense trajectories to capture both shape and motion features, and it defined an interpersonal area between the interacting subjects. Mironică *et al.* [31] created a new content representation pipeline for video classification, and it used Random Forest and a modified Vector of Locally Aggregated Descriptor (VLAD) with Fisher Kernel representation. Their experimental result showed that it got good performance on VSD2013 dataset [32]. Zhang *et al.* [33] proposed a robust violence detection (RVD) method, and got effective results on several benchmark datasets in terms of both detection accuracy and processing speed, even in crowded scenes. Zhang *et al.* [34] proposed a Motion Weber Local Descriptor (MoWLD), integrating the sparse coding method, and the experimental results demonstrated that the proposed method is effective for violence detection. Some other methods [39], [41], such as saliency detection [37], Low Rank Representation [38] are also used in the classification or the content analysis.

B. DEEP LEARNING METHODS

With the rise of deep learning, related research in video analysis domain boomed [36], [40], especially in human action recognition. Whereas related works in violent video detection are rarely to be published. Out of consideration for the importance of realistic security and the great breakthrough deep learning methods have made on visual recognition [21], [22], researchers come to employ deep learning methods to detect violence in video.

Lam *et al.* [35] evaluated the use of multiple features and their combination in a violent scenes detection system, providing an empirical foundation for selecting capable feature sets to deal with heterogeneous content comprising violent scenes in movies. Ding *et al.* [23] presented a 9 layers 3D-CNN for violent video detection, and get 91% on the Hockey fight dataset. However, their work uses 3D convolution, but adopts 2D pooling, result in losing temporal information of the input signals. Dai *et al.* [24] proposed a violent scene detecting method, and it concatenated two streams ConvNet to long short-term memory (LSTM) [25], and finally used SVM classifier to classify. Experimental

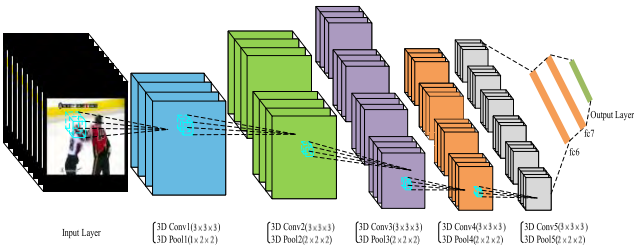

FIGURE 1. Adopted 3D convNet structure.

result shows that their method is superior to traditional methods. Zhou *et al.* [26] constructed a FightNet based on temporal segment networks (TSN) [27] to detect violent interaction. In addition, the authors collected a violent interaction dataset (VID) to train a pre-trained model, and then use it to extract violence features, get 97% on the Hockey fight dataset.

Although the mentioned traditional methods and deep learning methods showed good results on violent video detection, it is not discriminative enough. For this reason, this paper construct a modified 3D ConvNet, using 3D convolution and 3D pooling, exploring better methods for characterizing violence features. The distinguished contribution of this paper is adopting a new sampling method, protecting temporal information to a degree.

III. PROPOSED METHOD

A. 3D ConvNet

The proposed network is based on C3D, which is introduced by Tran *et al.* [4], which can learn spatiotemporal features of video. By using 3D convolution and 3D pooling, the temporal information of the input video is well preserved. The input of the network is successive clips with a length of 16 frames, sampled according uniform sampling method. As shown in Fig.1, the 3D ConvNet structure used in our implementation is composed of 5 convolution layers, 5 pooling layers, followed by 2 fully-connected layers, and a softmax layer.

1) NETWORK PARAMETERS

The main advantage of 3D ConvNet is 3D convolution and 3D pooling. As suggested in [4], We define the input clips with a size of $c \times l \times h \times w$, where c is the number of channels, l is the length of clip, h and w are the height and width of the frame, respectively. In this paper, the input size is $3 \times 16n \times 128 \times 171$, where n is 1 or 2. The kernel size of 3D convolution is set as $3 \times 3 \times 3$. The numbers of kernels for 5 convolution layers from layer 1 to layer 5 are 64, 128, 256, 256, and 256, respectively. All pooling layers are max pooling with kernel size $2 \times 2 \times 2$ (except first pooling layer, kernel size is $1 \times 2 \times 2$, in consideration of protecting the temporal information). Each fully connected layer of the net has 2048 outputs.

B. NEW SAMPLING METHOD

In this paper, the most notable feature of the proposed scheme is that the preprocessing method of 3D ConvNet has been improved. In view of the lack of relevant research on the improvement of preprocessing method and the importance of it, this method is proposed. Typically, uniform sampling method will sample every frame, or it will sample frames by fixed interval. For 3D ConvNet, it employs original uniform sampling method divides videos into video chunks of 16-consecutive frame. If sample videos are shorter, uniform sampling method is simple and effective. However, for longer videos, the fixed sample method brings the problem of redundancy and the discontinuity of motions.

The hierarchical structure of the video contains scenes, shots, and frames. Frames in the same shot have great visual similarity. Key frames are usually well chosen samples that best represent the content of the shot. Therefore, we extract key frames from the video and then divide up video based on the extracted key frames.

By comparing the effectiveness of different key frame extraction algorithms on experimental results, we adopt the gray centroid algorithm proposed in [28]. Based on the priori knowledge of the similarity of frames in the same shot, the gray centroid algorithm is adopted to evaluate the similarity between adjacent frames with gray-center position changes.

The adopted key frame extraction algorithm can be described as follow:

Firstly, convert RGB frames to grayscale, and calculate the gray centroid as:

$$\begin{aligned} x_c &= \sum_{i=1}^{h}\sum_{j=1}^{w} I(i,j) \times j \sum_{i=1}^{h}\sum_{j=1}^{w} I(i,j) \\ y_c &= \sum_{i=1}^{h}\sum_{j=1}^{w} I(i,j) \times j \sum_{i=1}^{h}\sum_{j=1}^{w} I(i,j) \end{aligned} \tag{1}$$

where, $\boldsymbol{I}$ stands for video frame image. i, j represents the row and the column of the frame, respectively. The similarity between adjacent frames can be described by the change of centroid position. For the gray frame images, $\boldsymbol{I}^{(k)}$ and $\boldsymbol{I}^{(k+1)}$ correspond to the adjacent frames, $(x_c^{(k)}, y_c^{(k)})$ and $(x_c^{(k+1)}, y_c^{(k+1)})$ are gray centroids. Define $L(I^{(k)}, I^{(k+1)})$ as centroid distance, and it can be calculated as:

$$L(\boldsymbol{I}^{(k)}, \boldsymbol{I}^{(k+1)}) = \sqrt{(x_c^{(k)} - x_c^{(k+1)})^2 + (y_c^{(k)} - y_c^{(k+1)})^2} \tag{2}$$

We define relative distance as the similarity measurement of adjacent frames:

$$\begin{aligned} D(\boldsymbol{I}^{(k)}, \boldsymbol{I}^{(k+1)}) &= L(\boldsymbol{I}^{(k)}, \boldsymbol{I}^{(k+1)}) \left\| (x_c^{(k+1)}, y_c^{(k+1)}) \right\| \\ &= L(\boldsymbol{I}^{(k)}, \boldsymbol{I}^{(k+1)}) \sqrt{(x_c^{(k+1)})^2 + (y_c^{(k+1)})^2} \end{aligned} \tag{3}$$

Then begin with the second frame, we can distinguish whether frames are visually similar or not. Minimal number of frames that have visual similarity is set to m. Let $(\boldsymbol{I}^{(1)}, \cdots, \boldsymbol{I}^{(k)})$ be successive frames that have visual similarity, where k is the amount of frames. Assume that $(x_c^{(i)}, y_c^{(i)})$ is centroid of i^{th} frame and (x_c^*, y_c^*) is the average centroid

of k successive frames that satisfies the minimum variance condition. The equation for calculating (x_c^*, y_c^*) as follows:

$$(x_c^*, y_c^*) = \arg\min_{(x,y)} \sum_{i=1}^{k} ((x_c^{(i)} - x)^2 + (y_c^{(i)} - y)^2)$$
$$s.t.\ 1 \le x \le w, 1 \le y \le h \tag{4}$$

where, w and h are the width and height of the frame, respectively. Therefore, the solution of the above problem can be calculated as:

$$\begin{cases} x_c^* = \sum_{i=1}^{k} x_c^{(i)}/k \\ y_c^* = \sum_{i=1}^{k} y_c^{(i)}/k \end{cases} \tag{5}$$

Define $\varepsilon_1 > 1\%$ as the threshold of inter-frame similarity, $|x_c^{(k)} - x_c^{(k+1)}|/x_c^{(k+1)} < \epsilon_1$ and $|y_c^{(k)} - y_c^{(k+1)}|/y_c^{(k+1)} < \epsilon_1$ as decision conditions of whether two adjacent frames have visual similarity. It can be easily proved that $|x_c^{(k)} - x_c^{(k+1)}|/x_c^{(k+1)} < \epsilon_1$ and $|y_c^{(k)} - y_c^{(k+1)}|/y_c^{(k+1)} < \epsilon_1$ are sufficient conditions for $D(\boldsymbol{I}^{(k)}, \boldsymbol{I}^{(k+1)}) < \varepsilon_1$. Define $\varepsilon_2 \le 3\%$ as the threshold of the cumulative change of centroid, then $|x_c^{(k)} - x_c^{(k+1)}|/x_c^{(k+1)} < \epsilon_2$ and $|y_c^{(k)} - y_c^{(k+1)}|/y_c^{(k+1)} < \epsilon_2$ need to be satisfied simultaneously, the current frame and previous frames can be considered as belonging to the same successive frames that have visual similarity. After that, we filter the frame with the smallest distance between gray centroid and average gray centroid as the key frame of the sequence. Let $V_j$be the j^{th} frames that have visual similarity, let *keyframe*(j) be the key frame filtered from V_j, and it can be calculated as:

$$keyframe(j) = \arg\min_{i \in V_j}((x_c^{(i)} - x_c^*)^2 + (y_c^{(i)} - y_c^*)^2) \tag{6}$$

Finally, the full set of key frames is {$keyframe_1$, $keyframe_2$, …, $keyframe_m$}. After key frame extraction, based on the number of key frames in the video clips and the key frame spacing, we design the rules of frame sequence segmentation in this paper. We adopt 16 frames as the clip length, and we adopt random sampling between key frames. Concrete division rules are as Fig.2.

Given one video clip which has frames, the random sampling method we adopted is as follows:

$$fm_i = \frac{N}{S}(i + \frac{j}{2}) \tag{7}$$

where, fm_i is the index of the i^{th} sampled frame, and j is a random value sampled from the uniform distribution between -1 and 1. The temporal jitter can augment the dataset without disturbing the timing sequence of the sampled frames [30]. The sampling result can be represented as:

$$SF = \{fm_1, fm_2, \cdots, fm_S\} \tag{8}$$

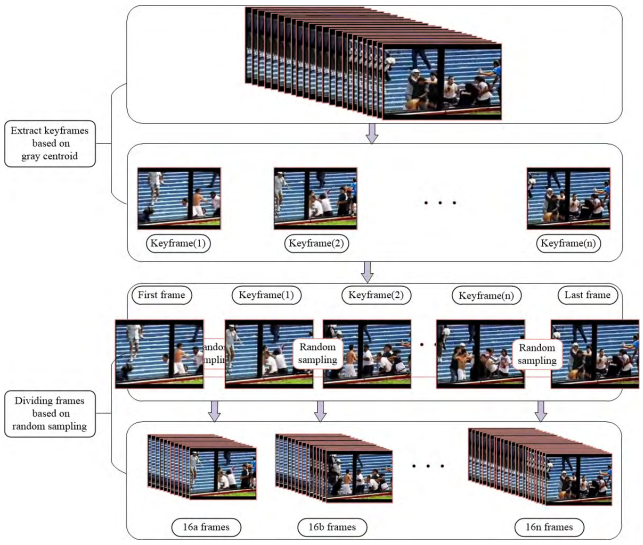

FIGURE 2. Sampling method based on key frames.

FIGURE 3. Frames captured from the hockey fight dataset. The frames in the first line are "non-fight" while "fight" in the second line.

IV. EXPERIMENTS

A. DATASETS

We evaluate the proposed method on three public datasets for violence detection, including Hockey fight, Movies, and Crowd violence.

Hockey fight [17]: This dataset was collected by Nievas *et al.* [17], and contains 1000 clips from hockey games of the National Hockey League (NHL), which were manually labeled as "fight" or "non-fight". Each clip last approximately 2 seconds, consists of about 41 frames with resolution 360×288. Fig.3 shows the frames captured from the Hockey fight dataset.

Movies [17]: This dataset contains 200 clips, as shown in Fig.4, 100 fight ones were extracted from action the Movies and 100 non-fight ones from public action recognition dataset.

Crowd violence [15]: This dataset was collected by Hassner *et al.* [15], and is mainly used to evaluate classifying quality of violence in crowded scenes. As shown in Fig.5, this dataset contains 123 videos, and is divided into 5 sets.

B. IMPLEMENTATION DETAILS

According to the varied length of the sample data of datasets, we conduct experiments separately. For all the experiments,

FIGURE 4. Frames captured from the movies dataset. The frames in the first line are "non-fight" while "fight" in the second line.

FIGURE 5. Frames captured from the crowd violence dataset. The frames in the first line are "non-fight" while "fight" in the second line.

the implementation is operated on a workstation with a NVIDIA GeForce GTX 1080 Ti GPU.

For Hockey fight dataset: video samples contain only about 41 frames, we present a 3D ConvNet directly. Under the assumption that longer clip length does less damage to the temporal structure of videos, we conduct a comparison experiment under the circumstances that the clip length is 16 and 32. In this paper, a uniform sampling method with an interval of 1 is used to select non-overlapping video segments with a fixed length as the input data of the network. Then, the size of frames is resized to 128×171 pixels, and we get input dimensions in this paper are $3 \times 16 \times 128 \times 171$, and $3 \times 32 \times 128 \times 171$, respectively. For network parameters, we train from scratch on the Hockey fight dataset for up to $20k$ iterations, and for $10k$ steps on the Hockey fight Dataset, with a $10\times$ reduction of learning rate. The batch size is 10. The initial learning rate is set among 0.0002 to 0.004, respectively. When training loss decays into a stable situation, the pre-training models are saved every 1000 iteration. Then the pre-training models that achieved high prediction accuracy are selected for later classification task. After training, we get clip accuracy, and the average of clip accuracy is calculated as the accuracy of video eventually.

For Movies dataset: we present a 3D ConvNet with uniform sampling. Considering the length of clips and the complexity of the network, we adopt a clip length of 16. The batch size is 10. The initial learning rate is set to 0.0003, and divided by 10 every 10k iterations. We stop the training at $10k$ iterations. We save pre-training models every 1k iterations, and finally choose best pre-training models for later classification task in similar circumstances with the Hockey fight dataset.

For Crowd violence dataset: As video samples are long-term, we conduct experiment using original diving method and our proposed method separately. Other parameters are same with the Movies dataset. After training, we perform the classification tests with five-fold cross validation.

TABLE 1. Comparison of optimal detection accuracy under varied learning rate and iterations with 16-frame clips on the hockey fight dataset.

Learning Rate	Iterations (k)	Video Accuracy (%)	Learning Rate	Iterations (k)	Video Accuracy (%)
0.0002	18	98.01	0.0008	20	98.85
0.0003	20	98.60	0.0009	18	98.83
0.0004	20	98.64	0.001	19	98.96
0.0005	19	98.76	0.002	17	98.35
0.0006	17	98.84	0.003	20	98.62
0.0007	19	98.44	0.004	20	97.32

TABLE 2. Comparison of optimal detection accuracy under varied learning rate and iterations with 32-frame clips on the hockey fight dataset.

Learning Rate	Iterations (k)	Video Accuracy (%)	Learning Rate	Iterations (k)	Video Accuracy (%)
0.0002	20	99.35	0.0008	18	99.62
0.0003	12	99.32	0.0009	16	99.27
0.0004	20	99.16	0.001	18	99.26
0.0005	14	99.45	0.002	10	99.42
0.0006	19	99.05	0.003	20	98.50
0.0007	20	99.36	0.004	15	97.20

C. EXPERIMENTAL RESULTS

In this part, we evaluate the modified 3D ConvNet on the three different dataset. Firstly, for Hockey fight dataset, we compare the performance of the network whilst varying the input clip length, initial learning rate, and max iterations. Table 1 shows the classification accuracy on the Hockey fight dataset with 16-frame clips. Table 2 shows the classification accuracy on the Hockey fight dataset with 32-frame clips.

As shown in Table 1, when the input clip length is set to 16, the initial learning rate is set to 0.001, and the iterations is set to 19k, we get 98.96% on the Hockey fight dataset, which is an optimal accuracy. Under similar condition, we get 99.62% on the Hockey fight dataset on the condition that the input clip length is set to 32, the initial learning rate is set to 0.0008, and the iterations is set to 18k.

Fig.6 plots video accuracy under input clip length of 16 and 32. It is clear to see that the performance of 32-frame clips outperforms the performance of 16-frame clips by 0.66%. This can be explained by the fact that longer clip do less harm to temporal structure of videos. The result shows that 32-frame clips are fitter for Hockey fight dataset.

We show a comparison of the performance of the presented 3D ConvNet and previous state-of-the-art methods on the Hockey fight dataset in Table 3. The traditional method, such as the HOF, HOG, MoSIFT, MoWLD, BoW, Sparse Coding, and some other method were tested on Hockey Fight Dataset, they got good performance. And the 3DCNN and the FightNet were also tested on this dataset, and the higher accuracy is 97% by FightNet. The architecture we adopted has

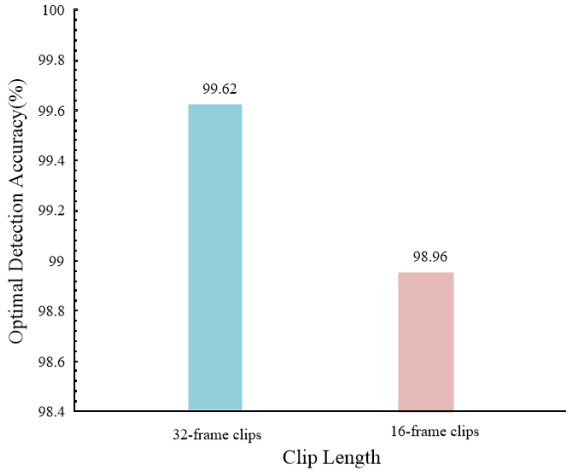

FIGURE 6. Comparison of video accuracy under the condition the clip length is set to 16 and 32, separately.

TABLE 3. Performance comparison with state-of-the-art on the hockey fight dataset.

Method	Accuracy (%)
HOF + BoW [17]	88.6
HOG + BoW [17]	91.7
MoSIFT + BoW [17]	90.9
MoWLD + BoW[34]	91.9
MoWLD+ Sparse Coding[34]	93.7
MoSIFT + KDE + Sparse Coding [19]	94.3
MoWLD + KDE + Sparse Coding[34]	94.9
MoIWLD + KDE + SRC [29]	96.8
3D-CNN [23]	91
FightNet [26]	97
3D ConvNet (16 frames)	**98.96**
3D ConvNet (32 frames)	**99.62**

TABLE 4. Comparison with optimal detection accuracy under varied iterations with 16-frame clips on the Movies dataset.

Iterations (k)	Video Accuracy (%)	Iterations (k)	Video Accuracy (%)
1	97.75	6	98.85
2	99.42	7	99.97
3	99.58	8	99.97
4	99.44	9	99.93
5	99.97	10	99.79

great advantage over previous models; and it brings overall performance to 99.62%. Apparently, the present 3D ConvNet has the advantage of accuracy compared to the other methods.

Secondly, for Movies dataset, we perform experiment with input clip length of 16. Table 4 shows the classification accuracy on the Movies dataset with 16-frame clips.

As shown in Table 4, we get optimal 99.97% on the Movies dataset. The result indicates that features extracted by 3D ConvNet are good features for Movies dataset. Table 5 shows a comparison of the performance of the presented 3D ConvNet and previous state-of-the-art methods on the Movies dataset. The traditional method, such as the HOF, HOG, and the MoSIFT did not got good performance.The performance of FightNet [26] on the Movies dataset reaches to 100%, outperforms 3D ConvNet by 0.03%. In terms of accuracy, there is a marginal difference between FightNet and our presented 3D ConvNet. Nevertheless, FightNet using three modalities, RGB images, optical flow fields and acceleration field as input, and these algorithms require a large quantity of extra space to store the input. Relatively speaking, the proposed 3D ConvNet just use RGB images as input, is a more applicable feature for violent video detection.

TABLE 5. Comparison of with state-of-the-art on the movies dataset.

Method	Accuracy (%)
HOF + BoW [17]	59
HOG + BoW [17]	49
MoSIFT + BoW [17]	84.2
Extreme Acceleration + SVM [16]	85.4
Extreme Acceleration + AdaBoost [16]	98.9
3D ConvNet (16 frames)	**99.97**
FightNet [26]	100

TABLE 6. Performance comparison with state-of-the-art on the crowd violence dataset.

Method	Accuracy (%)
LTP [15]	71.53
ViF [15]	81.3
MoWLD + BoW[34]	82.56
RVD[33]	82.79
MoWLD + SparseCoding[34]	86.39
MoSIFT + KDE + Sparse Coding [19]	89.05
MoWLD + KDE +SparseCoding[34]	89.78
MoIWLD + KDE + SRC [29]	93.19
3D ConvNet (uniform sampling method)	**93.5**
3D ConvNet (new sampling method)	**94.3**

Thirdly, for Crowd violence dataset, Fig.7 presents the comparison of classification results on five-fold cross validation test between origin dividing method and our proposed method. Fig.7 shows that our proposed sampling method is effective over original uniform sampling method. We use less clips and get a better result. For long-term video clips, one explanation for the better performance of our proposed method is that using key frames as diving node seems able to reflect the shift of video shots, making temporal information more consecutive.

As shown in Table 6, we compare our method with the other existing methods on the Crowd violence dataset. The LTP and the ViF in [15] showed that the accuracy was 71.53% and 81.3%, respectively. The accuracy of the method based on MoWLD was 82.56%. The tests also were performed on the MoWLD, MoSIFT and Sparse coding. However, our

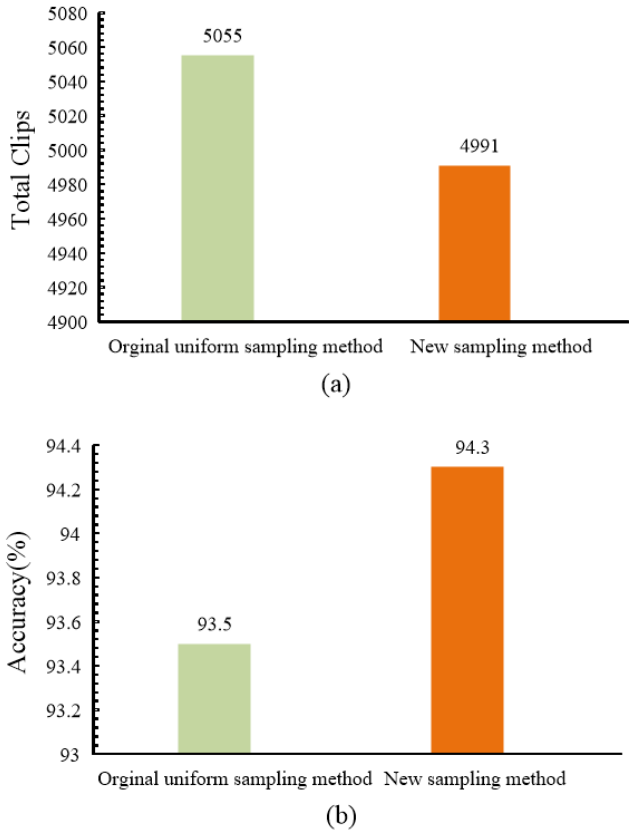

FIGURE 7. Comparison of origin uniform sampling method and our proposed method. (a) Comparison the number of total clips, and (b) Comparison of accuracy.

proposed sampling method is effective over previous state-of-the-art method. Our proposed method brings overall performance to 94.3%. By applying new sampled method, our approach outperforms original uniform sampling method.

V. CONCLUSIONS

In this paper, based on 3D ConvNet and key frame extraction algorithm, a novel violent video detection scheme is presented. In order to reduce redundancy and decrease the destruction of motion integrity caused by uniform sampling method, a new sampling method of frames is put forward, and the detection results showed that it is effective. For three different datasets, individualized strategies were studied to suit the detection of violence, and the results demonstrate that these modification is suitable. Performance comparisons with the existing schemes further demonstrate the effectiveness of the proposed approach. Our future work will focus on how to construct adaptive deep networks for violent video detection.

ACKNOWLEDGMENT

The authors would like to thank other students in the media computing laboratory, Minzu university of China, and the anonymous reviewers for their valuable comments and professional contributions to their improvement of this paper.

REFERENCES

[1] I. Laptev and T. Lindeberg, "On space-time interest points," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, Oct. 2003, pp. 432–439.
[2] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.
[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 568–576.
[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
[5] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in *Proc. 4th ACM Int. Conf. Multimedia*, 1996, pp. 21–30.
[6] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in *Proc. 5th ACM SIGMM Int. Workshop Multimedia Inf. Retr.*, 2003, pp. 109–115.
[7] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[8] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *Proc. Hellenic Conf. Artif. Intell.*, Berlin, Germany, 2006, pp. 502–507.
[9] C. Clarin, J. Dionisio, M. Echavez, and P. Naval, "DOVE: Detection of movie violence using motion intensity analysis on skin and blood," in *Proc. PCSC*, vol. 6, 2005, pp. 150–156.
[10] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *Proc. Int. Conf. Image Process.*, vol. 1, Oct. 1998, pp. 353–357.
[11] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *Proc. Pacific-Rim Conf. Multimedia*, Berlin, Germany, 2008, pp. 317–326.
[12] J. Lin and W. Wang, "Weakly-supervised violence detection in movies with audio and video based co-training," in *Proc. Pacific-Rim Conf. Multimedia*, Berlin, Germany, 2009, pp. 930–935.
[13] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," in *Proc. Hellenic Conf. Artif. Intell.*, 2010, pp. 91–100.
[14] A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," in *Proc. 16th IEEE Int. Conf. Pattern Recognit.*, vol. 1, Aug. 2002, pp. 433–438.
[15] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, 2012, pp. 1–6. doi: .10.1109/CVPRW.2012.6239348.
[16] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, vol. 2, Jan. 2014, pp. 478–485.
[17] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2011, pp. 332–339.
[18] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," CMU, Tech. Rep., 2009.
[19] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 3538–3542.
[20] P. Rota, N. Conci, N. Sebe, and J. M. Rehg, "Real-life violent social interaction detection," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3456–3460.
[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.
[23] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," in *Proc. ISVC*, Las Vegas, NV, USA, 2014, pp. 551–558.
[24] Q. Dai *et al.*, "Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning," in *Proc. MediaEval*, 2015, pp. 1–3.
[25] A. F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, vol. 2, Sep. 1999, pp. 850–855.
[26] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," *J. Phys., Conf. Ser.*, vol. 844, no. 1, Jun. 2017, Art. no. 012044.

[27] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 20–36.
[28] W. Song, P. Yang, J. Yu, and W. Jiang, "Terrorist video detection using visual semantic concepts," (in Chinese), *Netinfo Secur.*, vol. 16, no. 9, pp. 12–17, 2016.
[29] T. Zhang, W. Jia, X. He, and J. Yang, "Discriminative dictionary learning with motion Weber local descriptor for violence detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 696–709, Mar. 2017.
[30] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.
[31] I. Mironică, I. C. Duţă, B. Ionescu, and N. Sebe, "A modified vector of locally aggregated descriptors approach for fast video classification," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 9045–9072, 2016.
[32] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation," *Multimedia Tools Appl.*, vol. 74, no. 17, pp. 7379–7404, 2015.
[33] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools Appl.*, vol. 75, no. 12, pp. 7327–7349, 2016.
[34] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: A robust motion image descriptor for violence detection," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 1419–1438, 2017.
[35] V. Lam, S. Phan, D.-D. Le, D. A. Duong, and S. Satoh, "Evaluation of multiple features for violent scenes detection," *Multimedia Tools Appl.*, vol. 76, no. 5, pp. 7041–7065, 2017.
[36] Z. Xiong, Q. Shen, Y. Wang, and C. Zhu, "Paragraph vector representation based on word to vector and CNN learning," *CMC Comput., Mater. Continua*, vol. 55, no. 2, pp. 213–227, 2018.
[37] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.
[38] Q. Wu, Y. Li, Y. Lin, and R. Zhou, "Weighted sparse image classification based on low rank representation," *CMC Comput., Mater. Continua*, vol. 56, no. 1, pp. 91–105, 2018.
[39] Z. Pan, X. Yi, and L. Chen, "Motion and disparity vectors early determination for texture video in 3D-HEVC," *Multimedia Tools Appl.*, pp. 1–18, Nov. 2018. doi: 10.1007/s11042-018-6830-7.
[40] S. Zhou, W. Liang, J. Li, and J. U. Kim, "Improved VGG Model for Road Traffic Sign Recognition," *CMC Comput., Mater. Continua*, vol. 57, no. 1, pp. 11–24, 2018.
[41] Z. Pan, J. Lei, Y. Zhang, and F. L. Wang, "Adaptive fractional-pixel motion estimation skipped algorithm for efficient HEVC motion estimation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–12, 2018.

WEI SONG received the Ph.D. degree in traffic information engineering and control from Beijing Jiaotong University, in 2010. He was an Assistant Professor with the Research Institute of Information Technology, Tsinghua University, from 2011 to 2012. He was also a Visiting Scholar with the New Jersey Institute and Technology, from 2017 to 2018. He is currently an Associate Professor with the School of Information Engineering, Minzu University of China. He is also the Director of the Media Computing Laboratory. His research interests include multimedia networks, image processing, and natural language processing.

DONGLIANG ZHANG received the B.S. degree in computer science and technology from the Shandong University of Science and Technology, China, in 2014, and the M.S. degree in computer science and technology from the Minzu University of China, Beijing, China, in 2018. He is currently a Software Engineer with the Software Center, Bank of China. His research interests include deep learning and video recognition.

XIAOBING ZHAO received the M.S. degree from the College of Information Industry, Qingyun University, South Korea, in 2003, and the Ph.D. degree in computational linguistics from Beijing Language and Culture University, in 2007.

As a Principal Investigator, she has been taking charge of more than 20 projects, including the Natural Science Foundation Projects of China, and the High-Tech. Since 2007, she has been a Professor with the College of Information Engineering, Minzu University of China, Beijing, China. She is currently the Director of the National Language Resource Monitoring and Research Center of Minority Languages, Minzu University of China. She has published more than 50 papers in high-quality journals and international conferences. She has authored a book, and holds one national invention patent, two software copyrights, and one national invention patent. Her research focuses on machine translation, information retrieval, and other related areas of natural language processing.

Dr. Zhao is a member of the Chinese Information Processing Society of China (CIPS). She received the Qian Weichang First Prize of Chinese Information Processing Technology Progress. She is the Secretary General of the Chinese Minority Language Standardization Committee.

JING YU received the master's degree from the University of Science and Technology of China. He is currently pursuing the Ph.D. degree with Beijing Jiaotong University. He is also a Professor with the Department of Electronic Technology, Beijing Polytechnic. His research interests include image processing and electronic communication.

RUI ZHENG received the B.S. degree in automation from Tianjin University, Tianjin, China, in 2003, and the Ph.D. degree in control theory and control engineering from CASIA, Beijing, in 2009.

From 2009 to 2011, he was a Software Engineer with the Beijing Institute of Opto-Electronic Technology. From 2011 to 2012, he was a Research Assistant with CASIA. He is currently an Associate Professor with the School of Information Engineering, Minzu University of China. His research interest includes image processing.

ANTAI WANG received the B.S. degree in computational mathematics from Fudan University, China, the M.S. degree in mathematical statistics from York University, Canada, and the M.S. and Ph.D. degrees in statistics from the University of Rochester, in 2002. He was an Assistant Professor of biostatistics with Georgetown University, an Associate Professor of biostatistics with the University of Nebraska Medical Center (UNMC), and an Assistant Professor of clinical biostatistics with the Biostatistics Department, Columbia University, and also with the Herbert Irving Comprehensive Cancer Center, Columbia University. He has been an Associate Professor with the Department of Mathematical Sciences, New Jersey Institute of Technology (NJIT), since 2013. His main research interests include survival analysis and data analysis.