J Grid Computing (2020) 18:227–237
https://doi.org/10.1007/s10723-020-09506-2

Design and Implementation of Abnormal Behavior Detection Based on Deep Intelligent Analysis Algorithms in Massive Video Surveillance

Yan Hu

Received: 13 May 2019 / Accepted: 1 January 2020 / Published online: 1 February 2020

Abstract Aiming at the high complexity of existing crowd abnormal detection models, the inability of traditional CNN to extract time-related features, and the lack of training samples, an improved spatial-temporal convolution neural network is proposed in this paper. The algorithm firstly uses the aggregation channel feature model to process the surveillance image, and selects the suspected object region with saliency characteristics. Then, the scaled correction and feature extraction are performed on the obtained suspected object region. The corresponding low-level features are obtained and input into the deep network for deep feature learning so as to enhance the representation ability. Finally, the deep feature is input into the least squares SVM classification model to obtain the final abnormal behavior detection result. The embedded chip Hi353I is used as the hardware processor to realize the real-time abnormal behavior detection effect. Our proposed deep intelligent analysis algorithm is used as abnormal Behavior Detector in the board level test. The results show that most of abnormal behaviors can be detected and the alarming message can be timely transmitted in the real-time surveillance.

Keywords Abnormal behavior detection · Deep learning · Spatial-temporal convolution · Embedded platform · Aggregate channel feature

Y. Hu (✉)
School of Fang Chenggang, Guangxi University of Finance and Economics, Nanning 530003 Guangxi, China
e-mail: hustbingke@163.com

1 Introduction

The detection of abnormal behavior of pedestrians has gradually become a popular topic in the field of intelligent video surveillance [1]. Pedestrian safety is not only related to pedestrians but also affects the surrounding traffic system for the complex indoor and outdoor traffic environment [2, 3]. However, the amount of surveillance data and Internet data are increased more rapidly in recent years. Previous safety management schemes on artificial operation can not satisfy the era of big data [4]. At present, there is a great demand for intelligent video surveillance system with efficient and reliable, and the intelligentize of video monitoring system needs to be improved [5].

The development trend of science and technology is intelligent, which uses high technology to assist human beings to complete as much work as possible. It not only can liberate people from tedious, boring and repetitive work, but also complete work that is more conducive to social development and progress, which will improve basic work efficiency and accuracy [6]. Using computers to simulate the thinking of the human brain is the top priority of intelligent development research. The first technology to be implemented is the analysis and understanding of video information. By means of the advantages of computer computing, the obtained video content is analyzed, and then valuable information is extracted. In recent years, along with the safety problem becoming increasingly outstanding, video surveillance is particularly important. Traditional video surveillance is used to detect abnormalities through the observation

of the staff. This method is not only subjective, but also wastes manpower and inefficiency. Therefore, the intelligent video surveillance system for detecting and locating abnormal behaviors of people has important research significance and commercial value.

Recently, researchers at home and abroad have done a lot of research work on crowd abnormal behavior detection and have achieved certain results. Related methods are mainly divided into two categories [7]. The first category is based on local target detection, which typically uses dynamic models to detect crowd behavior. An improved GrabCut algorithm is proposed in literature [8] to segment the target and the background of image. The algorithm combines the GrabCut algorithm and the Mean Shift algorithm to solve the problems, but the GrabCut algorithm has too many iterations, which results in slow segmentation speed and the changes in the bandwidth of the Mean Shift algorithm affect the segmentation effect. The algorithm can achieve fast segmentation and can maintain image texture and boundary well. When the target and background colors are similar, it also has better segmentation results. Once the pedestrian is detected, the feature point matching algorithm is adopted to identify the falling behavior. Literature [9] computes the Normalized Moment of Inertia of the human body's characteristic points to count the number of matching points of the human body's feature points in the two adjacent images, and compares it with a certain threshold to determine whether the human body has fallen. The experimental results show that the algorithm can effectively distinguish the falling and non-falling behaviors of the human body, but changing the shooting angle or shielding part of the limb has some influence on the recognition result. Therefore, an improved dual-line detection algorithm is proposed in literature [10] to identify abnormal human behavior in complex areas. The algorithm calculates the number of feature points in the sensitive area and combines the time that the feature point stays in the area to determine whether someone has crossed into the area and makes a stay. The experimental results show that the algorithm has a good recognition effect in the case of single-target or multi-target, and is not affected by the factors such as mutual obstruction and random movement of the target, and it can be used in various scenarios. Such methods can effectively detect and locate abnormal population behavior, but the model construction is complex and the detection rate is not high.

The second type of model is based on the global statistics, extracting some features from the whole, such as corner points, gradients, optical flow, etc., and then through the feature classification method to achieve crowd abnormal behavior detection. Aiming to take account of itself consistent representation of the moving object and the motion information of the target in the time dimension, the anomaly detection algorithm based on the super-pixels time context is proposed in literature [11]. In the feature representation phase, each frame is processed with super-pixels algorithm, and then whether or not the super-pixel belongs to the foreground is judged according to the proportion of the foreground obtained by the Gaussian Mixture model of each super-pixel. Then, according to the gray histogram and position information of the super-pixels, the closest matching super-pixel in the adjacent frame is found. And the super-pixel is represented by the Multi-scale Histogram of Optical Flow feature mean value of the closest matching super-pixels. In the phase of anomaly detection, the super-pixel features of the training set is first learned by sparse combination learning algorithm. In the test stage, determine if the super-pixel is abnormal based on the minimum reconstruction error between the super-pixel feature and dictionary combination set in the test set is determined. Although these traditional algorithms can obtain satisfactory detection results for simple scenes, the detection effect of abnormal behaviors in complex scenes is still not ideal.

With the rapid development of the theory and research of computer vision and deep learning, a series of intelligent algorithms based on deep learning have been gradually applied in the field of public security. Convolutional Neural Network (CNN) is a representative network of deep learning. Compared with traditional neural networks, the recognition effect of convolutional neural networks has been greatly improved, and has in many areas of computer vision achieved success. Therefore, based on deep learning, it is of great significance to efficiently and accurately identify abnormal behavior in video. Many scholars have focused on the research and implementation of two aspects: a combined optical flow network and a two-stream structure based on deep residual network [12, 13]. In the extraction of optical flow characteristics, literature [14] proposed an improved deep learning networks to perform optical flow estimation on moving

objects in the video. In order to accurately extract the motion information of the foreground target in the video, the concept of end-to-end optical flow is introduced in the combined optical flow network including multiple optical flow networks [15], and in addition, a subnetwork for capturing small displacements is added. To describe the changes in the details of the action. The optical flow image of the adjacent video frames is finally calculated by the combined optical flow network, and used as the input of the time flow network [16].

In design of the two-stream structure based on the deep residual network, literature [17] divides the video into two parts and adopts spatial flow network and time flow network respectively by imitating the ventral and dorsal channels of the brain to process visual signals. To achieve two-stream feature extraction, the motion information is adopted. In order to accurately identify abnormal behaviors, literature [18] proposes a 101-layer deep residual network as the spatial stream network and the time stream network to solve the problem, which prone to gradient explosion caused the degradation of the overall network. We adopt strategies such as data gain [24], migration learning [25], and Dropout to solve the overfitting problem, which enhance the learning and generalization ability. However, training a deep network requires large-scale and diverse training samples, but it is often difficult to obtain enough samples in actual human abnormal behavior detection, resulting in unsatisfactory detection results. In addition, the complexity of deep model is too high to achieve real-time detection effect.

Aiming at the high complexity of existing crowd anomaly detection methods [26–31], the inability of traditional CNN to extract time-related features, and the lack of training samples, an improved spatial-temporal convolution neural network is proposed in this paper. The algorithm firstly uses the aggregate channel feature model to process the surveillance image, and selects the suspected target region with saliency characteristics. Then, the scaled correction and feature extraction are performed on the obtained suspected target region. The corresponding low-level features are obtained and input into the deep auto-encoder network for deep feature coding so as to enhance the representation ability. Finally, the coding feature is input into the least squares SVM classification model so as to obtain the final detection result. The embedded chip Hi353I is used as the hardware processor to realize the real-time abnormal behavior detection effect. The experimental results show that the proposed method has a higher detection rate than the existing methods.

2 Abnormal Behavior Detection System

Behavior can be divided into abnormal behavior and normal behavior. Abnormal behavior is a small probability event that occurs in behavior and is part of behavior recognition. Abnormal behavior recognition research and behavior recognition research are consistent. From the occurrence of effective behavior to the end of effective behavior, the behavior can be divided into short-term behavior and long-term behavior by the angle of time. Short-term behavior is mainly the behavior that takes less time during the duration of the behavior, such as running, walking, fainting and so on. The short-term behavior is characterized by greater discriminability, and short-term behavior is difficult to subdivide to the lower level. Long-term behavior mainly refers to behaviors that have a certain time span and are composed of multiple short-term behaviors mixed in different time series. Long-term behavior is a complex behavior that can be divided down into multiple short-term behaviors. From the perspective of the number of people involved in the behavior, the behavior can be divided into single-person behavior and multi-person interaction behavior. Multi-person interaction mainly refers to the behavior that requires two or more people to interact with each other. The behavior recognition usually consists of four steps: data set, data preprocessing, feature extraction and behavior classification. Different behavior recognition processing methods will have different steps, some will be merged in some steps, and some will separate different steps in a certain step.

Regardless of the feature extraction of abnormal behavior detection, the type of anomaly detection model, the main flow of the abnormal behavior detection system is shown in Fig. 1. First, spatial and temporal segmentation of the video is performed to extract features that can describe the characteristics of the target region; then, in the training phase, the normal event is modeled; in the test phase, the abnormality of the test feature is calculated for the normal event model that has been learned so as to determine whether the behavior is abnormal according to the set abnormal threshold. Among them, the two steps of feature extraction and abnormal behavior detection model have a great influence on the detection effect of abnormal behavior.

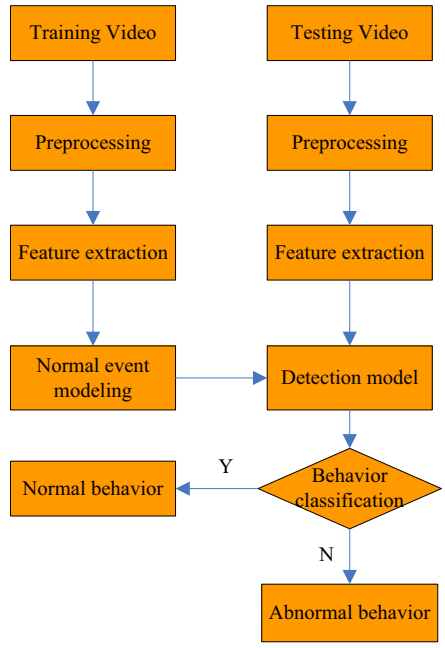


Fig. 1 Framework for abnormal behavior detection

As a kind of multi-layer neural network learning algorithm, deep learning technology can learn features through deep nonlinear network structure, and combines low-level features to form more abstract deep representations, and realize complex function approximation, so that you can learn the essential characteristics of the data set. This paper uses the deep feature to detect abnormal behavior.

3 Our Proposed Abnormal Behavior Detection Model

3.1 Deep Model and its Corresponding Symbol Description

It is Assumed that N_p positive samples I_p^i, $i \in 1, ..., N_p$ and N_q negative samples I_q^i, $\mathrm{i} \in 1, ..., N_q$ are selected in a single frame image, all samples are normalized to the same scale, and their corresponding gradient histogram features (HoG) are represented as H_p^i and H_q^i. Once the positive and negative samples are trained by the deep model, the obtained coding vectors are represented as $\left(w_p^i \in R^n, b_p^i \in R\right)$ and $\left(w_{\mathrm{q}}^i \in R^n, b_q^i \in R\right)$, respectively. In order to facilitate the subsequent model description, H_p^i, H_q^i, $\left(w_p^i, b_p^i\right)$ and $\left(w_{\mathrm{q}}^i, b_q^i\right)$ are rewritten as $x_p^i = \left[\mathrm{H}_p^{i\,\mathrm{T}}, 1\right]^{\mathrm{T}}$, $x_{\mathrm{q}}^i = \left[H_q^{i\,\mathrm{T}}, 1\right]^{\mathrm{T}}$, $y_p^i = \left[w_p^{i\,T}, b_p^i\right]^T$ and $y_{\mathrm{q}}^i = \left[w_q^{i\,\mathrm{T}}, b_q^i\right]^{\mathrm{T}}$, respectively. Once the deep features of the suspected sample x_E is extracted and encoded, the optimal regression function F can be learned by the proposed pedestrian detection model, that is $y_E = F(x_E)$, the pedestrian detection is realized in complex background.

3.2 Multi-Level Deep Feature

Although the low-level features can represent the texture information of the abnormal behavior, the acquired features are difficult to be directly used for classification due to differences in pedestrian posture, clothing, etc. At present, many abnormal behavior detection models adopt PCA [9], sparse [10], low-rank and other priors [11, 12, 19] to perform feature re-encoding so as to reduce interference noise and improve feature representation. This paper seeks to enhance the representation ability of abnormal behavior target features by deep coding of low-level features.

As for any samplesx_i, $i \in 1, ..., N$, it is given that $g_e(\theta, x)$ is coding function in DAN network, $g_d(\theta, x)$ is the corresponding decoding function, where θ is the corresponding model parameter; W_e, W_d represent the encoding and decoding matrix, respectively. For a multi-layer deep network, $g_e(\theta, x) = f_h\left(W_e^T x\right)$, $g_d(\theta, x) = f_o\left(W_d^T x\right)$ can be obtained, where $\theta = (W_e, W_d)$; the activation function of the hidden layer and the output layer can be denoted as $f_h(\cdot)$ and $f_o(\cdot)$, respectively. For any x^i, $i \in 1, ..., N$, its class label can be expressed as $l^i \in (-1, 1)$; $\{y_i\}$ is the result of the class label of all samples; therefore, the objective function of our proposed deep model can be written as follows,

$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left\|x^i - y_r^i\right\|_2^2 + h\left(1 - u^i, 0\right) + \frac{\lambda}{2}\left\|y_r^i\right\|_2^2 \quad (1)$$

where $y_r^i = g_d \cdot g_e(\theta, x^i)$ is represent as the deep feature vector of x_i; $u^i = l^i\left(y_r^i\right)^T x^i$; $h(x) = \max(o, x)$; $\lambda$is a regularization learning factor, which is used to regulate the generalization performance of deep model under different samples. According to the expression of Eq. (1) function, it can be seen that the objective optimization is to adjust Hinge Loss function on the basis of deep coding reconstruction error. In the training phase, loss

functions in inter-class and intra-class can ensure deep feature vectors y_r^i can accurately characterize samples x^i, $i \in 1, ..., N$.

The objective function represented by Eq. (1) represents a convex function. The optimization process can be carried out by the method proposed in reference [13], and is optimized by stochastic gradient descent (SGD), where the iteration update formula of parameter θ is shown as follows:

$$\theta_{n+1} = \theta_n - \alpha \frac{\partial L}{\partial \theta} \tag{2}$$

where n and α represent iteration times and learning factors, respectively. Therefore, the partial differential of Eq. (1) for parameters θ can be expressed as follows:

$$\frac{\partial \mathrm{L}}{\partial \theta} = \sum_{i=1}^{N} \left[-\left(y^i - y_r^i\right) - \mathrm{S}\left(1 - u^i\right) l^i x^i + \lambda y_r^i \right] \frac{\partial y_r^i}{\partial \theta} \tag{3}$$

where $\mathrm{S}(x) = 0.5(\mathrm{sgn}(x) + 1)$, $Sgn(x)$ represents a symbolic function. y_r^i is the deep characteristic of the output of the deep model, and its partial differential for θ can be calculated by error back propagation network [18]. It can be seen from Eq. (3) that the core step of calculation is the result obtained. Once the auto-coding network is trained, the samples can be transformed into low-dimensional deep feature vectors.

So the deep characteristics of the output of the deep model can be calculated by the error back propagation network [14]. It can be seen from Eq. (3) that the key step in calculating $\partial \mathrm{L}/\partial\theta$ is to get the result of $\partial y_r^i/\partial\theta$. Once the deep network is completely trained, the samples x_i can be transformed into low-dimensional deep feature vectors $z^i \in R^{n_c}, \forall i \in 1, ..., N$, namely, $z^i = f_h\left(W_e^T y^i\right)$.

3.3 Least Squares Support Vector Machine

The Least Squares SVM (LSSVM) is an improved model based on hyperplane two classification. Compared with the traditional support vector machine, the least squares support vector machine can overcome the short training time, the randomness of the training results and the lack of learning ability, enhance the classification accuracy of the model, reduce the calculation amount and shorten the calculation time. Given a linearly separable sample set: $\{(x_i, y_i) | \, x_i \in R^m, i = 1, 2, \cdots, n\}$, its least squares optimization model is expressed as:

$$\min\left(\frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{l}\varsigma_i^2\right) s.t.\ y_i(\mathrm{w}\cdot\Phi(\mathrm{x}_i) + \mathrm{b}) = 1 - \varsigma_i, i = 1, 2, \cdots, l \tag{4}$$

where C is a regularization parameter that balances the differences in inter-class and intra-class. Since pedestrian non-rigid objects are linearly inseparable in low-dimensional space classification, the accuracy of classification using only traditional SVM is not high. Literature [15] proposed that nuclear mapping is capable of mapping low-dimensional inseparable samples to high-dimensional separable feature spaces. Therefore, it can be seen that the optimization problem of the least squares support vector machine can be obtained by solving the linear equations, and the optimal classification surface is:

$$\min\left\{\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j K(x_i, y_i) + \sum_{i=1}^{n}\frac{\alpha_i^2}{2\gamma} - \sum_{i=1}^{n}\alpha_i y_i + b\sum_{i=1}^{n}\alpha_i\right\} \tag{5}$$

where the kernel function $\mathrm{K}(x_i, y_i)$ can use a linear kernel function, a polynomial kernel function, and so on. The kernel function can be used to directly calculate the inner product $\mathrm{K}(x_i, y_i) = \varphi(x_i)\varphi(y_i)^{\mathrm{T}}$ of high-dimensional space, but most kernel functions cannot fit all data. The least squares support vector machine transforms the objective function in the SVM model into a linear optimization problem, and improves the solution speed through linear solution, which is suitable for the real-time response requirement of abnormal behavior check in the large-scale video surveillance data [15].

3.4 Abnormal Behavior Detection and its Framework

As we all know, abnormal behavior detection algorithms mostly use a certain search strategy to generate a large number of candidate sample sets, and then use the response algorithm, such as correlation matching, detector, pattern recognition and other models to score the candidate area (Score), so as to find the best sample as the final abnormal detection result [16]. This is an exhaustive search mode with a high level of complexity. In addition, because abnormal behavior are non-rigid targets, their shape is greatly affected by factors such as scale and posture. On the one hand, only multi-scale screening targets can cover all target areas, and the complexity is too high. On the other hand, the strategy is an exhaustive search mode with considerable

complexity [17]. In order to reduce the complexity of abnormal behavior detection samples and enhance the efficiency of detection, this paper firstly uses the aggregate channel feature model to obtain the suspected object region and reduce the search time for single frame image. A large number of qualitative and quantitative simulation experiments show that the suspected target almost covers all possible target areas in the image after the processing of the aggregate channel feature model, which greatly reduces the suspected target detection time.

It is given that the pre-processed image has M suspected saliency regions, which can be expressed as $\{B^i \in R^{m_i \times n_i} | i = 1, 2, ..., M\}$. Because the scales of different suspected regions are different and the training parameters of the model are fixed, it is necessary to normalize the M suspected samples to a unified scale so as to facilitate the training of the model and the optimization of the parameters, $\{D^i \in R^{m \times n} | i = 1, 2, ..., M\}$.

Since the suspected sample D^i need to be normalized to a uniform scale, the corresponding gradient histogram feature is obtained and converted into feature vector d^i; then the trained deep model is used to obtain the deep feature vector d^i; The LSSVM classifies the deep features, and then finds the optimal abnormal behavior target, and reconstructs the classification vector c^i, where $c^i = g_d(\theta, v^i)$. Figure 2 shows the detection process of the abnormal behavior detection model proposed in this paper. It can be seen that the model can greatly reduce the sample size by preprocessing, and only needs the suspected area. In addition, deep feature enhances the representation ability of abnormal behaviors, and can refer to the accuracy of suspected behavior detection.

3.5 Hardware Architecture

The current research on intelligent video surveillance is mostly based on computers, and the video data is transmitted to the computer for processing through the monitoring terminal. The disadvantage of this is that the more numbers of monitoring terminals, the greater pressure of data processing on the computer. To solve above problem, the research content of this paper is that the video analysis is completed at the monitoring terminal based on Hisilicon Hi3531 embedded chip, and the result of analysis is transmitted to the server.

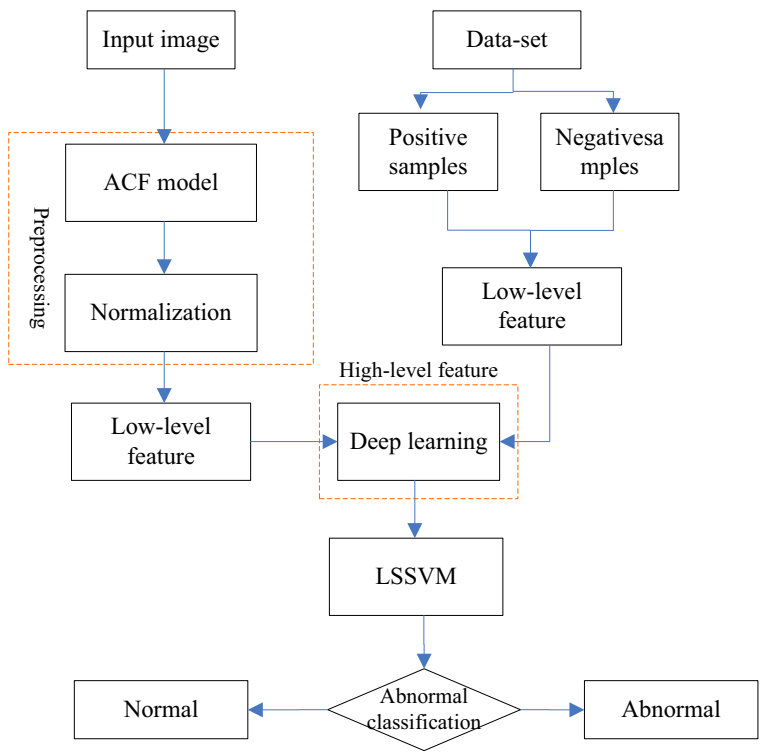


Fig. 2 Framework for abnormal behavior detection model

Firstly, the hardware parameters and interface of the chip Hi3531 and the media process platform (MPP) are introduced in detail. Then the necessity and steps of the establishment of cross-compilation and network file system are detailedly described in Hisilicon Linux development environment. Finally, this paper presents the design plan of intelligent monitoring system based on Hisilicon platform, which consists of video capture device, main processing chip Hi3531, monitoring server and video display device. Among them, the main processing chip Hi3531 completes the identification of the human fall behavior and the transmission of the recognition result to the monitoring server. At the same time, the monitoring server can receive the real-time monitoring video data compressed into format H.246.

4 Experimental Results and Analysis

4.1 Experimental Data Set

In order to effectively evaluate the performance of the deep feature abnormal behavior detection algorithm proposed in this paper, a common international detection data set UCF101 is selected. UCF101 is an action video data set built by the University of Center Florida for action classification tasks. The data set consists of 13,320 video data collected from YouTube, including a variety of action forms. The format of the video is unified during the construction of the data set.

The UCF101 dataset action video category is a more common scene in life, such as sports scenes, playing different instruments, etc., but there are few video data with abnormal behavior. In order to complete the establishment of the abnormal behavior data set, the CASIA Behavior Analysis Database is added in training. The CASIA Behavior Analysis Database has a total of 1446 video data, which are simultaneously captured by three uncalibrated cameras distributed in horizontal, oblique and overhead angles in an outdoor environment, providing experimental data for behavioral analysis. The data is divided into single-person behavior and multi-person interaction behavior. Single-person behavior includes walking, running, bending, jumping, squatting, fainting, squatting and braking. Each type of behavior has 24 people involved in the shooting, 4 times per person. Multi-person interactions include robbery, fighting, trailing, catching up and meeting. In order to verify the effectiveness of the deep network in the identification task of abnormal behavior, we first train the deep network based on the UCF101 data set with the rich action categories. After training the deep network model, the network will be transferred to the CASIA Behavior Analysis Database to complete the task of identifying and classifying abnormal behaviors.

4.2 Parameter Setup

In order to accelerate the convergence of the network, we pre-train the network on the ImageNet dataset and use the pre-trained deep network model as the improved deep network. The specific training parameters are as follows: Set the Batch_size of the UCF101 training set to 64. Due to the large data of the training set, in order to improve the performance, the data set should be trained in batches during the actual training process. The size of Batch_size represents the sample capacity contained in each batch, and also affects the training efficiency. GPU storage mode, Batch_size generally chooses 32, 64, 128, 256 and so on. In the experiment of this paper, the size of Batch_size is set to 640. Set the number of Epoch to 50. In the training process of the neural network, it is not enough to transmit only the complete data set once. In order to adjust the parameters in the network to the optimal parameters, the entire data set needs to be transmitted multiple times in the neural network, and the parameters are continuously adjusted for optimal state, so Epoch is set to 50 in this test. In the experiment, the training process can be terminated early according to the change curve of the loss value and the test accuracy.

4.3 Result Analysis

In order to verify that the proposed deep network can effectively identify abnormal behavior, the trained network model on the UCF101 dataset is transferred to the CASIA behavior database. The CASIA Behavior Database is a database provided by Chinese Academy of Sciences. It contains single-person behavior and multi-person interaction behavior. There are 14 types of behaviors, including 4 abnormal behaviors, namely fight, rob, faint and Punch, as shown in Fig. 3.

In this paper, the CASIA database is divided into training set and test set according to the ratio of 4:1. The database is constructed from the perspective of monitoring video. In order to realize the detection of abnormal behavior, a strategy for classifying each abnormal behavior is adopted. When the image data is

Fig. 3 Abnormal behavior in CASIA behavior database. **a** Fight; (**b**) Rob; (**c**) Faint; (**d**) Punch

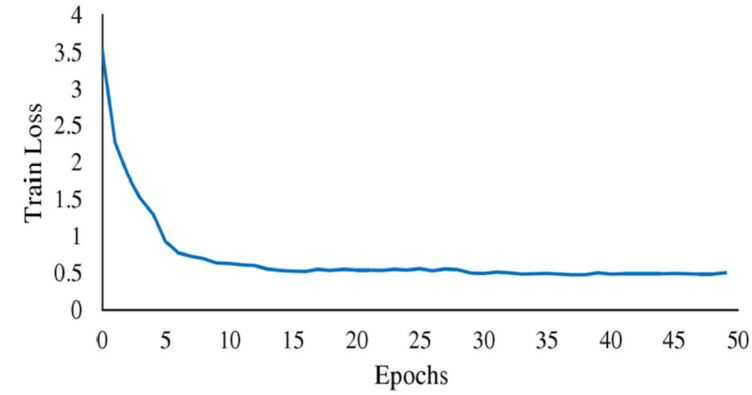
converted into a database file, only two labels are set, one is the abnormal behavior that needs to be learned, the other is the non-abnormal behavior, and the output feature of the Softmax layer is also changed to the two-dimensional feature, respectively predicting the target in the video whether the behavior is abnormal or not. We select DLHS [10], DLPD [20], CTLD [21], Yolo-SD [22], HOD-CCL [23] as comparison algorithms. During the training phase, the experimental software and hardware environment is set to: Xeon Bronze 3106-B 1.7GHz, 32GB RAM, Nvidia Geforce GTX 1080Ti, Ubuntu l6.04, 64-bit operating system; During testing, the hardware processor is the embedded chip Hisilicon Hi3531.

In order to facilitate visual analysis, Table 1 shows the accuracy of detection under different algorithms, which is convenient for visual analysis. The detection rate of the proposed algorithm is 67.79%, and the best detection result in the comparison algorithm is 65.01%. It can be seen from the quantitative results that under the same conditions, the proposed algorithm has the highest accuracy, which is 2.78% higher than the deep learning DLPD. In summary, the proposed algorithm achieves better detection results, mainly due to the direct coding of low-level histogram features, which increases the target representation ability and further enhances the generalization of the model. At the same time, the model abandons the traditional *softmax* for classification learning, but using the optimal linear optimal solution to obtain the least squares SVM classification algorithm, further improving the overall performance of the model detection.

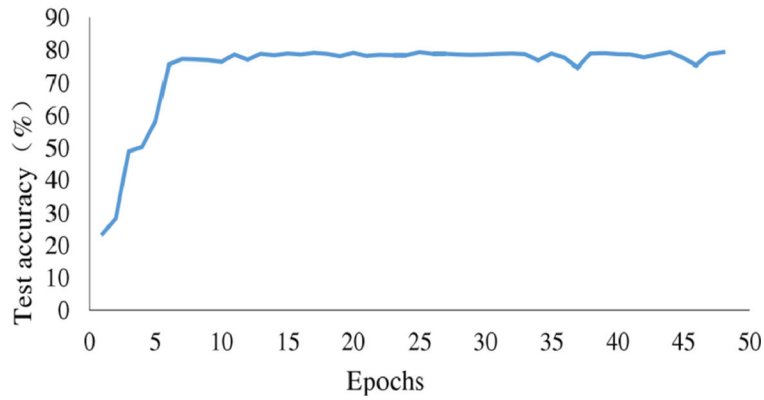
From the results in the table, the accuracy of HOD-CCL for fight, rob and faint behavior is higher than that of Yolo-SD for the four abnormal behaviors in the CASIA behavior database. Only the recognition accuracy of HOD-CCL with faint behavior is lower than that of Yolo-SD. Since the faint behavior is not repetitive, the

Table 1 Detection rate for different algorithms

Model	DLHS	DLPD	CTLD	Yolo-SD	HOD-CCL	Proposed
Fight	45.12%	57.1%	51.15%	52.34%	56.01%	67.79%
Rob	51.23%	59.27%	53.19%	48.36%	61.09%	68.04%
Faint	56.11%	57.66%	63.22%	61.06%	70.11%	72.00%
Punch	61.24%	67.21%	68.32%	63.59%	69.21%	70.13%



(a)Training Loss



(b) Testing accuracy

Fig. 4 Convergence curve in training process. **a** Training Loss. **b** Testing accuracy

key points of the faint behavior appear for a short time. After the fall, the human body no longer has dynamic information, and the surveillance distance is very far, the extracted optical flow information is very limited, and the recognition accuracy of the HOD-CCL network is low, only 56.11%. However, compared with the recognition accuracy of the Yolo-SD network, since the input information of the Yolo-SD network is an RGB frame

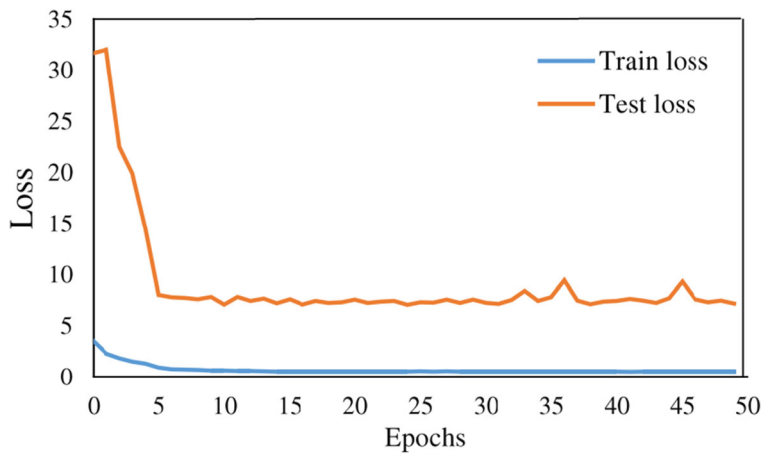


Fig. 5 Comparison of Loss value between training set and testing set

image, the recognition accuracy is high, which is 70.11%, while our result is 72.00%. This comparison fully demonstrates the effectiveness of our algorithm.

4.4 Convergence Performance Analysis

In our proposed deep network training process, the data set loss function value change trend is shown in Fig. 4a. The accuracy of the test set changes with the number of iterations as shown in Fig. 4b. As the number of iterations increases, The value of the loss function gradually decreases, the accuracy of the test set gradually increases, and the two curves tend to be stable around 15 Epoch. After 15 Epoch, the accuracy of the test set does not increase significantly, indicating that the network model has basically reached the optimal state.

During the training process of the HOD-CCL network, the loss value of the training set and the loss value of the test set increase with the number of iterations as shown in Fig. 5. It can be seen from the figure that the loss function value of the training set gradually decreases with the increase of the number of iterations. Although the loss function value of the test set is larger than the loss function value of the training set as a whole, the loss function value of the test set gradually decreases and tends to be stable, which basically converges synchronously, indicating that the network does not have an over-fitting phenomenon. The accuracy of our proposed deep network based on the UCF101 data set is compare in experiment. The accuracy of the training set is slightly higher than the accuracy of the test set. The training accuracy can reach 77.39%, and the test accuracy rate can reach 79.88%.

5 Conclusion

Aiming at the high complexity of existing crowd abnormal detection models, the inability of traditional CNN to extract time-related features, and the lack of training samples, an improved spatial-temporal convolution neural network is proposed in this paper. The algorithm firstly uses the aggregation channel feature model to process the surveillance image, and selects the suspected object region with low-level characteristics. Then, the scaled correction and feature extraction are performed on the obtained suspected target region. The corresponding low-level features are obtained and input into the deep network for deep feature learning so as to enhance the representation ability. Finally, the deep feature is input into the least squares SVM classification model so as to obtain the final abnormal behavior detection result. The embedded chip Hi353I is used as the hardware processor to realize the real-time abnormal behavior detection effect. The experimental results show that the proposed method has higher detection rate than the existing methods.

References

1. Li, C., Han, Z., Ye, Q., et al.: Visual abnormal behavior detection based on trajectory sparse reconstruction analysis[J]. Neurocomputing. **119**(16), 94–100 (2013)
2. Ko, K.E., Sim, K.B.: Deep convolutional framework for abnormal behavior detection in a smart surveillance system[J]. Eng. Appl. Artif. Intell. **67**, 226–234 (2018)
3. Xiong, G., Cheng, J., Wu, X., et al.: An energy model approach to people counting for abnormal crowd behavior detection[J]. Neurocomputing. **83**(23), 121–135 (2012)
4. Devroye, L., Wise, G.L.: Detection of abnormal behavior via nonparametric estimation of the support[J]. SIAM J. Appl. Math. **38**(3), 480–488 (1980)
5. Zhang, J., Wu, C., Wang, Y., et al.: Detection of abnormal behavior in narrow scene with perspective distortion[J]. Mach. Vis. Appl. 1–12 (2018)
6. Rasheed, N., Khan, S.A., Khalid, A., et al.: Tracking and abnormal behavior detection in video surveillance using optical flow and neural networks[C]. 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2014. IEEE Computer Society. 2014:61–66
7. Cheng G, Wang S, Guo T, et al. Abnormal behavior detection for harbour operator safety under complex video surveillance scenes[C]. 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). 2018: 28–33
8. Jeong, H., Chang, H.J., Choi, J.Y.: Modeling of moving object trajectory by spatio-temporal learning for abnormal behavior detection[C]// IEEE International Conference on Advanced Video & Signal Based Surveillance. IEEE Comput. Soc. 71–82 (2011)
9. Wang, Q., Ma, Q., Luo, C.H., et al.: Hybrid histogram of oriented optical flow for abnormal behavior detection in crowd scenes[J]. Int. J. Pattern Recognit. Artif. Intell. **30**(02), 14–23 (2016)
10. Sabokrou, M., et al.: Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. Comput. Vis. Image Underst. **172**, 88–97 (2018)
11. Bouttefroy P L M, Bouzerdoum A, Phung S L, et al. Local estimation of displacement density for abnormal behavior detection[C]. IEEE Workshop on Machine Learning for Signal Processing. 2008:29–2036
12. Lee J J, Kim G J, Kim M H. Trajectory extraction for abnormal behavior detection in public area[C].

International Conference & Expo on Emerging Technologies for A Smarter World. 2013:212–218
13. Li C L, Hao Z B, Li J J. Abnormal behavior detection using a novel behavior representation[C]. International Conference on Apperceiving Computing & Intelligence Analysis. 2011: 54–60
14. Xiang J, Fan H, Xu J, et al. Abnormal behavior detection based on spatial-temporal features[C]. International Conference on Machine Learning And Cybernetics, 2013: 871–876
15. Avinash, R., Vinod, P.: Tucker tensor decomposition-based tracking and Gaussian mixture model for anomaly localisation and detection in surveillance videos[J]. IET Comput. Vis. **12**(6), 933–940 (2018)
16. Hirsch, M., et al.: A two-phase energy-aware scheduling approach for CPU-intensive jobs in mobile grids. J. Grid Comput. **15.1**, 1–26 (2017)
17. Xia, K.-j., Yin, H.-s., Zhang, Y.-d.: Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and ResNet models combined with SIFT-Flow algorithm. J. Med. Syst. **43**, 2 (2019). https://doi.org/10.1007/s10916-018-1116-1
18. Xia, K.J., Yin, H.S., Wang, J.Q.: A novel improved deep convolutional neural network model for medical image fusion [J]. Clust. Comput. **3**, 1–13 (2018)
19. Qianyin J, Guoming L, Jinwei Y, et al. A model based method of pedestrian abnormal behavior detection in traffic scene[C]. IEEE International Smart Cities Conference, 2015: 1–6
20. Rezazadegan F, Shirazi S, Upcrofit B, et al. Action recognition: From static datasets to moving robots[C]. 2017 IEEE International Conference on Robotics and Automation (ICRA). 2017:3185–3191
21. Xia, K., Yin, H., Qian, P., Jiang, Y., Wang, S.: Liver semantic segmentation algorithm based on improved deep adversarial networks in combination of weighted loss function on abdominal CT images. IEEE Access. **7**, 96349–96358 (2019)
22. Qian, P., Xi, C., Xu, M., Jiang, Y., Kuan-Hao, S., Wang, S., Muzic Jr., R.F.: SSC-EKE: semi-supervised classification with extensive knowledge exploitation. Inf. Sci. **422**, 51–76 (2018)
23. Qian, P., Sun, S., Jiang, Y., Kuan-Hao, S., Ni, T., Wang, S., Muzic Jr., R.F.: Cross-domain, soft-partition clustering with diversity measure and knowledge reference. Pattern Recogn. **50**, 155–177 (2016)
24. Kajian, X.I.A., Jiangqiang, W.A.N.G., Yue, W.U.: Robust Alzheimer Disease classification based on Feature Integration Fusion Model for Magnetic[J]. J. Med. Imag. Health Inf. **7**, 1–6 (2017)
25. Fang, W., Beckert, U.: Parallel tree search in volunteer computing: a case study. J. Grid Comput. **4**, 1–16 (2017)
26. Rui, Y., Bing, L., Ye-Lin, H., et al.: A method for abnormal behavior recognition based on deep learning[J]. Journal of Wuyi University (Natural Science Edition). **12**(2), 112–122 (2018)
27. Fang, Z., Fei, F., Fang, Y., et al.: Abnormal event detection in crowded scenes based on deep learning[J]. Multimed. Tools Appl. **75**(22), 14617–14639 (2016)
28. Kovács, J., Kacsuk, P.: Occopus: a multi-cloud orchestrator to deploy and manage complex scientific infrastructures[J]. J. Grid Comput. **16**(1), 1–19 (2017)
29. Qian, P., Zhou, J., Jiang, Y., Liang, F., Zhao, K., Wang, S., Kuan-Hao, S., Muzic Jr., R.F.: Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes. IEEE Access. **6**, 28594–28610 (2018)
30. Ramon-Cortes, C., et al.: Transparent orchestration of task-based parallel applications in containers platforms. J. Grid Comput. **16.1**, 137–160 (2018)
31. Tighe, M., Bauer, M.: Topology and application aware dynamic VM Management in the Cloud. J. Grid Comput. **4**, 1–22 (2017)