Violence behavior recognition of two-cascade temporal shift module with attention mechanism

Qiming Liang[a], **Yong Li**[b,*], **Bowei Chen**[c] **and Kaikai Yang**[a]
[a]Engineering University of PAP, Graduate Student Brigade, Xi'an, China
[b]Engineering University of PAP, College of Information Engineering, Xi'an, China
[c]Engineering University of PAP, Key Laboratory of Network and Information Security, Xi'an, China

Abstract. Violence behavior recognition is an important research scenario in behavior recognition and has broad application prospects in the field of network information review and intelligent security. Inspired by the long-short-term memory network, we estimate that temporal shift module (TSM) may have more room for improvement in the feature extraction ability of long-term information. In order to verify the above conjecture, we explored based on TSM. After many attempts, it was finally proposed to connect the two TSMs in a cascaded manner, which can expand the receptive field of the model. In addition, an efficient channel attention module was introduced at the front end of the network, which strengthened the model's spatial feature extraction capabilities. At the same time due to behavior recognition prone to over-fitting, we extended and processed on the basis of some open-source datasets to form a larger violence dataset and solved the problem of over-fitting. The final experimental results show that the algorithm proposed can improve the model's feature extraction ability of violent behavior in the space and temporal dimension and realize the recognition of violent behavior, which verified the above point of view. [DOI: 10.1117/1.JEI.30.4.043009]

Keywords: violence behavior recognition; convolutional neural network; attention mechanism; dataset.

Paper 210152 received Mar. 29, 2021; accepted for publication Jul. 9, 2021; published online Jul. 21, 2021.

1 Introduction

With the rapid popularization of mobile terminals, the Internet is uploading massive amounts of video data all the time, and these video data are likely to involve violent scenes, which will have an adverse impact on the health of the network environment. In order to maintain social safety and stability, functional departments such as police agencies and security companies have broad application requirements for intelligent video recognition systems in the field of on-duty security. The intelligent recognition of scenes involving violence can promptly feedback emergency security incidents to rear duty personnel, facilitating timely handling of incidents. Therefore, the recognition of violent behavior plays an important role in maintaining the safety and health of society and cyberspace.[1]

According to the recognition process, behavior recognition mainly includes three steps: video preprocessing, feature extraction, and behavior classification.[2] According to the method of feature extraction, behavior recognition can be divided into traditional behavior recognition[3,4] and behavior recognition based on deep learning.[5–8]

Traditional behavior recognition methods mainly extract features manually, and the types of features mainly include global features and local features. The global feature extraction mainly includes two methods: silhouette and human joint points. For example, Bobick and Davis[9] established a motion energy map to classify behaviors based on background subtraction. Yang[10] established the three-dimensional contour of the human body for feature extraction by determining the coordinates of the joint points. Local feature extraction mainly includes two feature

*Address all correspondence to Yong Li, liyong@nudt.edu.cn

extraction methods: spatiotemporal interest points sampling and trajectory tracking. For example, the dense trajectory extraction related algorithms dense trajectories and improved dense trajectories proposed by Wang et al.[11] and Wang and Schmid.[12]

According to the different feature extraction models, the current common methods of behavior recognition based on deep learning can be divided into three categories: two-stream CNN model, temporal model, and spatiotemporal model. Among them, the two-stream CNN model mainly extracts spatiotemporal information through two parallel channels and uses appropriate channel fusion to achieve behavior classification. For example, Simonyan and Zisserman[13] first proposed the two-stream approach for behavior recognition. Wang[14] adopted the temporal segment network to realize the recognition of long-term motion. Inspired by the two-stream CNN model, Feichtenhofer C[15] designed a lightweight two-stream network Slowfast, which reduces the complexity of the model.

Temporal models mainly rely on recurrent neural networks and their variants to extract temporal information in behavior and convolutional neural networks to extract spatial information. For example, Donahue et al.[16] introduced convLSTM[17] to replace the traditional long-short-term memory (LSTM) to achieve the fusion of spatiotemporal information. Li et al.[18] merged convLSTM with attention LSTM and constructed a new network structure VideoLSTM. The spatiotemporal model mainly uses 3D convolution to extract the spatiotemporal information of behaviors at the same time. In recent years, some scholars have adopted appropriate video preprocessing methods so that the spatiotemporal model can also achieve behavior classification through simple 2D convolution. Ji et al.[19] first applied 3D convolution to video behavior analysis and realized the extraction of spatial and temporal features from the video. Tran et al.[20] integrated on the basis of 3D convolution and proposed to establish convolutional 3D (C3D). C3D realized the use of large-scale video dataset training to learn the spatiotemporal characteristics of video, which improved the generalization ability of related algorithms. The 3D model is implicitly pretrained on ImageNet, and the 3D convolutional pretrained model is obtained in kinetics. Lin et al.[21] proposed the temporal shift module (TSM). By shifting and splicing adjacent frames in the temporal dimension, using 2D convolution to extract spatiotemporal information at the same time, the effect of 3D convolution is realized, and the problems of 3D convolution in parameters and calculations are solved.

However, the long-term information acquired by TSM network during behavior recognition is limited, the network structure is too simple, and over-fitting is prone to occur in the process of feature learning. In order to solve the problems above and also to further improve the accuracy of behavior recognition, this paper improves on the basis of the TSM network and conducts experimental exploration. The main contributions of this paper are as follows.

(1) A simple two-cascade TSM is proposed, which expands the receptive field of temporal dimensions and realizes the enhancement of long-term information extraction capabilities.
(2) Introduce the efficient channel attention (ECA) module at the front end of the TSM network to improve the network's feature extraction ability of spatial information to a certain extent and reduce the impact of overfitting on network performance.
(3) Data collection and multimedia processing are performed on the existing open-source datasets, and an expanded violent behavior recognition dataset is established, which solves the problem of overfitting and verifies the performance of the algorithm in a larger sample condition.

2 Related Work

2.1 *Temporal Shift Module*

Behavior recognition mainly obtains spatial information and temporal information contained in data during feature extraction. Traditional 3D convolution uses a 3D convolution kernel to perform convolution operations between adjacent multiple frames at the same time, which can extract the spatiotemporal feature information in the video, but it will inevitably lead to an increase in calculation. The TSM uses a simple data preprocessing method to convert the invisible temporal information in a single frame into extractable spatial feature information.

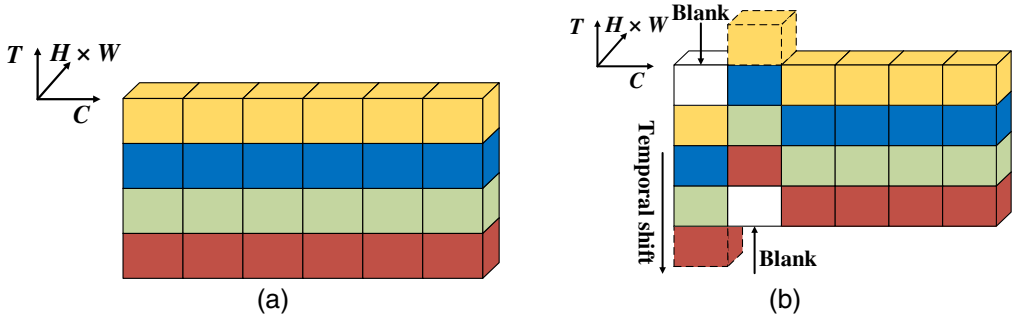

Fig. 1 TSM:[21] (a) original tensor and (b) shift module.

As shown in Fig. 1(a), several adjacent frames of images are stacked to form the original tensor, and the same color in the figure represents the same frame of image. Figure 1(b) shows the TSM. The TSM moves the channels forward and backward in the temporal dimension to perform simple feature fusion between adjacent frames. The fusion makes an independent single frame contain certain temporal information, and simple 2D convolution can be used to achieve spatiotemporal feature extraction.

The effect of convolution can be achieved through shift and multiply-accumulate operation, and that 3D CNN can be reduced in dimensionality in this way. For an infinitely one-dimensional vector $\boldsymbol{X}$ and a convolution kernel $W = (\, w_1 \quad w_2 \quad w_3 \,)$, the convolution operation is

$$y_i = w_1 x_{i-1} + w_2 x_i + w_3 x_{i+1}, \tag{1}$$

The above equation can also be decoupled by shift and multiply-accumulate operation:

$$x_i^{-1} = x_{i-1}, \quad x_i^0 = x_i, \quad x_i^{+1} = x_{i+1}, \tag{2}$$

$$Y = w_1 \sum x_i^{-1} + w_2 \sum x_i^0 + w_3 \sum x_i^{+1}, \tag{3}$$

$$\boldsymbol{Y} = w_1 \boldsymbol{X}^{-1} + w_2 \boldsymbol{X}^0 + w_3 \boldsymbol{X}^{+1}. \tag{4}$$

Among them, x_i represents the element in $\boldsymbol{X}$, y_i represents the result of convolution, X^{-1}, X^{+1} represent the infinite one-dimensional vector shifted back and forth by a unit, and $\boldsymbol{Y}$ represents the sum of the convolution results.

2.2 *Efficient Channel Attention Module*

The structure of the TSM behavior recognition network is too simple, and it is susceptible to interference from background information, causing serious over-fitting. In order to improve the network's feature extraction ability of spatial information, this paper introduces an ECA module.[22] As shown in Fig. 2, for the input tensor, the global average pooling is first performed

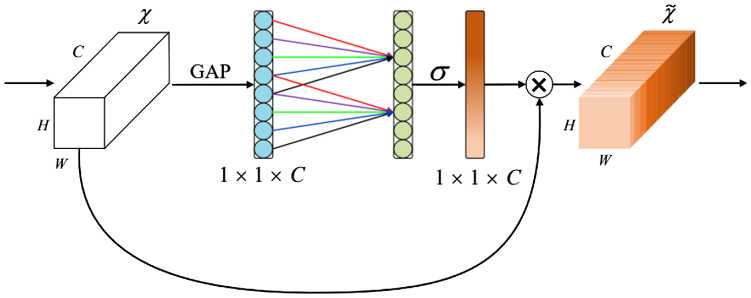

Fig. 2 ECA module structure diagram.[22]

without reducing the dimensionality, and then local cross-channel interaction is realized through one-dimensional convolution, and it is activated by the nonlinear function sigmoid. The result of activation is multiplied by the input tensor as the final output. The ECA module realizes local cross-channel interaction through one-dimensional fast convolution with adaptive size, which avoids channel dimensionality reduction and can reduce the interference of background information on feature extraction.

3 Module Design

3.1 *Intuition*

The TSM realizes the effective integration of spatiotemporal information in a single frame by performing simple channel shift in the temporal dimension. The shift of temporal dimension is similar to the function of RNN to a certain extent, which can realize the transfer of "memory" at different moments (Fig. 3).

The unidirectional TSM can be expressed mathematically as

$$\boldsymbol{Y} = w_1 \boldsymbol{X}^{-1} + w_2 \boldsymbol{X}^{0}. \tag{5}$$

The RNN can be expressed mathematically as

$$h^{(t)} = f(uh^{(t-1)} + wx^{(t)} + b). \tag{6}$$

Among them, $h^{(t)}$ is the state of the RNN at time t, u and w are the weights of the RNN nodes, and $x^{(t)}$ is the input at time t. Judging from the given network structure and mathematical formulas, there is a certain similarity between TSM and RNN, which is the source of inspiration for our follow-up work.

RNN cannot obtain long-term information when applied to behavior recognition, so some scholars have adopted a variant of RNN, LSTM,[23] to enhance the ability of the model to extract long-term information. Similarly, does the TSM have room for further improvement in the feature extraction capabilities of long-term information? This paper has launched an experimental analysis.

3.2 *Two-Cascade TSM Residual Module*

In order to strengthen the network's feature extraction capability for long-term information, it is simplest to move more channels forward and backward in the temporal dimension of the TSM. Based on the above ideas, this paper attempts to make various improvements to TSM.

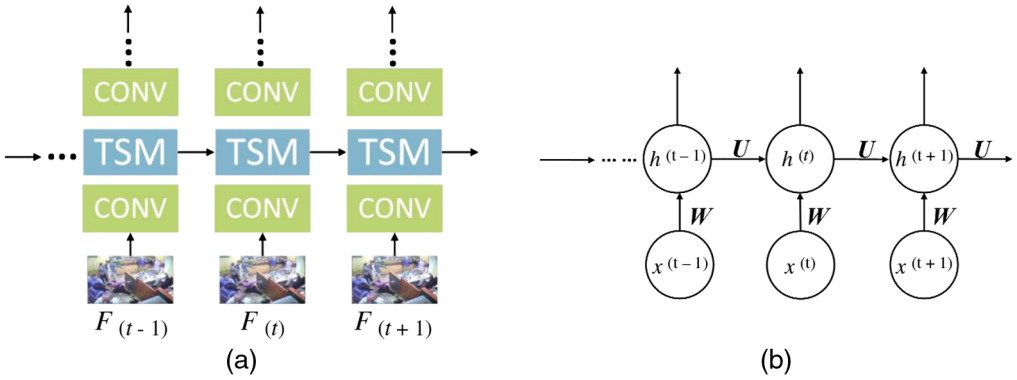

Fig. 3 TSM behavior recognition network and RNN are similar in structure and function, and both can realize the shift of information across different moments: (a) uni-directional TSM behavior recognition network and (b) the structure of RNN.

For example, introducing two temporal shifts in the channel dimension, changing the proportion of two temporal shifts in the tensor, and trying to manually add weights to various shifts. However, a large number of experiments have proved that these changes will not help improve the network's feature extraction ability for long-term information.

The above scheme unilaterally emphasizes the channel shift in the temporal dimension and ignores the overall feature fusion, resulting in the shift of temporal information only limited to the local area of the tensor, which destroys the integrity of the temporal and spatial information to a certain extent. Therefore, when strengthening the shift of temporal dimension, we must also consider the global fusion of spatio-temporal information. The TSM will reshape the data before and after the shift of the temporal dimension. This design is helpful to the integration of original data and shifted data, which is conducive to the global fusion of time and space information. Therefore, on the basis of the TSM behavior recognition network, this paper uses a simple two-cascade TSM, which strengthens the model's ability to extract temporal information to a certain extent and also realizes the effective integration of spatial–temporal information.

Similarly, suppose there are an infinite one-dimensional vector $\boldsymbol{X}$ and a convolution kernel $\boldsymbol{W}$ with a size of 1×3. Assume that the vector after a shift is $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \alpha \boldsymbol{X}^{-1} + \beta \boldsymbol{X}^{0} + \gamma \boldsymbol{X}^{+1}. \tag{7}$$

Among them, α, β, and γ are the weighting factors. Then after two cascades, the convolution result $\boldsymbol{Y}$ is

$$\boldsymbol{Y} = w_1 \boldsymbol{Z}^{-1} + w_2 \boldsymbol{Z}^{0} + w_3 \boldsymbol{Z}^{+1}, \tag{8}$$

$$\boldsymbol{Y} = w_a \boldsymbol{X}^{-2} + w_b \boldsymbol{X}^{-1} + w_c \boldsymbol{X}^{0} + w_d \boldsymbol{X}^{+1} + w_e \boldsymbol{X}^{+2}, \tag{9}$$

and

$$w_a = \alpha w_1, \tag{10}$$

$$w_b = \beta w_1 + \alpha w_2, \tag{11}$$

$$w_c = \gamma w_1 + \beta w_2 + \alpha w_3, \tag{12}$$

$$w_d = \gamma w_2 + \beta w_3, \tag{13}$$

$$w_e = \gamma w_3. \tag{14}$$

Then through inverse decoupling, the following conclusions can be drawn:

$$y_i = w_a x_{i-2} + w_b x_{i-1} + w_c x_i + w_d x_{i+1} + w_e x_{i+2}. \tag{15}$$

This realizes the convolution operation between the infinite one-dimensional vector $\boldsymbol{X}$ and the new convolution kernel $W' = (\, w_a \quad w_b \quad w_c \quad w_d \quad w_e \,)$. That is to say, without changing the original convolution kernel, a 1×3 convolution kernel can achieve a 1×5 convolution effect through the simple two cascades.

As shown in Fig. 4(a), based on the residual module, this paper adds two cascaded TSMs before the convolutional layer, forming a two-cascaded TSM residual module. It expands the receptive field of temporal dimension without changing the size of the convolution kernel. The experimental results show that the cascaded TSM independently shifts the temporal information, which improves the fusion of features in the temporal dimension and strengthens the model's feature extraction ability for long-term information. At the same time, the cascaded modules will restructure the shifted tensors and integrate spatiotemporal information before the second shift, avoiding the one-sided and fragmented temporal shift.

As shown in Figs. 4(b) and 4(c), this paper also tries to make more changes on the basis of the two-cascaded TSM, such as introducing short-cut in two TSM and expanding the cascade to three times. However, as shown in Fig. 5, the experimental results on the RWF-2000 dataset show that using different residual modules as the basic unit to construct a ResNet50[24] network

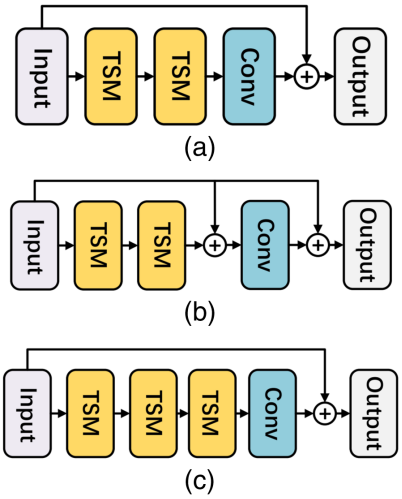

Fig. 4 Improved TSM residual module: (a) two-cascade TSM residual module, (b) shortcut two-cascade TSM residual module, and (c) three-cascade TSM residual module.

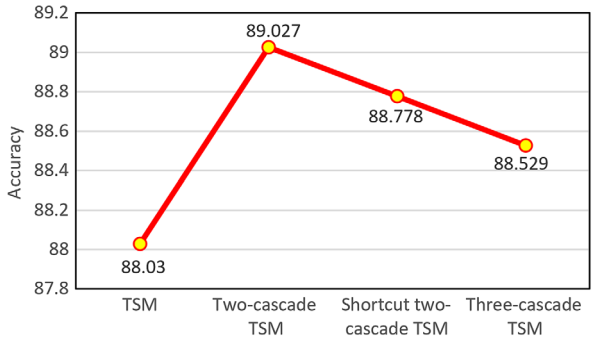

Fig. 5 Accuracy of different improvement schemes.

for violent behavior recognition, subsequent improvements to the two-cascaded TSM residual module will not help further improve the feature extraction ability of the model. Therefore, this paper chooses a simple two-cascaded TSM residual module as the basic unit to form a ResNet50 network for behavior recognition.

3.4 *Efficient Channel Attention Module*

The TSM network introduces the TSM into the residual module of ResNet50 and realizes the fusion of spatiotemporal information through simple data shift. Behavior recognition can be realized through the 2D convolutional neural network. This paper also uses the two-cascaded TSM as the basic unit to construct a two-cascaded TSM behavior recognition network on the basis of ResNet50. The specific structure is shown in Table 1. If a two-cascaded TSM is used, the two-cascade TSM is recorded as 1 otherwise it is recorded as 0.

This paper attempts to introduce the ECA module directly into the residual module of ResNet50 to form ECANet in the model construction, but the results show that this will greatly increase the amount of model parameters, and it will not help improve the accuracy of recognition.

As shown in Fig. 6, for the input video image F_i of the i'th frame, first extract the key information from the data through the attention module to complete the preprocessing of the information, which can reduce the interference caused by the background information to a certain extent. Then a 2D CNN network ResNet50 composed of two-cascaded TSM residual modules is used to realize feature extraction and classification of video frames that incorporate temporal and spatial information.

Table 1 Two-cascaded TSM behavior recognition network structure.

Conv layer	Structure	Output	Two-cascade TSM
conv1	7×764	112×112	0
conv2_x	3×3 maxpool	56×56	1
conv3_x	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$	28×28	1
conv4_x	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$	14×14	1
conv5_x	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 6$	7×7	1
	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$		
	Avgpool, fc	1×1	0

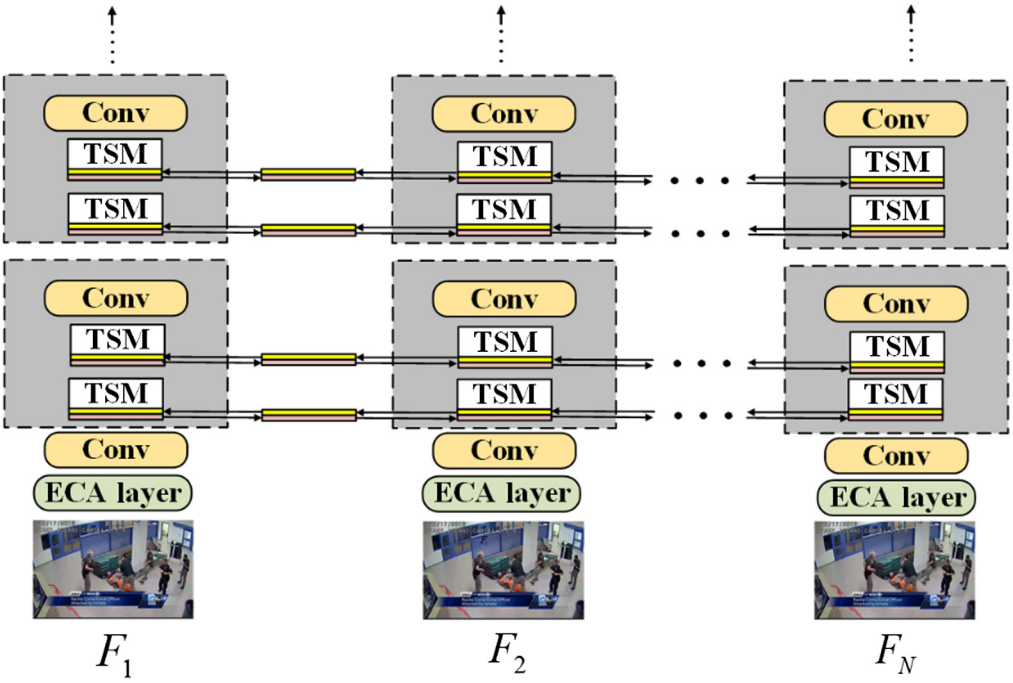

Fig. 6 ECA-two-cascaded TSM behavior recognition network structure diagram.

4 Experimental Design

4.1 *Dataset*

In order to fully test the performance of the algorithm and verify the proposed conjecture, this paper has conducted experiments on three open-source violent behavior recognition datasets and an expanded new dataset.

The crowd violence[25] dataset contains 246 video clips with a duration of 1.04 to 6.52 s, with an average duration of 3.6 s. This dataset mainly depicts scenes of crowd violence, but the scenes are relatively vague. The hockey dataset contains 1000 violent and non-violent videos collected from ice hockey game. The training set includes 800 video clips, and the validation set includes

200 video clips. The main content of the video is the violent actions in the ice hockey game. Each video is 2 s and contains 41 frames. Since the hockey dataset has a small number of videos, a single scene, and limited application value, it is difficult to meet the needs of deep neural network learning, so this paper introduces the latest RWF-2000[26] dataset. The dataset contains 2000 surveillance video clips collected from YouTube. The training set includes 1600 video clips, and the verification set includes 400 video clips. Each video clip is 5 s and contains 150 frames. It mainly includes violent behaviors such as two persons, multiple persons, and crowds. The scenes are rich and the recognition is difficult, and the video clips are all obtained by security cameras, without multimedia technology modification, which fits the actual scene and has high research value.

However, in the course of the experiment, this paper found that the TSM network has a serious over-fitting phenomenon in the RWF-2000 dataset, so this paper expands the dataset on the basis of the predecessors. Based on the open-source violence recognition dataset UCF-Crime, we collect hockey dataset, movies dataset, violent-flow dataset, HMDB51 dataset, and so on as the main scenes of violence in the video, and collect UCF101 and HMDB51 datasets as the main non-violent scenarios in the expanded dataset. The collected video is edited and processed by Adobe Premiere Pro, and the video clips that have nothing to do with behavior recognition are removed, and the data are unified into two kinds of video clips with a length of 1 and 5 s. Finally, this paper constructs a violence recognition dataset containing 5000 video clips, which greatly increases the number of samples, and the scene is richer than RWF-2000, which can solve the problem of over-fitting. Figure 7 shows the basic situation of the dataset.

Fig. 7 Basic situation of dataset: (a) crowd violence dataset, (b) hockey dataset, (c) RWF-2000 dataset, and (d) expanded dataset.

This paper selects 178 video clips from the crowd violence dataset as the training set, and the remaining 98 video clips as the validation set. 200 videos were randomly selected from the hockey dataset as the verification set, and each video was extracted into 41 consecutive images for experiment. Randomly select 400 videos from RWF-2000 as the verification set, and the rest are the training set. Every two frames are intercepted to form a 75-frame continuous image sequence. While reducing the amount of data, try to keep the temporal information in the data complete. For the expanded dataset, the video duration is mainly 1 and 5 s. 1000 video clips from 5000 video clips are randomly selected as verification sets, and all videos are intercepted as image sequences. After all the datasets are processed into continuous image sequences, the total size of the crowd violence dataset is 533 MB, the total size of the hockey dataset is 219 MB, the total size of the RWF-2000 dataset is 10.7 GB, and the size of the expanded dataset is 25.7 GB. Before loading the data into the model, we carry out random data preprocessing, such as clipping, scaling, and rotation, to realize the data transformation.

4.2 *Parameter Configuration*

The deep learning framework used in this paper throughout the training and testing process is Pytorch1.5, the operating system is Ubuntu 16.04, and the CPU is Intel I9-10920X. Use CUDA10.2 to accelerate the GPU and use two NVIDIA RTX2080super GPU with 8 GB of video memory for parallel computing. SGD is used to optimize the algorithm, and the TSM model trained on kinetics is used to reduce the risk of over-fitting and reduce the computational complexity of network training. In the comparative experiment, the experimental environment and dataset are set according to the introduction in this paper, and other basic configurations such as learning rate configuration, algorithm optimization method, and pretraining model are configured according to the instructions of the respective open source projects.

The learning rate adjustment method of the TSM algorithm is 100 epochs of training, the initial learning rate is 0.01, and the learning rate is adjusted to 10% when the training reaches 20 and 40 times. In this paper, when reproducing the original text experiment on the RWF-2000 dataset, it is found that the training loss value of the experiment decreases from the beginning of the training until it is stable, and the verification loss value of the experiment will drop rapidly before the training is started 20 epochs and keep increasing. This indicates that over-fitting occurred during the experiment. In response to the above problems, this paper designs a new learning rate adjustment method. The initial learning rate is 0.01, and the learning rate is adjusted to 90% of the original every two epochs. To a certain extent, this not only accelerates the adjustment speed of learning rate but also accelerates the rate of model learning. As shown in Fig. 8, the adjusted verification loss curve does not show a significant increase after 20 epochs, and the loss value is lower than the traditional method, indicating that the over-fitting problem in the experiment has been alleviated.

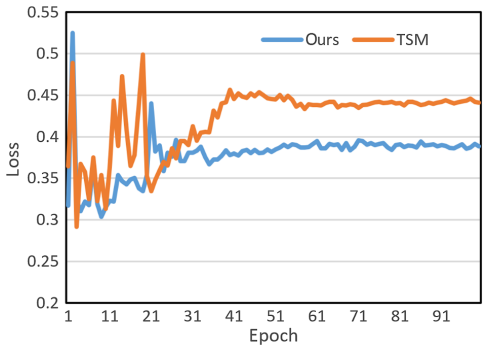

Fig. 8 The corresponding verification loss curve of different learning rate adjustment methods.

5 Results

After 100 epochs of training and verification of the model, Fig. 9 shows the accuracy curve of the experiment. The blue curve, green curve, and red curve in the figure are the verification accuracy curves of TSM algorithm, two-cascade TSM algorithm, and ECA-two-cascade TSM algorithm in each dataset, respectively.

As can be seen from Fig. 9, the accuracy of the two algorithms proposed in this paper is slightly higher than that of the traditional TSM algorithm. The accuracy curve is stable and the fluctuation is small, which shows that the algorithm can achieve effective feature extraction. Figure 9(a) shows the accuracy curve of various algorithms in the crowd violence dataset. It can be seen that the improved algorithm has a higher accuracy. Figure 9(b) shows the accuracy curve of various algorithms in the hockey dataset. It can be seen that the improved algorithm is obviously more accurate than the traditional algorithm, and the curve is more stable. Figure 9(c) shows the accuracy curve of the RWF-2000 dataset. From the graph, we can see that the accuracy of the improved algorithm is slightly higher than that of the traditional algorithm, but the accuracy decreases obviously after 20 epochs, which indicates that the algorithm has some over-fitting.

In order to solve the problem of over-fitting and to further verify the performance of this algorithm, as shown in Fig. 9(d), experiments are carried out in a larger dataset. The experimental results show that the accuracy of the three algorithms in the larger dataset is improved rapidly, and the accuracy curve is stable, which proves that the larger dataset does solve the problem of over-fitting, and further verifies the performance of this algorithm. Table 2 shows the specific situation of violence recognition by different algorithms.

As can be seen from Fig. 10, the algorithm proposed in this paper has a great improvement over the traditional algorithm. In the crowd violence dataset, the two-cascade TSM is 0.989% higher than the TSM, and the ECA-two-cascade TSM is 2.009% higher than the TSM. The two-cascade TSM in the hockey dataset is 0.55% higher than the TSM, and the ECA-two-cascade

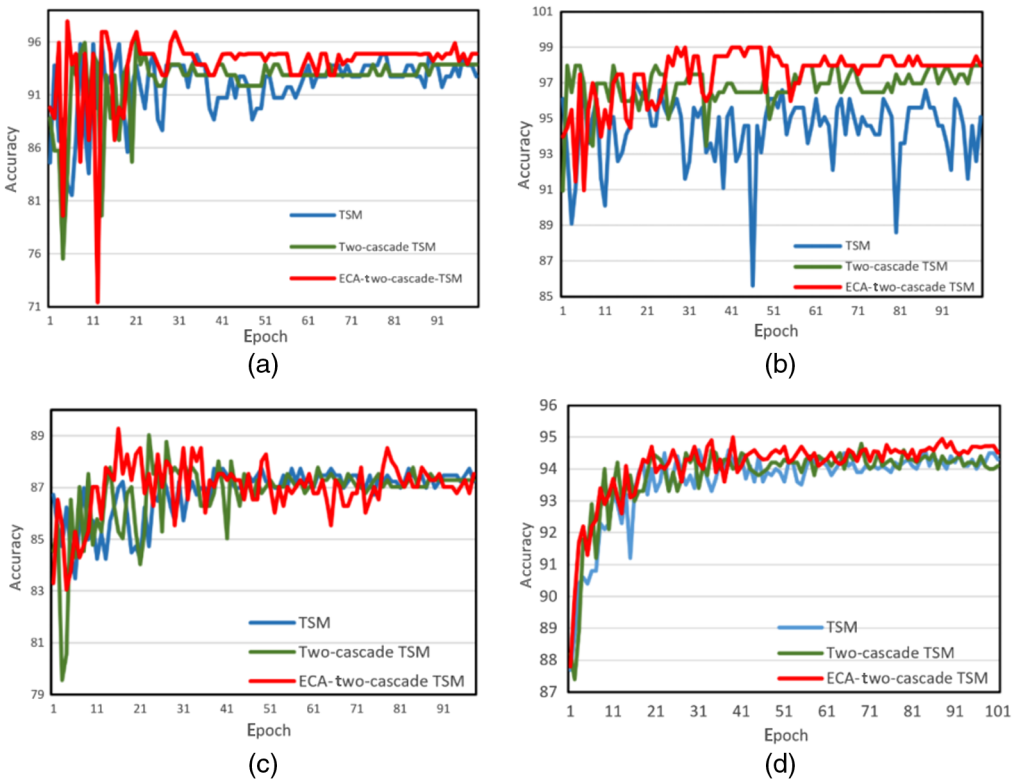

Fig. 9 Experimental verification accuracy curve: (a) crowd violence dataset accuracy curve, (b) hockey dataset accuracy curve, (c) RWF-2000 dataset accuracy curve, and (d) expanded dataset accuracy curve.

Table 2 Comparison of optimal accuracy.

Algorithm	Crowd violence	Hockey	RWF-2000	Expanded dataset
3D-CNN[3]	94.3	94.4	82.75	91.7
LRCN[21]	94.57	97.1	77	92.3
I3D[14]	88.89	97.5	85.75	93.3
AR-Net[27]	95.918	97.2	87.3	92.8
TSM[16]	95.95	97.5	88.03	94.6
TEA[28]	96.939	97.7	88.5	93.8
Two cascade TSM (ours)	96.939	98.05	89.027	94.8
ECA-two cascade TSM (ours)	97.959	98.995	89.277	95

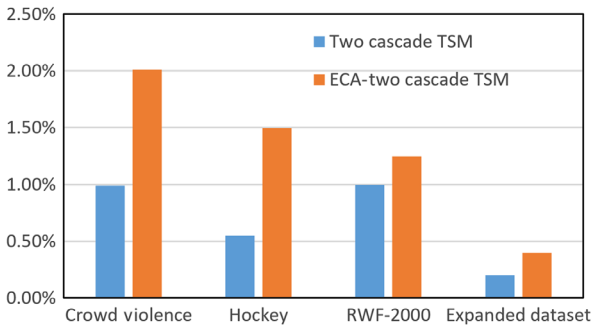

Fig. 10 The improvement of the algorithm in this paper compared with TSM.

TSM is 1.495% higher than the TSM. In the RWF-2000 dataset, the two-cascade TSM is 0.997% higher than TSM. The ECA-two-cascade TSM is 1.247% higher than TSM. In the expanded dataset, the two-cascade TSM is 0.2% higher than TSM. The ECA-two-cascade TSM is 0.4% higher than TSM.

The above results show that the two-cascade cascade of TSM modules can expand the model's receptive field in the temporal dimension, which also proves that there is still room for improvement in the feature extraction capabilities of the TSM module for long-term information. At the same time, it also suppresses the interference of background information through the ECA module and finally improves the performance of violence recognition.

6 Discussion

In order to recognize violence behavior in videos, this paper makes improvements on the TSM behavior recognition network. Inspired by LSTM, in order to strengthen the feature extraction ability of TSM module for long-term information, this paper proposes a two-cascaded TSM behavior recognition network, which expands the model's receptive field in the temporal dimension. In order to suppress the interference of background information, an ECA module is inserted at the front end to enhance the sensitivity of the model to spatial information. At the same time, in order to solve the over-fitting problem of some datasets in the experiment, this paper carries on the data expansion and multimedia processing on the basis of the existing datasets. Verification experiments in multiple datasets show that the proposed algorithm can achieve higher accuracy than the traditional algorithms. This means that the algorithm proposed in this paper can improve the ability of the network to understand the characteristics of time and space, solve the problem of over-fitting in the experiment, and realize the effective recognition of violence behavior.

Acknowledgments

This work was supported by the National Educational Science 13th Five-Year Plan Project (No. JYKYB2019012), the Basic Research Fund for the Engineering University of PAP (No. WJY201907), and the Basic Research Fund of the Engineering University of PAP (No. WJY202120).

References

1. S. Fischer, "Automatic violence detection in digital movies," *Proc. SPIE* **2916**, 212–223 (1996).
2. S. Cheng, "Research on feature extraction and recognition of human actions in video sequences," PhD Thesis, University of Electronic Science and Technology of China, China (2020).
3. W. Peng et al. "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 34, pp. 341–346 (2020).
4. G. Zhang et al., "Weighted score-level feature fusion based on Dempster-Shafer evidence theory for action recognition," *J. Electron. Imaging* **27**(1), 013021 (2018).
5. A. Mumtaz et al., "Fast learning through deep multi-net CNN model for violence recognition in video surveillance," *Comput. J.* **07**, bxaa061 (2020).
6. C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based LSTM networks," *Appl. Soft Comput.* **86**, 105820 (2020).
7. Q. Xiong et al. "Transferable two-stream convolutional neural network for human action recognition," *J. Manuf. Syst.* **56**, 605–614 (2020).
8. H. Ou and J. Sun, "Spatiotemporal information deep fusion network with frame attention mechanism for video action recognition," *J. Electron. Imaging*, **28**(2), 023009 (2019).
9. A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(3), 257–267 (2001).
10. X. D. Yang and Y. L. Tian, "Effective 3D action recognition using Eigen joints," *J. Vis. Commun. Image Represent.* **25**(1), 2–11 (2014).
11. H. Wang et al., "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision* **103**(1), 60–79 (2013).
12. H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. 2013 IEEE Int. Conf. Comput. Vision*, IEEE, pp. 3551–3558 (2013).
13. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, MIT Press, Vol. **1**, 568–576 (2014).
14. L. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *Eur. Conf. Comput. Vision*, Amsterdam, Springer, pp. 20–36 (2016).
15. C. Feichtenhofer et al., "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, Springer, pp. 6202–6211 (2019).
16. J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Springer, pp. 2625–2634 (2015).
17. X. Shi et al., "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," in *29th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, MIT Press, Vol. **1**, 802–810 (2015).
18. Z. Li et al., "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vision Image Understanding* **166**, 41–50 (2018).
19. S. Ji et al., "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1):221–231, (2012).
20. D. Tran et al., "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, Springer, pp. 4489–4497 (2015).
21. J. Lin, C. Gan, and S. Han, "TSM: temporal shift module for efficient video understanding," in *Proc. IEEE Int. Conf. Comput. Vision*, Springer, pp. 7083–7093 (2019).

22. Q. Wang et al., "ECA-Net: efficient channel attention for deep convolutional neural networks," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, IEEE, pp. 7083–7093 (2020).
23. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
24. K. He et al., "Deep residual learning for image recognition," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, IEEE, pp. 770–778 (2016).
25. T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: real-time detection of violent crowd behavior," in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit. Workshops*, IEEE, pp. 260–269 (2012).
26. M. Cheng, K. Cai, and M. Li, "RWF-2000: an open large scale video database for violence detection," in *25th Int. Conf. Pattern Recognit. (ICPR)*, IEEE (2020).
27. Y. Meng et al., "AR-net: adaptive frame resolution for efficient action recognition," in *Eur. Conf. Comput. Vision*, Springer, pp. 86–104 (2020).
28. Y. Li et al., "TEA: temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 909–918 (2020).

Qiming Liang is a MS degree candidate at the Engineering University of PAP (EUPAP). His research interests include behavior recognition.

Yong Li is an associate professor at the EUPAP. His research interests include pattern recognition.

Bowei Chen is a MS degree candidate at the EUPAP. His research interests include side channel attack.

Kaikai Yang is a MS degree candidate at the EUPAP. His research interests include information standard.