IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 6, December, 2022, pp. 2003-2010
Journal homepage: http://iieta.org/journals/ts

Intelligent Recognition of Key Frame Target Behavior in Video Surveillance Based on Lightweight Convolution Neural Network

Chuanzhong Mao[1*], Cuicui Wu[2], Xiangqun Sun[3], Ronghua Ji[2], Jin Zhang[1]

[1] Laboratory Management Centre University, Shandong Agriculture and Engineering University, Jinan 250100, China
[2] Business school University, Shandong Agriculture and Engineering University, Jinan 250100, China
[3] School of Information Science and Engineering University, Shandong Agriculture and Engineering University, Jinan 250100, China

Corresponding Author Email: z2013374@sdaeu.edu.cn

https://doi.org/10.18280/ts.390612

Received: 5 July 2022
Accepted: 3 November 2022

Keywords:
lightweight convolution neural network, video surveillance, target behavior recognition

ABSTRACT

In the analysis and processing of massive surveillance videos, target behavior recognition is an important task. Most researchers pay more attention to the lightweight of convolution operators in intelligent recognition systems or increase the complexity of lightweight modules, but lack of lightweight research on point-by-point convolution modules which occupy a large number of parameters and computation. For this reason, this article carries out the research on intelligent recognition of key frame target behavior in video surveillance based on lightweight convolution neural network. The three-dimensional position information of bone joints is extracted as the target behavior feature. Based on local vector aggregation descriptor, it makes a more compact representation of key frames of the surveillance video, and gives the generation process of local vector aggregation descriptor. After the structured pruning of the filter, the memory occupation of the processed network model is significantly reduced, and the lightweight of the model is realized. Experimental results verify the effectiveness of the model.

1. INTRODUCTION

With the development of science and technology, the application of Internet technology in real life is becoming more and more extensive. Under the fast-paced living environment, the emergence of multimedia transmission such as video, music and documents, and social applications such as instant messaging, email, WeChat and QQ facilitates the work and living [1-6]. At the same time, people's demands for security are getting higher and higher, and the development and research of security technology under the Internet environment has gradually gained the attention of experts and scholars [7-16]. In the analysis and processing of massive surveillance videos, target behavior recognition is an important task, which can prevent the occurrence of unexpected events such as fights, medical violence and school bullying. At the same time, it has the advantages of quick response and high execution efficiency that traditional security technology does not have [17-25]. Therefore, intelligent recognition of target behavior in video surveillance has great commercial value and practical significance.

The study of Aouayeb et al. [26] focuses on the difficulty of measuring the sleep stage of elephants from the surveillance video of elephant bran at night. To help zoologists, it’s suggested to use deep learning technology to automatically locate elephants in every camera monitoring, and then map the detected elephants to the barn plan. Generally speaking, the proposed method can monitor elephants in barns with high accuracy. The working environment of mechanical operators is complex, and the amount of video information in process monitoring is large, which brings serious interference to the abnormal behavior of human target recognition. The algorithm is more complex and the reaction time is longer. The study of Luo et al. [27] analyzes and compares RGB model, YUV model and gray level processing of images, and converts RGB color images into gray level images which is beneficial to image recognition when segmenting video sequence images. It analyzes the filtering and denoising effects of Gaussian low-pass filter and median filter. It is very suitable for image preprocessing and denoising after recognition. Before target detection, the computer graphics method is used to preprocess the target video sequence, which can highlight useful image features, remove useless image information and improve the detection effect. Most of the traditional monitoring systems are monocular cameras, whose main function is to play and store videos in the monitoring area, and to collect and judge the information in the videos through manual observation. The system constructed in the study of Li et al. [28] uses panoramic camera to capture 360-degree video, and uses deep learning to process the video to obtain the posture of human body; then processes it through human behavior recognition network to identify abnormal human behavior in real time. Through the double warning of local and remote clients and the artificial re-judgment of received abnormal behavior videos, managers can find and deal with abnormal behaviors in time. To solve the problem that it is difficult to detect the dangerous behavior of escalator passengers in real time and accurately, Du et al. [29] proposes an algorithm to identify the abnormal behavior of escalator passengers based on video surveillance. Thirdly, Hungarian allocation algorithm based on bone distance is used to assign passenger identification numbers in video. Finally, taking key points as inputs, the abnormal behavior of passengers is identified by graph convolution neural network. The results on GTX1080GPU show that the proposed recognition algorithm can accurately identify the abnormal behavior of passengers on escalators in real time.

To sum up, in order to obtain the fast response of the intelligent recognition system of video surveillance target behavior, it is necessary to lighten the model structure. Most researchers pay more attention to the lightweight of convolution operators in intelligent recognition systems or increase the complexity of lightweight modules, and lack of lightweight research on point-by-point convolution modules which occupy a large number of parameters and computation, resulting in unsatisfactory lightweight optimization results. For this reason, this article carries out the research on intelligent recognition of key frame target behavior in video surveillance based on lightweight convolution neural network. In the second chapter, it's decided to intelligently recognize the target behavior of key frames of the surveillance video based on bone data, that's, to extract the three-dimensional position information of bone joints as the target behavior features. In the third chapter, based on the BOW model and Fisher vector model, a local vector aggregation descriptor is proposed to represent the key frames of the surveillance video more compactly, and the generation process of local vector aggregation descriptor is given. In the fourth chapter, the filter structure pruning process is applied to the recognition model, and the memory occupation of the network model is significantly reduced after processing, thus realizing the lightweight of the model. Experimental results verify the effectiveness of the model.

2. LOCAL FEATURES EXTRACTION OF TARGET BEHAVIOR

Any complex behavior of video surveillance target can be regarded as a combination of human body part movement behavior sequence, and any form of human body part movement will drive the position of its bone nodes to change. In this article, the target behavior of key frames of the surveillance video is intelligently recognized based on bone data, that's, the three-dimensional position information of bone joints is extracted as the target behavior feature.

In this article, the joint position is described based on two kinds of position information. One is the normalized joint information, that's, absolute position information. Assuming that the coordinates of the *i*-th node in the *r*-th key frame of a bone sequence corresponding to a video surveillance are represented by $t_i(r)$, where $i\in\{1,...M\}$ and the number of joint nodes is represented by *M*, then the following absolute position information expression:

$$t_i(r)=(a_i(r),b_i(r),c_i(r)) \tag{1}$$

The other is the position information of bone paired nodes, that's, the relative position information. For each node *i* in the key frame, the following formula gives the calculation formula of its relative position distance with node *j*:

$$t_{i,j}=t_i-t_j \tag{2}$$

The three-dimensional node features of each node *i* in the key frame will be defined by the following formula:

$$t_i=\{t_{ij}\mid i\neq j\} \tag{3}$$

If the relative information of all joints of each bone map corresponding to key frames of the surveillance video is exhaustively listed by the above method, the extracted target behavior features will be redundant and even contain some worthless information. Because the center of bone is often in the position of marrow joint, the displacement of marrow joint is small in most complex behaviors of video surveillance targets, so this article sets the position coordinates of marrow joint as relative joint. Assuming that the position coordinates of the marrow joint are represented by $t_1(p)$, then the following definition of three-dimensional joint feature in frame *p*:

$$b(p)=\{t_i(p)-t_1(p)\mid i=2,...,15\} \tag{4}$$

Only extracting relative position features cannot fully describe the complex behavior of video surveillance targets. In order to achieve better behavior recognition effect, this article fuses the relative position features and relative displacement features of target bone nodes in key frames of the surveillance video. That's to say, the local features of target behavior in bone data are divided into two categories: relative displacement vector and relative position vector. The relative displacement vector can be generated based on the relative position of a single node in the bone sequence $r=\{1, 2,...p\}$, which represents the motion of a specific part of the target body in the surveillance video. Assuming that the coordinates (*a*,*b*,*c*) of node *i* in skeletal sequence *r* are represented by t^r_p, the time interval between two skeletal sequences *r*+1 and *r*-1 is represented by ΔP, and the number of key frames of a given sequence is represented by *p*, the following formula gives the specific definition formula:

$$u_i^r=\frac{t_i^{r+1}-t_i^{r-1}}{\Delta P},1<r<p \tag{5}$$

The relative position vector is also a three-dimensional vector, which represents the relevant information of the target body position in the surveillance video, and can be generated based on the row-pair position of nodes. Assuming that the coordinates (*a*,*b*,*c*) of different nodes in the same bone sequence are represented by t^r_i and t^r_l, the following formula gives its specific definition:

$$\theta_{i,l}^r=t_i^r-t_l^r \tag{6}$$

3. LOCAL FEATURES AGGREGATION OF TARGET BEHAVIOR

In order to encode the target behavior features in the key frames of the surveillance video, this article studies BOW model and Fisher vector model. It is found that the information loss cannot be ignored when the two methods are aggregated. Therefore, based on the two models, this article proposes a more compact representation of key frames of the surveillance video based on local vector aggregation descriptor, which can accurately describe the distribution features of target behavior features of key frames of the surveillance video. Figure 1 shows the local vector aggregation descriptor generation process.

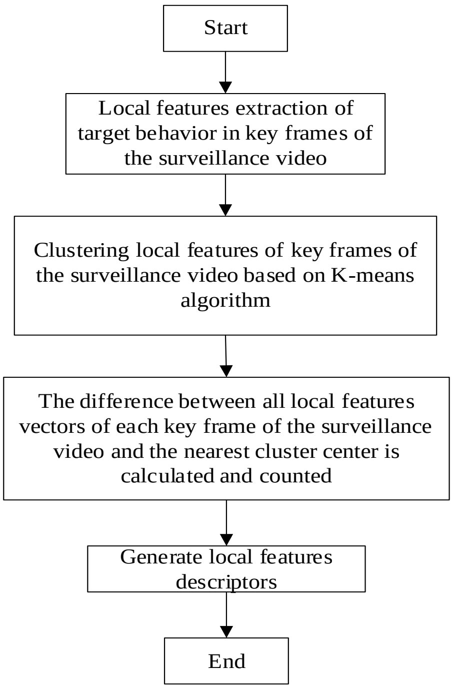
Start
Local features extraction of target behavior in key frames of the surveillance video
Clustering local features of key frames of the surveillance video based on K-means algorithm
The difference between all local features vectors of each key frame of the surveillance video and the nearest cluster center is calculated and counted
Generate local features descriptors
End

Figure 1. Flow chart of local vector aggregation descriptor generation process

Before the local vector aggregation descriptor is generated, it is necessary to cluster the local features of the target behavior in the key frames of the surveillance video. The clustering algorithm selected is K-means algorithm. *l* clustering centers, that's, codewords, can be obtained by this algorithm. The codebook can be generated after *l* codewords are aggregated, i.e.,

$$codebook=\{d_1,d_2,...,d_l\} \tag{7}$$

Furthermore, the difference between each local feature of the target behavior in the key frames of the surveillance video and the codebook can be calculated, and the local vector aggregation descriptor is the linear combination of all difference vectors. The specific steps of its generation are described in detail below:

STEP1: Assuming that the number of codewords is represented by *l*, the extracted local feature vector dimension of target behavior is represented by *c*, and *u* is a zero vector of dimension *l*c*, the initialization of descriptor is completed:

STEP2: Based on K-means algorithm, the local features *a* of each key frame of the surveillance video is clustered, and the nearest clustering center to *a* is searched. Assuming that the cluster center to which *a* belongs is denoted by *d*(*a*), assigning the vector *a* to the *i*-th cluster center is denoted by $d(a)=d_i$, that's, assigning *a* as follows:

$$d(a)=d_i=\left\{\begin{matrix} d_i\in codeboook|\ \forall d_j\in codebook\backslash d_i\|a-d_i\|\leq\|a-d_j\| \\ 1\leq l\leq l,1\leq j\leq l \end{matrix}\right\} \tag{8}$$

STEP3: Calculate and count the difference between all local features vectors of each key frame of the surveillance video and the nearest cluster center, and generate local features descriptors. The difference and u_i are calculated as follows:

$$u_i=\sum a-d_i \tag{9}$$

The local features descriptor vector of the *l*c* dimension can be represented by the following formula:

$$u=\left[u_1,u_2,...,u_l\right] \tag{10}$$

Assuming that the number of bone local features in each key frame is represented by *m* and the frame number of bone sequence is represented by $r=\{1,2,...p\}$, the extracted local features of target behavior in each key frame of the surveillance video can be represented by $g_{m,r}$. Next, at the stage of local features aggregation, that's, it's necessary to build a feature vector with fixed size dimensions for each behavior feature of the target in the key frame.

K-means clustering is performed for each group *m*, and the number of clusters is *n*, represented by $\{v_{m,n}\}$, where $n=\{1,2,...l\}$. Then principal component analysis is used to reduce the dimension of local features, and finally feature aggregation is carried out based on local features aggregation descriptor. Assuming that the set of the *n*-th cluster after initialization in the local features group *m* is represented by $R_{m,n}$, the feature vectors with fixed values are formed as follows:

$$R_{m,n}=\left\{g_{m,r}\mid n=\arg\min{}_t\left\|g_{m,r}-\lambda_{m,t}\right\|\right\} \tag{11}$$

The difference between the local features $g_{m,r}$ and the cluster center can be calculated by the following formula:

$$u_{m,n}=\sum_{g_{m,r}\in R_{m,n}}\left(g_{m,r}-\lambda_{m,n}\right) \tag{12}$$

The local vector aggregation descriptor of the *f*-th local feature of each bone sequence in each key frame target of the surveillance video can be composed of $u_{m,n}$, i.e.,

$$G_m=\left\lfloor u_{m,1},...,u_{m,l}\right\rfloor \tag{13}$$

The size of every G_m is determined by the number of cluster centers. By connecting G_m sequentially, a feature vector which can represent each kind of target behavior and action features can be finally generated.

4. LIGHTWEIGHT INTELLIGENT RECOGNITION MODEL OF TARGET BEHAVIOR

In order to reduce the model parameters and complexity, it is necessary to optimize the convolution neural network lightweight. At present, the mainstream method is pruning method. In order to obtain better recognition effect of key frames of target behavior in the surveillance video and avoid the limitation of traditional unstructured pruning, this article carries out filter structured pruning processing on the constructed recognition model, and the memory occupation of the processed network model is significantly reduced and can be applied to general hardware platform without restriction. Figure 2 shows the principle block diagram of filter pruning.

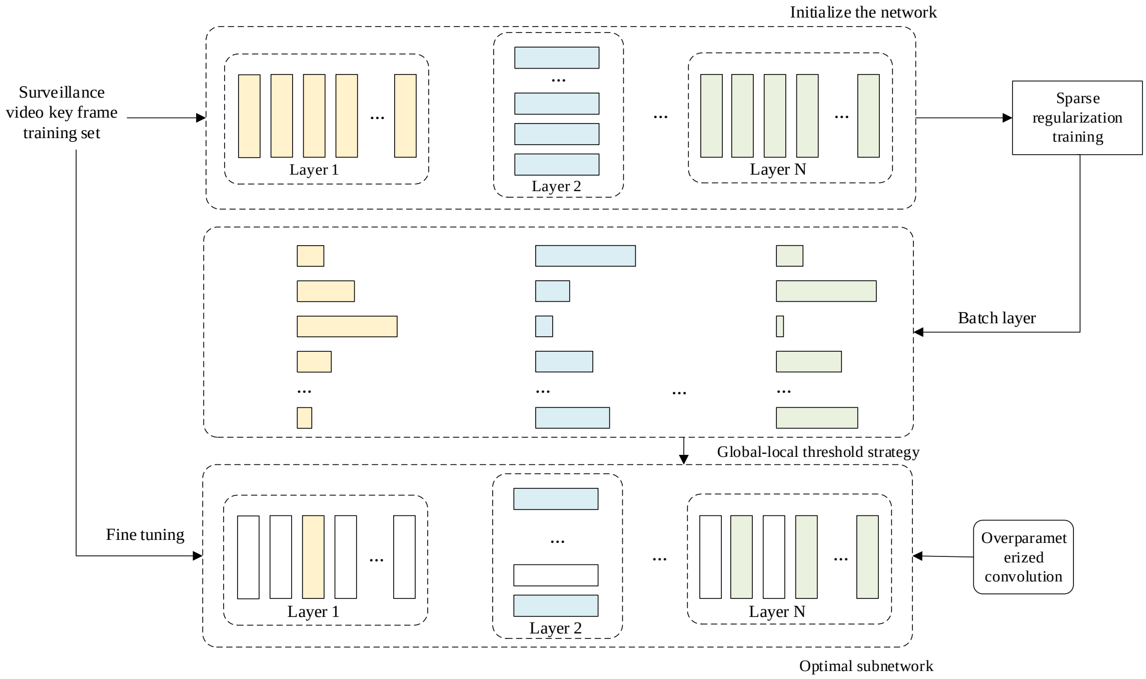
Initialize the network
Surveillance video key frame training set
Layer 1
Layer 2
Layer N
...
Sparse regularization training
Batch layer
Global-local threshold strategy
Fine tuning
Overparameterized convolution
Optimal subnetwork

Figure 2. Schematic diagram of filter pruning

The basic principle of filter pruning is to search for the optimal configuration of the number of channels in the feature graph in the model, and to reduce the model parameters to the greatest extent on the premise that the model has ideal recognition accuracy. Specifically, the scale factor of batch layer is used as the scale factor of feature map channel for sparse regularization, and the scale factor corresponding to some feature maps tends to 0 as the training target to iteratively adjust the weights and scale factors of the model, and finally realize the lightweight of the model by cutting off the feature map channels with small scale factors. If the input and output characteristic maps of the batch layer are represented by $c_{BN\text{-}in}$ and $c_{BN\text{-}out}$, the average value and standard deviation value of $c_{BN\text{-}in}$ on the current smallest batch image data Ω are represented by λ_Ω and ε_Ω, and the trainable linear transformation parameters are represented by α and γ, where the scale factor is represented by α and the offset term is represented by γ, then:

$$\dot{c} = \frac{c_{BN-in} - \lambda_\Omega}{\sqrt{\varepsilon_\Omega^2 + \kappa}}; c_{BN-out} = \alpha\dot{c} + \gamma \tag{14}$$

In order to prevent model over-fitting, this article introduces L_2 regularization operation. Assuming the model parameter is q, whose loss function is represented by $SE(q)$, the regularization operation process is given by the following formula:

$$\begin{aligned} J_{L1}(q) &= SE(q) + \mu|q| \\ J_{L2}(q) &= SE(q) + \mu q^2 \end{aligned} \tag{15}$$

Assuming that the derivative of $SE(q)$ at 0 is represented by c_0, then:

$$\frac{\partial SE(q)}{\partial q}\Big|_{q=0} = c_0 \tag{16}$$

The derivative of L_2 regularization at 0 can be calculated by the following formula:

$$\frac{\partial J_{K2}(q)}{\partial q}\Big|_{q=0} = c_0 + 2 \times \mu \times q = c_0 \tag{17}$$

The derivative of L_1 regularization at 0 can be calculated by the following formula:

$$\begin{aligned} \frac{\partial J_{K1}(q)}{\partial q}\Big|_{q=0^-} &= c_0 - \mu \\ \frac{\partial J_{K1}(q)}{\partial q}\Big|_{q=0^+} &= c_0 + \mu \end{aligned} \tag{18}$$

It can be seen from the above calculation process that the loss function of the model after L_2 regularization is derived at 0, and its derivative value is still d_0. However, if $L1$ regularization is performed, the derivative value of the loss function at 0 is not d_0, but has a sudden change from d_0-λ to d_0+λ. If the polarities of d_0-λ and d_0+λ are different, the loss function of the model will get a minimum at 0.

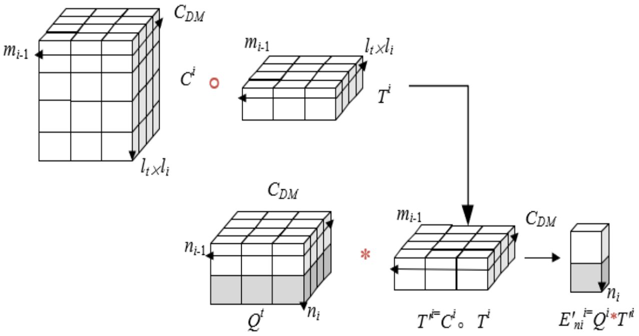

Figure 3. Schematic diagram of over-parameterized convolution layer calculation process

Because a large number of pruning operations will cause the model to lose the recognition accuracy of target behaviors of some key frames in the surveillance video, on the premise of adding no or a small amount of parameters, properly increasing the depth of the network can improve the expression ability of the model to a certain extent, thus improving the accuracy of target recognition. Therefore, this article introduces an over-parameterized convolution layer, which can improve the training speed and recognition performance of the model through deep convolution. Figure 3 shows a schematic diagram of the calculation process of the over-parameterized convolution layer.

Assuming that a patch of the *i*-th layer of traditional convolution layer is a two-dimensional tensor $T^i \in R^{(li \times li) \times mi\text{-}1}$, the size of convolution kernel is represented by l_i, and the number of channels of feature graph in *i*-1-th layer is represented by $m_{i\text{-}1}$. The output of convolution operators of the traditional convolution layer can be obtained by calculating the following formula:

$$E_{m_i}^{i} = \sum_{t=1}^{l_i \times l_i \times m_{i-1}} Q_{m_i p}^{i} T_t^{i} \tag{19}$$

Contrary to the traditional convolution layer, the deep convolution in the over-parameterized convolution layer introduced in this article will be calculated with each dimension channel of *T*, and finally C_{DM} dimension feature map is produced. C_{DM} is the depth multiplier of the feature map. For the over-parameterized convolution layer, the convolution kernel can be expressed as a three-dimensional tensor $Q^i \in R^{CDM(li \times li) \times mi\text{-}1}$. Therefore, the output of the deep convolution operators in the over-parameterized convolution layer is a $C_{DM} \times m_{i\text{-}1}$ dimensional feature map with the following formula:

$$E_{c_{mul} m_{i-1}}^{i} = \sum_{t=1}^{l_i \times l_i} Q_{t c_{mul} m_{i-1}}^{i} T_t^{i} \tag{20}$$

The over-parameterized convolution layer introduced in this article consists of a deep convolution with a trainable kernel $C^i \in R^{CDM(li \times li) \times mi\text{-}1}$ and a traditional convolution with a trainable kernel $Q^i \in R^{CDM(li \times li) \times mi\text{-}1}$, where $C_{DM} \geq (l_i \times l_i)$. The output of this convolution layer is an $m_{i\text{-}}$dimensional feature map with the following formula:

$$\begin{aligned} E_{m_i}^{i} &= \left(C^i, Q^i\right) \otimes T^i \\ &= Q^i * \left(C^i \circ Q^i\right) \ (x) \\ &= \left(C^i \circ Q^i\right) * T^i \ (y) \end{aligned} \tag{21}$$

5. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 4 and Figure 5 show the performance changes of the model under different quantities of relative displacement features. It can be seen from the figures that the relative displacement feature can well characterize the behavior and action features of the key frame target in the surveillance video. And with the increase of the number of extracted features, the recognition rate and accuracy of target behavior are significantly improved, and the time consumed by extracting features is also significantly increased. Compared with relative displacement features, the relative position feature has a little poor ability to describe the action feature of the target behavior in the key frame of the surveillance video. When all three groups of displacement features are used, the recognition accuracy increases slowly and the feature extraction time increases rapidly, so it is just divided into three groups. When each group of nodes does not coincide, the extraction time is the shortest and the recognition effect is very good. With the increase of the number of extracted features, the accuracy and recognition rate of target behavior have been improved, but the growth rate is slow. After careful consideration, the relative displacement features are set to 4 and the relative position features are set to 6.

Figure 6 shows the recognition rate of target behavior under three forms of features: relative displacement feature, relative position feature and fusion feature. As can be seen from Figure 5, the recognition performance of the model is the worst when only extracting the relative position features of the target behavior in the key frames of the surveillance video, and the accuracy of behavior recognition is only about 60%. Compared with extracting a single form feature, the fusion of relative displacement features and relative position features has a better description effect on the target behavior and action features, and the accuracy of behavior recognition reaches more than 80%.

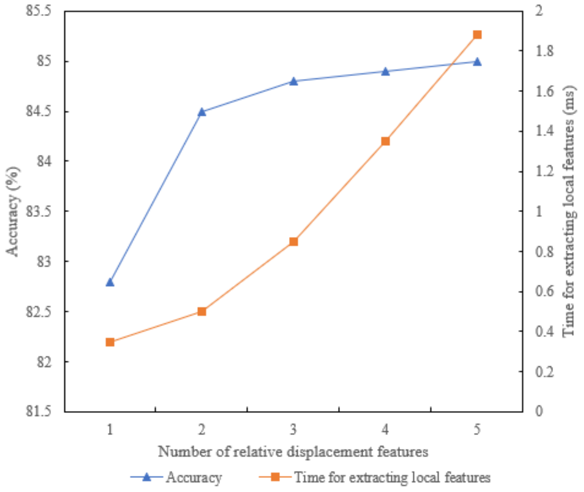

Figure 4. Change of model performance under different quantities of relative displacement features

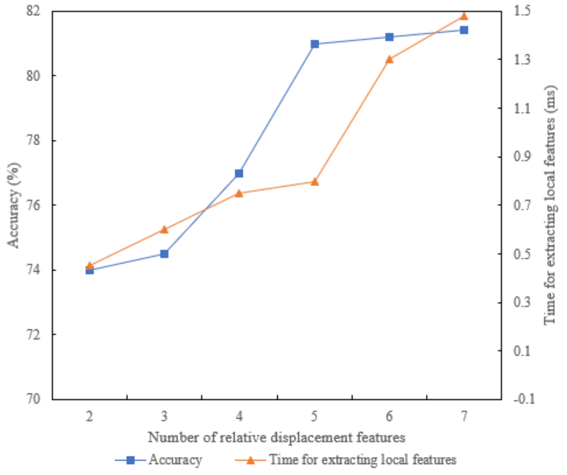

Figure 5. Change of model performance under the different quantities of relative position features

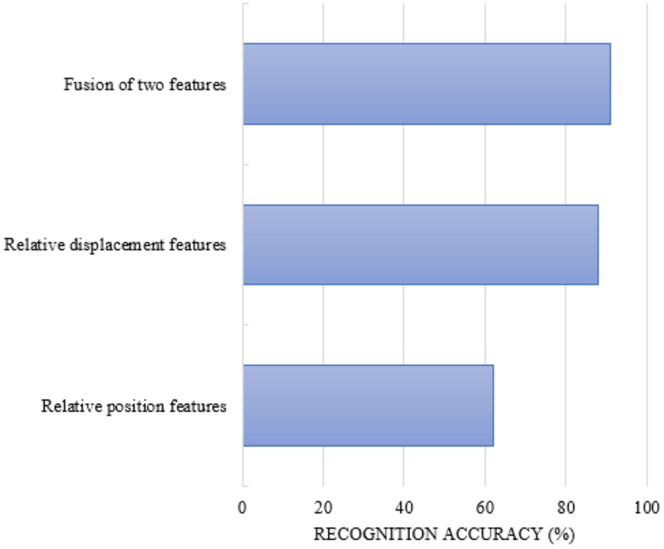

Figure 6. Recognition rate of target behavior under different formal features

In this article, three popular neural network architectures, *VGG*-19 (backbone 1), *ResNet*-50 (backbone 2) and *DenseNe*t-121 (backbone 3), are used to evaluate the proposed method of intelligent recognition of target behavior in key frame of the surveillance video. Table 1 shows the number of parameters, operation speed and classification accuracy of the model built with three backbone networks as the core architecture.

Table 2 summarizes the results of pruning experiments on *VGG*-19 (backbone 1), *ResNet*-50 (backbone 2) and *DenseNet*-121 (backbone 3). When 65% of the channels of *VGG*-19 (backbone 1) are pruned, the number of parameters is reduced by nearly 4/5, and the number of floating-point operations per second is reduced to 2/5. When 80% of the channels are pruned, the number of parameters and floating-point operations per second are reduced to 4/5 and 1/2 of the original ones, respectively, so the efficiency of the proposed algorithm is more ideal.

Compared with *VGG*-19 (backbone 1), the parameters of *ResNet*-50 (backbone 2) and the number of floating-point operations per second are reduced, but the amplitude is not large, which is mainly caused by the selection channel of its bottleneck structure. The parameters and the number of floating-point operations per second of *DenseNet*-121 (backbone 3) are reduced by 2/5, 1/5 and 4/5, 4/5 respectively when the pruning is 50% and 65%. It can be seen that the pruning algorithm proposed in this article can greatly improve the execution efficiency of the target behavior recognition model, achieve a larger compression and speedup ratio, and obtain obvious recognition performance advantages and satisfactory lightweight effect.

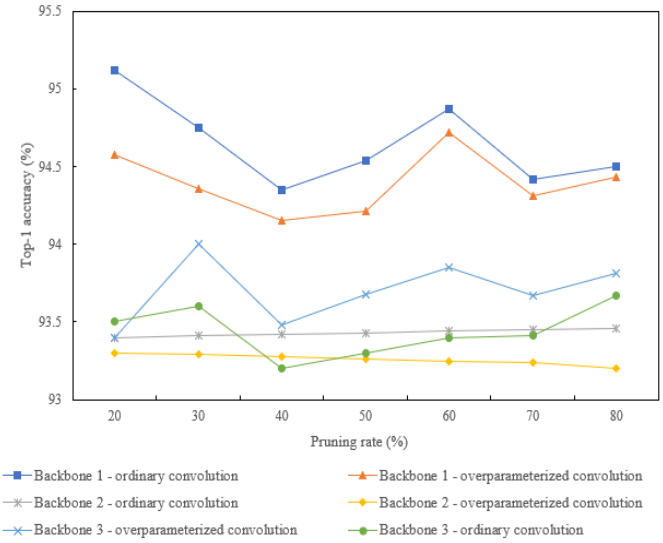

Figure 7. Performance comparison of models before and after introduction of overparameterized convolution under different pruning rates

Table 1. Number of parameters, operation speed and classification accuracy of different models

Model	Number of parameters	Floating-point operations per second	*Top*-1 classification accuracy (%)
Backbone 1 before optimization	21.50*M*	362.15*M*	95.41
Backbone 1 after optimization	21.50*M*	347.15*M*	91.03
Backbone 2 before optimization	1.85*M*	292.28*M*	94.27
Backbone 2 after optimization	1.85*M*	241.04*M*	95
Backbone 3 before optimization	1.63*M*	295.38*M*	92.48
Backbone 3 after optimization	1.63*M*	263.47*M*	96.35

Table 2. Pruning results of different models

	Method	Pruning ratio	Number of parameters	Floating-point operations per second	*Top*-1↑(%)
Backbone 1	*CART* algorithm	65%	2.14*M*	348*M*	0.13
	Algorithm	65%	4.61*M*	241.74*M*	0.04
	Algorithm	75%	3.74*M*	162.38*M*	0.18
Backbone 2	*CART* algorithm	50%	1.25*M*	315*M*	0.31
	Algorithm	50%	1.04*M*	241.93*M*	0.85
	CART algorithm	65%	1.63*M*	224*M*	0.16
	Algorithm	65%	1.47*M*	127*M*	0.57
Backbone 3	*CART* algorithm	50%	0.63*M*	369*M*	0.95
	Algorithm	50%	0.61*M*	214.57*M*	0.24
	CART algorithm	65%	0.31*M*	262*M*	0.42
	Algorithm	65%	0.11*M*	52.69*M*	0.29

In order to further verify the effectiveness of lightweight optimization of target behavior recognition model by introducing overparameterized convolution module in this article, *VGG*-19 (backbone 1), *ResNet*-50 (backbone 2) and *DenseNet*-121 (backbone 3) are still used in this article to carry out a series of experiments on a small sample set of the surveillance video images. The performance comparison results of the models are given in Figure 7. It can be seen that no matter what kind of backbone model, the target behavior recognition model with overparameterized convolution module is always better than the network model without overparameterized convolution module in recognition accuracy.

6. CONCLUSION

This article studies intelligent recognition of key frame target behavior in video surveillance based on lightweight convolution neural network. The three-dimensional position information of bone joints is extracted as the target behavior feature. Based on local vector aggregation descriptor, it makes a more compact representation of key frames of the surveillance video, and gives the generation process of local vector aggregation descriptor. After the structured pruning of the filter, the memory occupation of the processed network model is significantly reduced, and the lightweight of the model is realized. The experimental results show the performance changes of the model under different quantities of relative displacement features, and summarize the recognition rates of target behavior under three forms of features: relative displacement features, relative position features and fusion features. It is verified that the fusion of relative displacement features and relative position features is better for describing target behavior and action features. Three popular neural network architectures, *VGG*-19 (backbone 1), *ResNet*-50 (backbone 2) and *DenseNet*-121 (backbone 3), are used to evaluate the proposed method of intelligent recognition of target behavior in key frames of the surveillance video. It's verified that the pruning algorithm proposed in this article can greatly improve the execution efficiency of target behavior recognition model. Finally, the performance comparison of the models before and after the introduction of overparameterized convolution under different pruning rates is completed, which verifies the effectiveness of the introduction of overparameterized convolution module.

REFERENCES

[1] Wan, B., Xu, C. (2021). Application of computer information technology in internet. In 2021 4th International Conference on Information Systems and Computer Aided Education, pp. 2085-2089. https://doi.org/10.1145/3482632.3484104

[2] Chander, A. (2021). Protecting the global internet from technology cold wars. Communications of the ACM, 64(9), 22-24. https://doi.org/10.1145/3473606

[3] Yan, G. (2019). Simulation analysis of key technology optimization of 5G mobile communication network based on Internet of Things technology. International Journal of Distributed Sensor Networks, 15(6). https://doi.org/10.1177/155014771985145

[4] Wang, S.Y. (2021). Online learning behavior analysis based on image emotion recognition. Traitement du Signal, 38(3): 865-873. https://doi.org/10.18280/ts.380333

[5] Ning, Z., Yang, Y., Zhang, Y. (2019). Research on the trusted protection technology of internet of things. Cluster Computing, 22(6): 14339-14348. https://doi.org/10.1007/s10586-018-2294-9

[6] Cai, S., Wang, X., Zhao, Y. (2018). Revenue model of supply chain by internet of things technology. IEEE Access, 7: 4091-4100. https://doi.org/10.1109/ACCESS.2018.2888952

[7] Fujimoto, S., Kogure, J. (2019). ConnectionChain: Security technology for securely linking blockchains. Fujitsu Scientific & Technical Journal, 55(5): 47-52.

[8] Xie, Y.F., Zhang, S., Liu, Y.D. (2021). Abnormal behavior recognition in classroom pose estimation of college students based on spatiotemporal representation learning. Traitement du Signal, 38(1): 89-95. https://doi.org/10.18280/ts.380109

[9] Devi, E.A., Joany, R.M., Yogalakshmi, S., Therase, L.M. (2019). Digital video steganography technology for security application. In IOP Conference Series: Materials Science and Engineering, 590(1): 012057. https://doi.org/10.1088/1757-899X/590/1/012057

[10] Wawage, P., Deshpande, Y. (2022). Real-Time Prediction of Car Driver's Emotions Using Facial Expression with a Convolutional Neural Network-Based Intelligent System. Acadlore Transactions on AI and Machine Learning, 1(1): 22-29. https://doi.org/10.56578/ataiml010104

[11] Fan, P., Liu, Y., Zhu, J., Fan, X., Wen, L. (2019). Identity management security authentication based on blockchain technologies. Int. J. Netw. Secur., 21(6): 912-917. https://doi.org/10.6633/IJNS.201911 21(6).04

[12] Pérez, S., Garcia-Carrillo, D., Marín-López, R., Hernández-Ramos, J.L., Marín-Pérez, R., Skarmeta, A.F. (2019). Architecture of security association establishment based on bootstrapping technologies for enabling secure IoT infrastructures. Future Generation Computer Systems, 95: 570-585. https://doi.org/10.1016/j.future.2019.01.038

[13] Mansfield-Devine, S. (2019). The state of operational technology security. Network Security, 2019(10): 9-13. https://doi.org/10.1016/S1353-4858(19)30121-7

[14] Sun, J., Zhang, N. (2019). The Mobile payment based on public-key security technology. In Journal of Physics: Conference Series, 1187(5): 052010. https://doi.org/10.1088/1742-6596/1187/5/052010

[15] Si, H., Sun, C., Li, Y., Qiao, H., Shi, L. (2019). IoT information sharing security mechanism based on blockchain technology. Future Generation Computer Systems, 101, 1028-1040. https://doi.org/10.1016/j.future.2019.07.036

[16] Takafumi, K., Lee, K.A. (2019). Safety, security, and convenience: The benefits of voice recognition technology. NEC Technical Journal, 13(2): 83-86.

[17] Liu, F., Chen, Z., Wang, J. (2019). Video image target monitoring based on RNN-LSTM. Multimedia Tools and Applications, 78(4): 4527-4544. https://doi.org/10.1007/s11042-018-6058-6

[18] Morev, K., Kazanskaya, A., Nalesnaya, Y. (2019). Intelligent video monitoring system: technical and economic aspects. In 2019 Ural Symposium on Biomedical Engineering, Radioelectronics and

Information Technology (USBEREIT), pp. 272-275. https://doi.org/10.1109/USBEREIT.2019.8736675

[19] Zhang, F., Sun, H., Ge, J., Zhai, Y. (2019). Automatic monitoring & early warning of video receiving system. In 2019 International Conference on Robots & Intelligent System (ICRIS), pp. 191-193. https://doi.org/10.1109/ICRIS.2019.00057

[20] Zhen, S., Qingdang, L., Lu, W., Junfei, W. (2019). A distributed video monitoring system architecture for semantic retrieval. In 2019 2nd International Conference on Safety Produce Informatization (IICSPI), pp. 574-577. https://doi.org/10.1109/IICSPI48186.2019.9096038

[21] Bouvier, C., Balouin, Y., Castelle, B., Holman, R. (2019). Modelling camera viewing angle deviation to improve nearshore video monitoring. Coastal Engineering, 147: 99-106. https://doi.org/10.1016/j.coastaleng.2019.02.009

[22] Shilin, A.N., Dementyev, S.S. (2019). Designing video measuring device for monitoring ice deposits on overhead power lines. Russian Electrical Engineering, 90(5): 407-411. https://doi.org/10.3103/S1068371219050122

[23] Feng, Y., Zhao, S., Liu, H. (2019). Target tracking based on multi feature selection fusion compensation in monitoring video. Automatic Control and Computer Sciences, 53(6): 522-531. https://doi.org/10.3103/S0146411619060051

[24] Barthélemy, J., Verstaevel, N., Forehead, H., Perez, P. (2019). Edge-computing video analytics for real-time traffic monitoring in a smart city. Sensors, 19(9): 2048. https://doi.org/10.3390/s19092048

[25] Iozza, L., Lázaro, J., Cerina, L., Silvestri, D., Mainardi, L., Laguna, P., Gil, E. (2019). Monitoring breathing rate by fusing the physiological impact of respiration on video-photoplethysmogram with head movements. Physiological Measurement, 40(9): 094002. https://doi.org/10.1088/1361-6579/ab4102

[26] Aouayeb, S., Desquesnes, X., Emile, B., Mulot, B., Treuillet, S. (2022). Intelligent video surveillance for animal behavior monitoring. In: Mazzeo, P.L., Frontoni, E., Sclaroff, S., Distante, C. (eds) Image Analysis and Processing. ICIAP 2022 Workshops. ICIAP 2022. Lecture Notes in Computer Science, vol 13374. Springer, Cham. https://doi.org/10.1007/978-3-031-13324-4_31

[27] Luo, H., Liu, Z., Liu, X. (2021). Research on video image preprocessing for monitoring abnormal behavior of mechanical operators. In 2021 International Conference on Electronics, Circuits and Information Engineering (ECIE), pp. 216-219. https://doi.org/10.1109/ECIE52353.2021.00053

[28] Li, J., Xie, H., Zang, Z., Wang, G. (2020). Real-time abnormal behavior recognition and monitoring system based on panoramic video. In 2020 39th Chinese Control Conference (CCC), pp. 7129-7134. https://doi.org/10.23919/CCC50068.2020.9188630

[29] Du, Q., Huang, L., Tian, L., Huang, D., Jin, S., Li, M. (2020). Recognition of passengers' abnormal behavior on escalator based on video monitoring. Huanan Ligong Daxue Xuebao/Journal of South China University of Technology (Natural Science), 48(8): 10-21. https://doi.org/10.12141/j.issn.1000-565X.200010