JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 777-796 (2023)
DOI: 10.6688/JISE.202307_39(4).0005

Violence Detection Method Based on Convolution Neural Network and Trajectory

JIANXIN LI[1,11], JIE LIU[2], CHAO LI[3,+], WENLIANG CAO[4], BIN LI[5], FEI JIANG[6], JINYU HUANG[7], YINGXIA GUO[8] AND YANG LIU[9]
[1,2,4,5,9]*School of Electronic Information*
Dongguan Polytechnic, Dongguan, 523808 P.R. China
E-mail: 279149042@qq.com[1]; 1123261349@qq.com[2]; caowl22@163.com[4]; libin_dgpt@foxmail.com[5]; id09161819@qq.com[9]
[3]*School of Information Engineering*
Guangzhou Sontan Polytechnic College, Guangzhou, 511300 P.R. China
[+]*E-mail: gzlc666@sohu.com*
[6]*School of Modern Circulation*
Guangxi International Business Vocational College, Nanning, 530000 P.R. China
E-mail: jiangfei02@sd.taylors.edu.my
[7]*Facial Clinic*
[8]*Gastroscopy Department*
Dongguan Hospital of Integrated Traditional Chinese and Western Medicine
Dongguan, 523000 P.R. China
E-mail: 764601231@qq.com[7]; 1197104552@qq.com
[11]*Department of Public Relations*
Guangdong Only Network Science and Technology Co., Ltd., 523000 P.R. China
E-mail: 279149042@qq.com

The safety of people's lives and property is the primary factor for the success of urban construction. Therefore, in order to better maintain social stability and harmony, relying on computer technology to effectively detect violence and to make decision support has important theoretical and practical significance. Aiming at the shortcomings of traditional manual design feature extraction methods, this paper proposes a super automatic violence detection method based on the combination of Deep Learning and trajectory in AI systems. Firstly, aiming at the problem of complex time and high accuracy of traditional manual feature extraction, a deep spatiotemporal violence detection method based on three-dimensional convolution and trajectory in AI systems is proposed. We improve the IDT algorithm to extract the target trajectory, and carry out three-dimensional convolution and pooling operation to calculate the deep-seated temporal and spatial information in the video frame, so as to realize peer-to-peer detection in AI systems. Secondly, in order to further improve the acquired deep-seated time and space information and utilization rate and achieve high detection rate, the feature fusion of double stream convolution and three-dimensional convolution is proposed, and the feature extraction of continuous video frame sequence is carried out by three-dimensional convolution neural network (C3D), which can effectively extract the fusion feature information of time and space in the classification layer, so as to obtain the final classification result. Finally, in order to solve the problem of too deep network level and slow convergence, dense convolution is introduced, which reduces the parameters of the network model and time complexity. Experimental results show that compared with other mainstream algorithms, this method is more effective and stable, and can be applied to the detection of violent abnormal behavior in video. Meanwhile, the method proposed in this paper has important theoretical value and practical significance for decision support of video surveillance system in AI systems.

Keywords: AI, decision support, violence detection, machine learning, convolutional neural network, IDT algorithm

Received May 25, 2022; revised June 20 & July 1, 2022; accepted August 11, 2022.
Communicated by Mu-Yen Chen.
[+] Corresponding author.

1. INTRODUCTION

With the development of the international situation, terrorist forces in some areas are rising day by day, and violent incidents occur from time to time [1, 2]. Therefore, security construction has become particularly important. Especially at present, our country is building "Safe City". With the installation of monitoring equipment in various public places, it is possible to use intelligent video processing technology to detect and analyze violence in time, which is also a research hotspot focused by scholars. Intelligent video processing technology integrates vision technology, image analysis and processing technology and AI technology to describe, analyze and understand the behavior in the monitoring picture, and realize the automatic early warning function. Therefore, the core of intelligent video processing technology lies in video analysis technology. On the one hand, analyze the process of the event and alarm the abnormal behavior in time, so as to avoid the development of the situation in a more serious direction. On the other hand, it provides convenience for surveillance personnel to quickly retrieve and locate target segments in a large amount of video data [3]. With the successful application of Deep Learning in various fields, many scholars began to try to use Deep Learning to solve the problem of abnormal behavior in video, especially violence detection. Firstly, the continuous improvement of the performance of hardware devices such as GPU solves the problem of large-scale calculation of deep neural network model, so as to realize the analysis of abnormal behavior in video monitoring. In addition, the development of the Internet, the construction of major video websites and the construction of multiple public data sets provide enough high-resolution training samples for the training of deep neural network detection, and improve the performance and accuracy of the training model.

Based on people's increasingly urgent needs for public safety and life and property safety, this paper uses computer vision and Deep Learning methods to study violence detection methods based on surveillance video [4, 5]. Our research goal is to realize intelligent and automatic monitoring video violence detection, analyze and understand the violence in the video picture, and give timely early warning of violence, so as to reduce the harm of violence and reduce the loss of people's life and property. The main research contents of this paper include: (1) aiming at the problem of complex time and high accuracy of traditional manual feature extraction, a deep spatiotemporal violence detection method based on three-dimensional convolution and trajectory is proposed. We improve the IDT algorithm to extract the target trajectory, and carry out three-dimensional convolution and pooling operation to calculate the deep-seated temporal and spatial information in the video frame, so as to realize automatic peer-to-peer detection; (2) In order to further improve the acquired deep-seated time and space information and utilization rate and achieve high detection rate, it is proposed to fuse the features of double stream convolution and three-dimensional convolution, and three-dimensional convolution neural network (C3D) extracts the features of continuous video frame sequence, which can effectively extract the fusion feature information of time and space in the classification layer, so as to automatically obtain the final classification result; (3) In order to solve the problem of too deep network level and too slow convergence, dense connected convolution layer is introduced, so as to reduce the parameters of the network model and reduce the time complexity. The research content of this paper has important theoretical and practical significance for China's video surveillance system to enter intelligent automation as soon as possible. It

has great economic value to escort the construction of "safe city".

The study is arranged as follows. Section 2 introduces some related work similar to our algorithm, including traditional violence detection technology, violence detection technology based on temporal and spatial characteristics and violence detection technology based on deep learning. Section 3 briefly introduces the CNN model and the framework model of this method. The feature extraction method based on depth trajectory is introduced in Section 4. Section 5 describes the experimental setup. Section 6 introduces the experimental results and analysis. The last section is the conclusion.

2. RELATED WORK – DEEP LEARNING VIOLENCE DETECTION TECHNOLOGY

Ding [6] first proposed a violence detection method based on 3D convolutional neural network in 2017. In this method, the convolution kernel is transformed from two-dimensional to three-dimensional, and the video features are extracted directly without a priori processing. Compared with the previous methods, the detection effect of this method is further improved. Dong [7] introduced the acceleration characteristics of optical flow field into the multi flow neural network model, and input the extracted multi-source dynamic information into LSTM. However, due to the complex structure, the execution efficiency is low. In recent years, Deep Learning has made breakthroughs in artificial intelligence [8] and other fields [9]. Deep Learning has three benefits. The first point is that the network structure of Deep Learning may increase the depth of the network through the adjustment of parameters. At the same time, we have accumulated rich training experience to reduce gradient loss. Secondly, the era of big data provides enough training samples for deep neural network model training, and the data is not easy to over fit. Third, with the continuous improvement of hardware technology, the emergence of high-performance GPU reduces the difficulty of network model training. This paper is based on the candidate video generation model of 3D Convolution Neural Network and the DEC3D Network positioning model in the second stage. By extracting the characteristics of spatial and temporal sequence, the C3D Network model can perform convolution operation in space and deconvolution operation in temporal sequence, and realize the accurate positioning of violence temporal to the frame level to improve the retrieval accuracy of target behavior in long temporal sequence video.

Before Deep Learning, IDT algorithm is one of the best machine learning algorithms applied to behavior detection and recognition. The advantage of behavior detection and recognition algorithm based on Deep Learning is that it can extract more complete video features through network model, and the key is that it needs a large amount of training data to correct the detection model. Simonyan and Zisserman [10] combine spatial flow and temporal flow to extract the motion information of the object by overlaying the images in the video, and finally get the classification result. Compared with IDT algorithm, this method has obvious progress. Wang [11] proposed time periodic network (TSN) on the basis of predecessors. In view of the shortcomings of the traditional dual flow network in deep feature extraction, the model introduces multiple dual flow convolution layers into the network and fuses the motion information features of multiple time series. LAN *et al.* [12] weighted the short-term motion information and integrated it into the TSN network to

improve the accuracy of detection. Zhou *et al.* [13] added three full connection layers to the TSN network to adjust the weight of the length of multiple video frames to obtain the classification results. Tran *et al.* [14] proposed that the stack of continuous video frames is used as the input of 3D convolution neural network and convoluted in video blocks or stacked cubes, which improves the performance of the algorithm and is better than the dual stream convolution algorithm in terms of speed and recognition rate. Therefore, three-dimensional convolution algorithm has great research value. Xu *et al.* [15] proposed the R-C3D network, integrated the regional concept and RCNN idea into the C3D network, carried out three-dimensional convolution operation first, then RCNN convergence and collected candidate areas, and finally classified and regressed the boundary. The model can detect any length of video end-to-end, with fast detection speed and good applicability. Qiu *et al.* [16] proposed P3D network based on C3D, which can extract the behavior characteristics in spatial flow and time flow at the same time. Carreiar *et al.* [17] integrated the idea of dual flow and inflation to form a dual flow inflatable 3D network. In 2018, Feichtenhofer proposed [18] to apply two parallel 3D convolutional neural networks with different volumes to the same video clip, forming a new fast slow network structure. The network has made outstanding achievements in dynamics-400 [19], AVA [20] and dynamics-600 [19].

Two stream and three-dimensional convolutional neural networks have excellent modeling ability in long and short-term memory (LSTM) [21]. LSTM has been studied and supported by many scholars. Long *et al.* [22] conducted in-depth research on the mechanism of attitude attention in Du *et al.* [23]. They combined LSTM and CNN, which is very effective in temporal and spatial feature extraction. Wang [24] and Hinton *et al.* [25] combined local operation with unsupervised learning to realize Boltzmann machine model. The model can quickly learn the action features in video. The behavior detection and recognition algorithm based on Deep Learning mainly includes two stream convolution and three-dimensional convolution model. However, although the current three-dimensional convolution model has advantages in speed, its accuracy is not high. In general, compared with machine learning algorithm, deep neural network model algorithm has higher performance and has great advantages in dealing with complex background and large class changes.

3. CNN MODEL

The traditional violence detection algorithm is mainly based on the behavior characteristics designed by hand, the design process is complex, and cannot well describe the violence, especially in the case of complex background or occlusion. Considering the advantages of deep neural network in behavior detection and recognition, this paper adopts the combination of manual features and deep features. When extracting features, this paper integrates the idea of IDT algorithm into VGGNet model to realize the feature extraction of violent behavior in continuous time. Then, the extracted fusion features are input into support vector machine to realize the classification of violence. Finally, our method is verified on three public data sets. The accuracy of our method in three public data sets is 93.2%, 93.9% and 98.7% respectively.

This method effectively combines artificial features with deep features and trajectory features, and uses IDT and VGGNet deep neural network to extract temporal and spatial

features in video, which can improve the accuracy of violence detection, and is applicable to occluded scenes, crowded scenes and clear single scenes. At the same time, the method is tested on NVIDIA Tesla k30m GPU with 8GB video memory by MATLAB. The data frame processing speed in video is 42 frames per second, which takes into account the real-time and robustness, and the speed is fast. Based on the collected public data set, a self-made HD video data set is proposed to evaluate the robustness of the proposed violence detection method in various scenes and different resolutions. Through a large number of experiments, the parameters of the proposed violence detection algorithm are constantly modified, and compared with some existing methods, and finally a robust and accurate violence detection algorithm is obtained. The framework of violence detection system based on the combination of deep feature and IDT trajectory includes training stage and testing stage. The overall flow chart of the system is shown in Fig. 1.

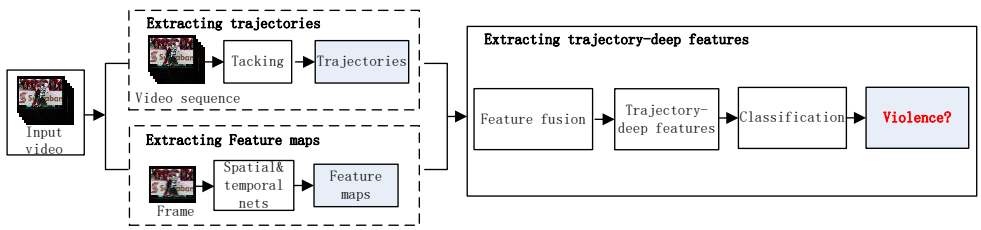

Fig. 1. Feature extraction method flow based on combination of deep feature and trajectory.

As shown in Fig. 1, after the video input, the deep feature extraction and trajectory feature extraction stages are performed. In the training stage, on the one hand, we track the optical flow according to the IDT algorithm; On the other hand, double stream convolution is used for time and space, and then fused with trajectory extraction to obtain a mature VGGNet model. In the test phase, the test video is input into the dual flow trajectory network, and the dense extraction is carried out according to the SIFT interest points, and the optical flow field is tracked to obtain the deep trajectory features, which are used as the input of support vector machine, and finally the classification results of violence are obtained. The specific design idea of the algorithm is introduced below.

3.1 CNN Model Based on Spatial Features

The two-dimensional information in the image is called spatial feature. Obviously, the key is to extract the target and background information existing in the video. Violence usually occurs in specific scenes. People usually judge violence based on the relationship between people and the background in the video. For example, if a person holding a knife is about to stab another person, the violence is related to the knife. The change of scene and the position and posture between people can be used as the basis for judging violence. At present, two-dimensional convolution model has strong advantages in processing the spatial features of images. Therefore, this paper uses VGGNet to extract behavior and scene features. VGGNet has five parameter groups and four convolution levels. In order to better extract video features from still images, this paper improves VGGNet, changes the 19 layer network structure to 21 layers, and adds two convolution layers, as shown in

Table 1. In which the symbol kernel refers to the convolution core size of each layer, the stripe refers to the sliding step size of convolution, the channel refers to the number of channels of each layer, and the ratio refers to the reduction after convolution of each layer Scale, conv for convolution layer, pool for pooling layer, FC for full connection layer. The model has been pre-trained in UCF101 video database, and the input is RGB still image (224×224×3). As shown in Fig. 1, CNN model is trained for extracting Spatial-Temporal features between the object and the scene in the video frame. Therefore, based on the original model, this paper deletes two full connection layers to better extract key features. For video V, the convolution feature layer can be represented by the following formula:

$$F(V) = \{F_1^s, F_2^s, \ldots, F_M^s\}. \tag{1}$$

Among them, the feature layer of the mth ($m \in \{1, 17\}$) spatial network is $F_m^s \in Z^{H_m \times W_m \times L \times N_m}$. Its height and width are H_m and W_m respectively. The length of a video frame is L. The number of channels is N_m.

Table 1. Structural parameters of neural network model.

Layer name	Convolution kernel parameters			Output parameters		
	number	size	step	length	width	depth
Input-0	–	–	–	224	224	3
Conv-1-2	64	3	1	224	224	64
Pool-3	–	2	2	112	112	64
Conv-4-5	128	3	1	112	112	128
Pool-6	–	2	2	56	56	128
Conv-7-8-9	256	3	1	56	56	256
Pool-10	–	2	2	28	28	256
Conv-11-12-13	512	3	1	28	28	512
Pool-14	–	2	2	14	14	512
Conv-15-16-17	512	3	1	14	14	512
Pool-18	–	2	2	7	7	512
Fully-19-20	–	–	–	1	1	4096
Fully-21	–	–	–	1	1	1000
Output-22	–	–	–	1	1	3

3.2 CNN Model Based on Temporal Characteristics

Time feature refers to the information in three-dimensional space-time. For the dynamic actions in violence, such as kicking or hitting, this paper uses time characteristics to describe them. When building the behavior model of running objects, time characteristics can effectively the state and direction of moving objects. The advantage of dual stream architecture in the field of motion description is that it can capture the timing of motion behavior in video. In this paper, we use the velocity of moving objects, that is, optical flow, to input into the neural network model to obtain time information. Optical flow is similar to the continuous change of motion information on the human visual membrane. It expresses the state of the moving target by recording the direction and speed of the target in consecutive frames in the video. Optical flow field is the change of motion state infor-

mation of the same target in two consecutive videos. We compare the mainstream optical flow processing algorithms, and finally choose farneback optical flow extraction algorithm with good performance and efficiency. This algorithm is used to extract the optical flow information in the X and Y directions in the dense optical flow field. The X direction contains the horizontal velocity field of the pixel, and the Y direction contains the vertical velocity field of the pixel. In this paper, the optical flow fields in two directions are superimposed to form a 224×224×21 dimensional VGGNet model structure. The architecture of the model is the same as the spatial feature extraction model summarized above. The model structure is shown in Table 2. When the model is trained, it can be used to extract time features. For video V, the convolution feature layer can be expressed by the following formula:

$$F(V) = \{F_1^t, F_2^t, \ldots, F_M^t\}. \tag{2}$$

Among them, the feature layer of the mth ($m \in \{1, 17\}$) spatial network is $F_m^t \in Z^{H_m \times W_m \times L \times N_m}$. Its height and width are H_m and W_m respectively. The length of a video frame is L. The number of channels is N_m.

4. TRAJECTORY EXTRACTION

4.1 Principle of Trajectory Extraction Algorithm

In 2011, Wang *et al.* [26] proposed DT algorithm, which was more accurate than other algorithms at that time. Then in 2013, in order to eliminate the influence of camera motion, Wang *et al.* [27] proposed IDT algorithm based on DT algorithm. The process of the algorithm is as follows: Firstly, each frame of the video is divided equally by eight scales, and then each scale is sampled according to a fixed pixel step size to filter out the points of interest with smaller eigenvalues of the autocorrelation matrix. The dense optical flow field of the current frame is calculated after the points of interest are obtained by dense sampling. The optical flow field is a two-dimensional vector field, which represents the instantaneous gray change rate of the pixel and contains the instantaneous velocity information of the target. Therefore, the dense sampling points obtained by optical flow tracking can be used to form the trajectory, as shown in Eq. (3). Where (x_t, y_t) is the position of the point of interest in the current frame t, and (x_{t+1}, y_{t+1}) is the position of the next frame. M is the median filter core of 3×3 and ω_t is the optical flow field.

$$G_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(x_t, y_t)} \tag{3}$$

Then the points of interest in each frame are connected in series to form a dense track (G_t, G_{t+1}, G_{t+2}, …). in order to prevent the occurrence of drift, the length of the track is limited to 15 frames. After the trajectory is obtained, the shape descriptor of the static trajectory is obtained by making the difference between the two frame trajectory coordinates, that is, the sequence (ΔG_t, …, ΔG_{t+L-1}) of the displacement vector. The shape descriptor of the dense trajectory itself is 30 dimensions. Considering the camera jitter, IDT algorithm integrates the improved RANSAC operator into DT algorithm to calculate the projection matrix, so as to avoid the redundant background optical flow caused by camera jitter. See Algorithm 1 for IRANSAC algorithm.

Algorithm 1: IRANSAC algorithm

1. BEGIN
2. Suppose that the grayscale image at time t and time $t + 1$ is $image_t$ and $image_{t+1}$ respectively
3. Calculate formula $image_{t+1} = m \times image_t$ to obtain the projection change matrix M
4. Using M^{-1}, the image gray level $Image_{t+1}^{wrap}$ of $t + 1$ without camera motion is obtained
5. The optimized optical flow is obtained by calculating the difference between $Image_{t+1}^{wrap}$ and $Image$
6. END

In the above Algorithm 1, firstly, time and time gray image are assumed; Then, the gray value of the image without camera motion is obtained by calculating the projection change matrix; Finally, the optimized optical flow is obtained by calculating the difference between the gray level and the image.

In the IDT algorithm, assuming there is a video V, $BT(V)$ is the behavior track of the target in the video, as shown below:

$$BT(V) = \{BT_1, BT_2, \ldots, BT_k\}. \tag{4}$$

Where k represents the total number of tracks extracted from the video, and T_k represents the kth track, as follows:

$$BT_k = \{(x_1^k, y_1^k, z_1^k), (x_2^k, y_2^k, z_2^k), \ldots, (x_i^k, y_i^k, z_i^k,), \ldots, (x_i^k, y_i^k, z_i^k)\}. \tag{5}$$

Where (x_i^k, y_i^k, z_i^k) represents the position of the point on the kth track in the ith video frame, and L is the length of a track.

4.2 3D Trajectory Deep Feature

Based on the above IDT algorithm integrated with RANSAC operator, the trajectory of video frame is extracted, and the trajectory information such as hog is obtained. Next, entering the feature extraction stage, this paper uses VGGNet to extract deep-seated information from IDT track, that is, deep track features. The spatial and temporal characteristics of the object contain the information of the object. Since the deep spatiotemporal features can only be obtained in the original image, not in the eight scales in IDT algorithm. Therefore, this paper improves the IDT algorithm, abandoning 8 scales and tracking only the images in the original scale, so as to speed up the speed of trajectory extraction. In addition, in order to ensure that the size of the feature layer is the same as the range of track extraction, the position of the track needs to be reduced according to the scale in Table 1. Suppose there is a video V, after extracting the deep track feature, track T and feature layer C are obtained in turn. Feature layer C includes temporal features and spatial features. The above feature extraction program is based on 13 convolution layers, and the calculation formula is as follows,

$$E(T_k, C_m^n) = \sum_{l=1}^{L} (x_l^k \times ratio, y_l^k \times ratio, z_l^k). \tag{6}$$

In the above formula, $E(T_k, C_m^n)$ is the deep trajectory feature extracted by our algorithm, and the proportion ratio of the *m*th feature layer is *ratio*.

5. EXPERIMENT SETUP

(1) Dataset

In order to investigate the accuracy of our algorithm, we conducted experiments on authoritative violence data sets, namely hockey data set [12], crowd violence data set [9] and our own high-definition fighting video.

Fig. 2. (a) Hockey dataset; (b) Homemade HD dataset; (c) Some video samples in crowd violence dataset.

(a) Hockey dataset

The hockey game consists of 1000 players with a low video resolution of 360×228, about 30 minutes, 60 frames per second, a total of 1.8 million frames. Some Samples are shown in Fig. 2 (a). Among them, 500 videos contain violence, which is normal game or standing action. This video is very challenging because it has a complex background, including occlusion scenes and other changes.

(b) Self-made data set in this paper

At present, the resolution of most public test data sets is not high, and video violence detection technology can detect high-definition video. Therefore, in order to verify the effectiveness of this algorithm in high-resolution video, this paper makes a high-definition violence format video data set, including 100 violence videos and 100 non-violence videos. A video shot by two or more people, usually with a high resolution of 1080720. Some Samples are shown in Fig. 2 (b).

(c) Crowd violence dataset

These videos are from real scenes on authoritative websites, with a total of 250 and a resolution of 320×240 pixel video. Among them, there are 125 violent videos and 125 non-violent videos respectively. This video contains a lot of crowd scenes, almost all of them are crowded people, so the recognition task is challenging. Some Samples are shown in Fig. 2 (c).

(2) CNN Model Training Scheme Based on Spatial Features

The training sample of CNN model based on spatial features adopts Hockey Fight video. The training server is xeone-E6, 8GB GPU, and the development platform adopts Caffe (Jia *et al.* 2014) open source framework under Linux system. In the algorithm test, the training batch value is 50 and the SGD value is 0.9. The pixels of the video frame are adjusted to 330×256, 224×224 pixel frames are used as input data. Firstly, the VGGNet model is pre-trained. Input the UCF101 data into the model for pre-training, and then take the hockey game data set as the training set to optimize the model on four GPUs. Based on the learning rate of 0.001, we conducted 10 rounds of training with 500 iterations per round. In order to verify the correctness of the model, we used 300 Hockey videos as a test, and the recognition rate reached 91.6%.

(3) CNN Model Training Scheme Based on Time Series Feature

In this paper, the optical flow field is used as the input data of the algorithm. We superimpose 10 optical flow fields and input them into the network model based on spatial features. The superposition calculation of optical flow field adopts Farneback algorithm. Based on the learning rate of 0.006 and 200 hockey videos, 224×224×20 sub regions are trained for 100 iterations, and finally 87% recognition accuracy is obtained. In addition, this paper tests the combination of temporal and spatial feature model and spatial feature model, and the best recognition accuracy is 93.2%.

(4) Feature Coding Process

The trajectory of the moving target in the video changes (including direction, state, *etc.*). Based on this, the dimensions of deep trajectory features extracted from each video are different. In order to ensure the uniqueness of deep trajectory feature dimension, it is necessary to encode the deep trajectory feature in each video. In recent years, Fisher vector [28] has become an effective feature coding method. In this paper, Fisher vector coding is compared with bow sparse matrix model. After testing, it is found that Fisher vector coding is better than sparse matrix model in improving the dimension of data frame, and is more suitable for image classification and recognition. Therefore, this paper selects Fisher vector for feature coding. See Algorithm 2 for Fisher vector coding process.

Algorithm 2: Fisher vector coding process

Input:

– Image features $I = \{i_t \in S^D, t = 1, \ldots, T\}$

– Gaussian mixture model $(GMMs)\lambda = \{\omega_k, \mu_k, \sigma_k, k = 1, \ldots, K\}$

– Fisher vector coding features $\mathcal{G}_\lambda^I \in S^{K(2D+1)}$

Begin

1. Initialize K value.
2. Based on the image feature $I(I_i \in I)$, calculate value λ.
3. Using I and a priori parameters λ to obtain the Fisher vector coding features.
4. Repeat Steps 2 and 3 for the images in the training set until all Fisher vector training sets are obtained.
5. SVM or other classifiers are used for training and classification.

End

Firstly, the image and Gaussian mixture model are used as inputs; Then the *K* value is initialized, and the prior parameter value is calculated according to the image characteristics; Then the vector coding feature is calculated according to the prior value; Repeat the above steps until you get all the training sets. Finally, the classifier is used for classification.

The training of GMMs depends on the characteristics of lower dimension, while the dimension of deep trajectory is higher, which does not meet the training requirements, so it is necessary to reduce the dimension appropriately. Therefore, before feature coding, this paper uses PCA to reduce the dimension to D dimension, which is introduced in Subsection (5). Then, in the first step of the algorithm, the K value is initialized to 512. 869000 features are obtained from the training, which will be used as the input of SVM classifier. The C value of the classifier is set to 2.

(5) Feature Dimension Reduction

PCA dimensionality reduction was initially applied to face recognition. Its principle is that in the process of data projection, the original high-dimensional to low-dimensional samples are transformed, and the original covariance matrix is sorted to obtain a new characteristic matrix. In this paper, the dimension of the deep trajectory obtained in Step 4 is reduced by using the in MATLAB tool. During the experiment, different principal component analysis dimensions were selected for testing. We test the dimensions of spatial and temporal features on conv4-2 convolution layer. As shown in Fig. 3, the size of 256 achieves the best effect in two convolution layers. Therefore, 256 dimensions are selected as the dimensionality reduction index of PCA in this paper.

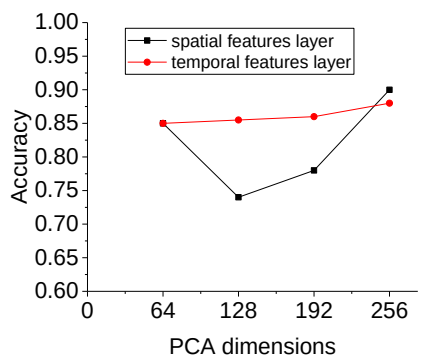

Fig. 3. The accuracy of different PCA dimensions.

(6) Experiment of Selecting Network Layer

Because the deep trajectory feature is extracted by convolution layer on the basis of optical flow trajectory. Therefore, the selection of convolution layer has a great impact on feature extraction. Therefore, a comparative experiment of convolution layer is carried out in this paper. Among the 17 convolution layers of the spatio-temporal model, we selected the data of SCL1-2, SCL2-2, SCL 3-2, SCL4-3, SCL5-3 and SCL 6-3 layers for algorithm verification. As shown in Table 2, SCL *n* is the spatial feature layer *N* (*n* = 1, ..., 6), and TFL *n* is the temporary feature layer (*n* = 1, ..., 6). For spatiotemporal networks, SCLl4-3 has the best performance, because the high-level deep features can better learn violence

scenes and motion information. Therefore, in the next experiment, we use SCL4-3 feature layer to extract the trajectory.

Table 2. Accuracy of different feature layers in spatial and temporal networks.

	Spatial Feature Layer						Temporal Feature Layer					
Accuracy	SCL 1	SCL 2	SCL 3	SCL 4	SCL 5	SCL 6	TFL 1	TFL 2	TFL 3	TFL 4	TFL 5	TFL 6
	78.8%	83%	83.8%	90.9%	77%	87.3%	70%	79.9%	83.9%	87.9%	82%	89.6%

6. EXPERIMENTAL RESULTS AND ANALYSIS

(1) Experiment on Crowd Dataset

In order to better verify our proposed algorithm, based on the dictionary size value of 100, we compare this algorithm with classical algorithms, such as MOSIFT, HOF feature, HOG, VIF, OHOF and so on.

From the results in Table 4, although the deep trajectory model in this paper is trained in the hockey video with lower resolution, the method in this paper is better than the current mainstream methods (such as MOSIFT, HOF, *etc.*). When the number of people in the video picture is large, the performance of the temporal feature descriptor decreases, which is due to the large amount of scene and motion information. Because the spatial model and time model in this paper adopt the deep trajectory feature extraction method, they are superior to the traditional algorithm in performance, and have higher recognition rate, even in the case of complex background. In addition, we also test the combination of time model and space model. As shown in Table 3, the combined model performs better than the original model. The above experimental results show that the deep trajectory feature extraction method proposed in this paper can effectively improve the detection and recognition rate of violence and reduce the false detection rate and missed detection rate.

Table 3. Experimental results on crowd dataset.

Algorithm	Handmade Features					IDT+			Spatial CNN Model	Temporal CNN Model
	OHOF	HOG	HOF	MoSIFT	VIF	Spatial + Temporal CNN	Spatial CNN	Temporal CNN		
Accuracy	82.7%	55.6%	55.9%	56.5%	81.4%	93.2%	91.5%	88.1%	66.4%	68.9%

(2) Experiments on Hockey Fight Dataset

In the classification stage, we put 270 videos on the model for training. We compare this method with the current mainstream methods, including Multi-stream CNN + LSTM [29], STIP(HOF) feature [30], MOSIFT [31], 3D CNN [32], VIF [33] and IDT [27]. As shown in Table 4, the proposed method performs very well, and the best experimental result is 98.7%.

As shown in Table 4, the recognition rate of spatiotemporal descriptors (HOG, HOF, Mo SIFT) is higher than that of VIF, because the violence perspective recorded in this data set is relatively short, the actions are more concentrated, and the pictures are not very crowded. For spatiotemporal feature descriptors, learning these scene and action information is relatively simple. VIF is mainly a scene change oriented method, so its perfor-

mance is not as good as spatiotemporal descriptor. As shown in Table 4, the efficiency and stability of the method in this paper are higher than those in the literature [7, 15, 34, 35]. This also shows that compared with single artificial feature or other deep network methods, the performance of the combination of artificial feature and deep feature in violence detection is greatly improved.

Table 4. Experiments on hockey fight dataset.

Algorithm	Handmade Features					IDT+			R-C3D CNN
	IDT	Multi-streams + LSTM	HOF	MoSIFT	VIF	Spatial + Temporal CNN	Spatial CNN	Temporal CNN	
Accuracy	91.2%	92.9%	78.9%	88.9%	70.8%	98.7%	97.5%	95.8%	90.9%

(3) Experiment on High Definition Video Dataset

In order to verify the effectiveness of our method in high-resolution video violence detection, we carried out experiments on self-made high-definition data sets. As shown in Table 5, compared with other methods, this method is still better, and the best experimental result is 93.9%.

Table 5. Experiments on high definition datasets.

Algorithm	Handmade Features					IDT+			R-C3D CNN
	IDT	HOG	HOF	MoSIFT	VIF	Spatial+Temporal CNN	Spatial CNN	Temporal CNN	
Accuracy	92.9%	83.9%	81.9%	88.9%	92.8%	93.9%	89.5%	92.3%	91.4%

The above experiments compared this method with other methods, including HOG, HOF characteristics [30], MOSIFT [31], 3D CNN [16], VIF [33], IDT [27], R-C3D [15]. Comparing the effects of the methods in Tables 3-5, the effect of this method in the self-made HD data set is still stable and better than other methods. The performance of this method in HD data set is slightly worse than that in hockey data set. This is because the training data of the model comes from the hockey data set. Other methods, such as artificial features such as spatio-temporal descriptors (HOG, HOF, MOSFIT), VIF and IDT, R-C3D and other deep models, perform better in the HD data set than the other two data sets. This is because the HD data set provides more detailed information, and the behavior features and scene information are easier to be extracted by the deep model and artificial feature model.

(4) Analysis of Experimental Results

Considering the performance of various methods on the three data sets, the deep trajectory convolution feature method proposed in this paper is the most effective and stable. When the scene in the video is complex and changeable, the recognition rate of spatiotemporal descriptors (HOG, *etc.*) will decline greatly. The VIF descriptor is suitable for crowded scenes, but its performance will decline when there are few people in the data frame. Although IDT algorithm and existing Deep Learning model are suitable for various scenarios and occupy advantages in behavior recognition, they are still not as effective as the method in this paper.

In addition, to verify the quality of the final classifier, this paper draws ROC curves on three datasets. ROC curve reveals the relationship between sensitivity and specificity through combination method, which is often used to evaluate the performance of binary classification model [36, 37]. The closer the curve is to the upper left corner, the better the performance of the classification model. The results in Fig. 4 show that our algorithm is very effective for violence detection, even when the crowd is crowded.

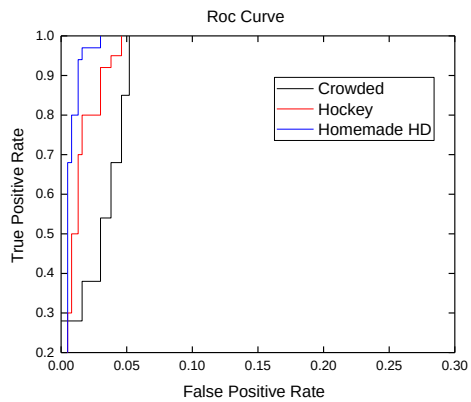

Fig. 4. ROC curve.

(5) Analysis of Experimental Efficiency

The structure of the deep convolution trajectory violence detection algorithm proposed in this paper is not end-to-end [38,39,40]. It stores the track features and the deep spatiotemporal features extracted based on the track features into memory, and then uses the classifier to classify the behavior categories. However, this method only needs about 1GB of memory. The track storage space required for video with an average frame rate of 25FPS (frames per second) is 0.88mb, and the storage space required for deep track features is 7.5mb. Trajectory feature extraction is quite time-consuming. The time consumption of this method is also concentrated here. However, due to the improvement of IDT algorithm in this paper, 8 scales are abandoned and the original scale is used for trajectory extraction, which is much more efficient than the original IDT algorithm. Matlab experiment is carried out on 8GB GPU. For 42 frames of video, the whole process of feature extraction, classification and coding only takes one second.

To sum up, in violence detection, the traditional machine learning method [41] based on manual design features not only has a complex design process, but also can not well describe violence. Deep learning models usually have deep convolution, which can automatically and effectively learn the scene and target behavior characteristics in video, but the current Deep Learning models do not have time continuity. To solve this problem, this paper proposes to integrate the IDT trajectory into the VGGNet deep convolution neural network, so that the long-term motion information of the target in the video can be extracted. Experiments show that the deep trajectory feature of this method can effectively improve the accuracy of violence detection, and the detection speed can reach 42 frames per second. At the same time, the method proposed in this paper can accurately identify the scene with occlusion change, dense crowd and clear single scene. Therefore, the method

is robust and real-time. The violence detection method proposed in this paper can help video surveillance personnel detect violence timely and accurately, give early warning, and improve the efficiency of dealing with emergencies.

7. CONCLUSION

In view of the low detection rate caused by the manual features of the traditional violence detection algorithm, this paper combines the actual application scenarios of violence detection, and carries out the following research: (1) a violence detection method based on dual stream convolution neural network and trajectory features is proposed. Aiming at the low efficiency and accuracy of abnormal behavior detection in monitoring system, the IDT algorithm is improved and an IRANSAC algorithm is proposed. The trajectories in the original image are extracted at different scales as the input of VGGNet dual flow network, and then the deep spatial-temporal features are extracted and put into SVM classifier for classification and recognition. This method combines the advantages of artificial features and deep learning features, combines the motion trajectory and convolution features, and obtains new features, which can be used as a basis for judging and improve the efficiency of violence detection. The processing speed of the whole detection process reaches 42 frames per second; (2) Self-made high-definition violence data sets, combining violence data sets in different scenes. The robustness of this method is verified in different scenes and different resolutions. Through a large number of experiments, the algorithm parameters are constantly modified, and finally a robust violence detection algorithm is obtained. Experimental results show that the average accuracy of the algorithm on three data sets is 93.4%, which is higher than the current research level. The improved IDT algorithm proposed in this paper is not optimal, and can be further optimized. The next step will continue to engage in this research work.

FUNDING

This paper is supported by Dongguan Science and Technology of Social Development Program in 2020 (2020507156694), Special Fund for Science and Technology Innovation Strategy of Guangdong Province in 2021 (Special Fund for Climbing Plan) (pdjh2021 a0944), Various Scientific Research Projects Carried Out in 2020 in Colleges and Universities of the Education Department of Guangdong Province (No. 2020KTSCX320), 2020 School-level Research Fund Key Project of Dongguan Polytechnic (2020a19), Special Projects in Key Fields of Colleges and Universities in Guangdong Province in 2021 (2021ZD ZX1093), Dongguan Science and Technology Commissioner Project (20201800500362), 2021 Special projects in Key Fields of Colleges and Universities in Guangdong Province (2021ZDZX1092), 2021 Engineering Technology Center of Colleges and Universities in Guangdong Province (2021GCZX016), Special Projects in Key Fields of Colleges and Universities in Guangdong Province in 2021 (2021ZDZX1146), 2021 Special projects in Key Fields of Colleges and Universities in Guangdong Province (2021ZDZX1119), Dongguan Science and Technology of Social Development Program in 2021 (20211800900252), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2021 (ZXYYD001), Special fund for

electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2021 (ZXF002), 2022 Guangxi University Teachers' Basic Research Ability Improvement Project' Construction of multimodal transport smart logistics system based on the new western land-sea corridor under the influence of RCEP (2022KY1252), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022 (ZXB202203), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022 (ZXC202201), Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic in 2022 (ZXD202204), Mid-term research result of the 2022 Guangxi University Teachers' Basic Research Ability Improvement Project' Construction of multimodal transport smart logistics system based on the new western land-sea corridor under the influence of RCEP (2022KY1252), Dongguan Science and Technology Commissioner Project (202018005 00362), Violence Detection Method Based on Convolution Neural Network and Trajectory (2022H21). Dongguan Social Science and Technology Development Project under Grant (2020507156696), Dongguan Science and Technology of Social Development Program (20231800900011), Special Fund for Science and Technology Innovation Strategy of Guangdong Province (Special Fund for Climbing Plan)(pdjh2023b1020) , Special fund for electronic information engineering technology specialty group of national double high program of Dongguang Polytechnic (ZXD202303), Dongguan Science and Technology Ombudsman Project in 2023 (20231800500451), Dongguan Science and Technology Ombudsman Project in 2023(20231800500282), Dongguan Science and Technology of Social Development Program (20231800903592).

REFERENCES

1. S. Amraee, A. Vafaei, and K. Jamshidi, *et al.*, "Anomaly detection and localization in crowded scenes using connected component analysis," *Multimedia Tools and Applications*, Vol. 77, 2018, pp. 14767-14782.
2. Y. Qi, P. Lou, J. Yan, *et al.*, "Surveillance of abnormal behavior in elevators based on edge computing," in *Proceedings of International Conference on Image and Video Processing, and Artificial Intelligence*, 2019, No. 11321: 1132114.
3. H. Cheng, L. Wu, R. Li, *et al.*, "Data recovery in wireless sensor networks based on attribute correlation and extremely randomized trees," *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, 2021, pp. 245-259.
4. T. Bao, S. Karmoshi, C. Ding, *et al.*, "Abnormal event detection and localization in crowded scenes based on PCANet," *Multimedia Tools and Applications*, Vol. 76, 2017, pp. 23213-23224.
5. M. Ravanbakhsh, M. Nabi, H. Mousavi, *et al.*, "Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection," in *Proceedings of IEEE International Winter Conference on Applications of Computer Vision*, 2018, pp. 1689-1698.
6. C. H. Ding, "Research on violence detection and face recognition based on deep learning," Master Thesis, Department of Computing, University of Science and Technology of China, 2017.
7. Z. Dong, J. Qin, and Y. Wang, "Multi-stream deep networks for person to person

violence detection in videos," in *Proceedings of Chinese Conference on Pattern Recognition*, 2016, Vol. 662, pp. 517-531.
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Image net classification with deep convolutional neural networks," *Communications of the ACM*, Vol. 60, 2017, pp 84-90.
9. Y. Dai, S. Wang, X. Chen, *et al.*, "Generative adversarial networks based on Wasserstein distance for knowledge graph embeddings," *Knowledge-Based Systems*, Vol. 190, 2020, No. 105165.
10. K. Simonyan and A. Zisserman, "Two-steam convolutional networks for action recognition in videos," *Neural Information Processing Systems*, Vol. 1, 2014, No. 568576.
11. L. Wang, Y. Xiong, Z. Wang, *et al.*, "Temporal segment networks: towards good practices for deep action recognition," in *Proceedings of the 14th European Conference on Computer Vision*, 2016, pp. 20-36.
12. Z. Lan, Y. Zhu, A. G. Hauptmann, *et al.*, "Deep local video feature for action recognition," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1219-1225.
13. B. L. Zhou, A. Andonian, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of European Conference on Computer Vision*, 2018, No. 831946.
14. D. Tran, L. Bourdev, R. Fergus, *et al.*, "Learning spatiotemporal features with 3D convolutinoal networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 4489-4497.
15. H. J. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 5794-5803.
16. Z. Qiu, T. Yao, and T. Mei, "Learning spatiotemporal representation with pseudo-3D residual networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 5534-5542.
17. J. Carreiar, E. Noland, A. Banki-horvath, *et al.*, "A short note about Kinetics-600," *arXiv Preprint*, 2018, arXiv:1808.01340.
18. C. Feichtenhofer, H. Fan, J. Malik, *et al.*, "Slow fast networks for videos recognition," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6201-6210.
19. J. Carriera and A. Zisserman, "Quo Vadis. Action recognition? A new model and the kinetics dataset," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4724-4733.
20. C. Gu, C. Sun, D. A. Ross, *et al.*, "AVA: A video dataset of spatiotemporally localized atomic visual actions," in *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, 2018, pp. 1012-1026.
21. F. A. Gers and J. Schmidhber, "Recurrent nets that time and count," in *Proceedings of IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 2000, Vol. 3, pp. 189-194.
22. X. Long, C. Gan, D. M. Gerard, *et al.*, "Multimodal keyless attention fusion for video classification," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Vol. 32, 2018, No. 72027209.
23. W. Du, Y. Wang, and Y. Qiao, "RPAN: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proceedings of IEEE International*

Conference on Computer Vision, 2017, pp. 3745-3754.
24. X. L. Wang, R. Girshick, A. Gpta, *et al.*, "Non-local neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, No. 7794703.
25. G. A. Hinton, "Practical guide to training restricted Boltzmann machines," *Mementum*, Vol. 9, 2010, pp. 926-947.
26. H. Wang, A. Klaeser, C. Schmid, *et al.*, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, Vol. 103, 2013, pp. 60-79.
27. H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 3551-3558.
28. J. Nchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: theory and practice," *International Journal of Computer Vision*, Vol. 105, 2013, pp. 222-245.
29. T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications*, Vol. 75, 2016, pp. 7327-7349.
30. E. B. Nievas, O. D. Suarez, G. B. García, R. Sukthankar, "Violence detection in video using computer vision techniques," *Computer Analysis of Images and Patterns*, Chapter, 2011, LNCS, Vol. 6855, pp. 332-339.
31. Y. Cheng, H. Jiang, F. Wang, *et al.*, "Using high-bandwidth networks efficiently for fast graph computation," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 30, 2019, pp. 1170-1183.
32. C. Clarin, J. Dionisio, M. Echavez, *et al.*, "Dove: Detection of movie violence using motion intensity analysis on skin and blood," in *Proceedings of Pacific Coast Softball Conference*, Vol. 6, 2006, pp. 150-156.
33. T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: real-time detection of violent crowd behavior," in *Proceedings of Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1-6.
34. A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, 2013, pp. 2782-2795.
35. N. N. Chen, X. T. Gong, Y. M. Wang, C. Y. Zhang, and Y. G. Fu, "Random clustering forest for extended belief rule-based system," *Soft Computing*, Vol. 25, 2021, pp. 4609-4619.
36. Y. G. Fu, J. F. Ye, Z. F. Yin, *et al.*, "Construction of EBRB classifier for imbalanced data based on fuzzy C-means clustering," *Knowledge-Based Systems*, Vol. 234, 2021, No. 107590.
37. Y. G. Fu, J. H. Zhuang, Y. P. Chen, *et al.*, "A framework for optimizing extended belief rule base systems with improved ball trees," *Knowledge-Based Systems*, Vol. 210, 2020, No. 106484.
38. X. Y. Li, W. Lin, X. Liu, *et al.*, "Completely independent spanning trees on BCCC data center networks with an application to fault-tolerant routing," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 33, 2022, pp. 1939-1952.
39. S. Wang, Z. Wang, K. L. Lim, G. Xiao, and W. Guo, "Seeded random walk for multiview semi-supervised classification," *Knowledge-Based Systems*, Vol. 222, 2021, No. 107016.

40. G. Liu, Z. Chen, Z. Zhuang, *et al.*, "A unified algorithm based on HTS and self-adapting PSO for the construction of octagonal and rectilinear SMT," *Soft Computing*, Vol. 24, 2020, pp. 3943-3961.
41. Y. Zhang, Z. Lu, and S. Wang, "Unsupervised feature selection via transformed auto-encoder," *Knowledge-Based Systems*, Vol. 215, 2021, No. 106748.

Jianxin Li received MS degree in School of Computing at Guangdong University of Technology. He is a Lecture at University of Dongguan Polytechnic. His research interest focuses on machine vision and behaviour recognition

Jie Liu is a Research Assistant. His research interests include behaviour recognition and high-performance computing.

Chao Li is a Lecture in Department. His research interests include behaviour recognition, cloud computing.

Wen Liang Cao is an Associate Professor. His research interests include data mining and behaviour recognition.

Bin Li is an Associate Professor. His research interests include grid computing, behaviour recognition.

Fei Jiang is an Associate Professor. Her research interests include behaviour recognition.

Jinyu Huang is a Research Assistant. Her research interests include medical data mining and behaviour recognition.

Yingxia Guo is a Research Assistant. Her research interests include medical data mining, behaviour recognition.

Yang Liu is a Research Assistant. Her research interests include behaviour recognition and high-performance computing.