Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.020240
Article

An Improved Two-stream Inflated 3D ConvNet for Abnormal Behavior Detection

Jiahui Pan[1,2,*], Liangxin Liu[1], Mianfen Lin[1], Shengzhou Luo[1], Chengju Zhou[1], Huijian Liao[3] and Fei Wang[1,2]

[1]School of Software, South China Normal University, Foshan, 528225, China
[2]Pazhou Lab, Guangzhou, 510330, China
[3]Department of Mechanical, Materials and Manufacturing Engineering, University of Nottingham, Nottingham, NG7 2RD, United Kingdom
*Corresponding Author: Jiahui Pan. Email: panjh82@qq.com
Received: 16 May 2021; Accepted: 18 June 2021

Abstract: Abnormal behavior detection is an essential step in a wide range of application domains, such as smart video surveillance. In this study, we proposed an improved two-stream inflated 3D ConvNet network approach based on probability regression for abnormal behavior detection. The proposed approach consists of four parts: (1) preprocessing pretreatment for the input video; (2) dynamic feature extraction from video streams using a two-stream inflated 3D (I3D) ConvNet network; (3) visual feature transfer into a two-dimensional matrix; and (4) feature classification using a generalized regression neural network (GRNN), which ultimately achieves a probability regression. Compared with the traditional methods, two-stream I3D feature extraction technology is better able to extract visual features and retain the optical flow and red, green and blue (RGB) information in the video. Probabilistic regression technology is better able to quantify data and provide a more intuitive visual experience. The experimental results on 50 detection cases from the UCF-Crime dataset show that the developed model obtains a high average abnormal behavior recognition accuracy. The improved I3D models obtain an average accuracy of 93.7% based on UCF-101, which outperforms the state-of-the-art methods, verifying the robustness and effectiveness of our approach.

Keywords: Abnormal behavior detection; two-stream inflated 3D convnet network; general regression neural network; python

1 Introduction

With the continuous increase in the population base, there are still a large number of violent events in underdeveloped countries, such as theft, robbery, and explosions. How to maintain social public security has become a national focus. In most public places, people increase the number of cameras and conduct real-time detection through intelligent monitoring systems to ensure social safety. Therefore, studying intelligent surveillance systems is highly important. It is anticipated that intelligent surveillance systems will

eventually be able to analyze surveillance videos using computer vision [1] and image processing and perform identification through artificial intelligence technologies. Then, measurements can be conducted according to the analysis results.

This study aims to propose improved two-stream inflated 3D ConvNets using probability regression to identify abnormal behaviors in surveillance videos. The inflated 3D (I3D) network model was first proposed by Koshti and You [2,3]. Inspired by their work, improved I3D networks were adopted to build the model proposed in this study.

The American Defense Advanced Research Projects Agency (DARPA) establish ed a major visual surveillance research project called Video Surveillance and Monitoring (VSAM). The project was launched in 1997 with the participation of famous universities and research institutions such as Carnegie Mellon University, the Massachusetts Institute of Technology, and the University of Maryland. VSAM was the first project to implement intelligent video surveillance. The project mainly studies methods of automatic video understanding for battlefields, future cities, and common civilian scenes. CitiLog, a company in France, also developed an automated video detection system that can detect traffic incidents that occur in the surveillance area through background adaptive technology of dynamic images and vehicle image tracking technology. Microsoft, GE, IBM, and other companies in the United States have also studied intelligent surveillance systems, such as GE's VideoIQ products, which have been tested in a variety of scenarios.

We have made several improvements or contributions to the traditional model. (1) We use max pooling instead of average-pooling layers in the new model. (2) We process the video feature output as a two-dimensional matrix. (3) We use the generalized regression neural network (GRNN) classifier instead of the original softmax classification layer. At the same time, our improved model is divided into three steps. First, the two improved I3D networks are used for feature extraction. Second, a red, green and blue (RGB) video and its superimposed optical flow form two separate inputs into a 3D convolutional network, which outputs the fused results. Finally, a deep learning network is used to perform probability regression. In this study, three comparative experiments were conducted on the UCF dataset to prove the superiority of our improved model.

The remainder of this paper is organized into five sections. Section 1.1 introduces the related work. Sections 2.1 and 2.2 present the two-stream inflated 3D ConvNets and the GRNN regression network algorithms. Section 2.3 describes the proposed two-stream network fusion detection algorithm. In Section 3, the experimental results are presented and analyzed. A discussion is presented in Section 4, and finally, Section 5 provides a conclusion.

1.1 Related Work

The earliest abnormal behavior detection technology was proposed to detect violent human behavior by analyzing human movements and limb position [4]. They defined an acceleration measure vector (AMV) and used jerk as the temporal derivative. Video and audio data were used to detect attacks in surveillance video [5]. They use the fusion of audio and video to estimate the attack level of the surrounding environment through a dynamic Bayesian network (DBN). Zhou et al. [6] constructed a violence stream descriptor to detect violence in crowd videos. They used two low-level features to represent violence behaviors, the local histogram of oriented grade (LHOG) and the local histogram of optical flow (LHOF). Xiang et al. [7] used CNN and LSTM models to extract features. Experiments proved that this method can effectively extract temporal features from historical activities. However, these methods result in large detection errors. Recently, a new approach [8] based on behavior heuristics has been developed to classify violent and nonviolent videos. In addition to the difference between violent and nonviolent modes, tracking has been proposed to track normal human movements; then, deviations from such

normal motions are detected as abnormal [9]. Zhang et al. [10] used two-level optimization classification to classify features. They use a two-level optimization model to make better use of the relationship between sparse representation and the desired feature projection. Chen et al. [11] applied Facebook's open [10] source C3D framework and ranking model to identify abnormal behavior. The advantage of this approach is that it can detect a substantial number of abnormal behaviors, and it is not limited to human abnormal behaviors. Luo et al. [12] used Faster RCNN to detect objects in images and proposed a novel mapping rule based on filtered robust object labels. Xia et al. [13] used Weber local binary descriptors to detect features. The method consists of two components: the local binary differential excitation component and the local binary gradient orientation component.

1.2 Definition of Abnormal Behavior

From the psychological point of view: abnormal behavior can be defined as disturbing, socially unacceptable, distressing, maladaptive behavior, and is usually the result of distorted thoughts or cognition. From a behavioral perspective: abnormal behavior is caused by erroneous or ineffective learning and conditioning. From the perspective of cognition: people engage in aberrant behavior because of certain thoughts and behaviors, which are often based on their wrong assumptions. We defined a large number of recurring common behaviors in daily life as normal behaviors, and the rest are abnormal behaviors.

2 Methods

In this study, a video feature extraction method based on the two-stream-I3D network model is proposed. In addition, this approach uses a GRNN to perform feature classification and regression to identify abnormal behaviors. The developed model achieved satisfactory results on the UCF-Crime dataset and UCF-101 behavior dataset. A flowchart of our algorithm is shown in Fig. 1.

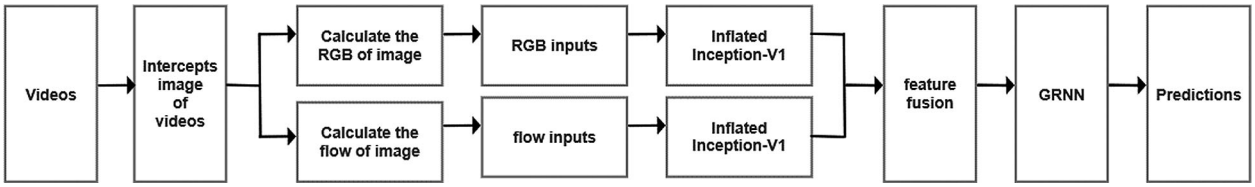


Figure 1: Algorithm flowchart for abnormal behavior detection

2.1 Visual Feature Extraction

Detecting abnormal behaviors in videos is a challenging task because both abnormal human behaviors and other types of abnormal situations are easily affected by factors such as lighting conditions and movement size. Low-cost city cameras capture only low-resolution videos; therefore, it is important to achieve high-performance abnormal behavior detection from such surveillance videos.

In a complex environment, Tomas et al. [14] proposed the OpenPose model, which can track the trunk, limbs, and fingers of multiple people, thereby providing an approach for implementing a method to detect abnormal human behaviors by detecting changes in human body behavior. However, OpenPose cannot detect abnormal behavior outside the human body, such as blasts and fires. The I3D network is better for extracting video features. A two-stream I3D [15] model can also store optical flow information from the video to enable better video understanding. In this paper, we adopt a two-stream I3D to extract visual features from videos.

The I3D model [16] is an improvement on the Inception-V1 (2D) model; its fundamental model structure is called Inflated Inception-V1. Inception-V1 (2D) is a relatively mature image detection model. The core idea of the initial Inception model was to replace a portion of the larger convolutional layers of GoogLeNet with a smaller convolution operation to reduce the number of weight parameters. Inception-V1 (2D) improves the subroutine for the Inception module and adds a 3×1×1 convolutional layer to reduce the convolution size of the original module. The I3D model includes a few improvements compared with the Inception-V1 (2D) model. First, the I3D network expands the 2D basic convolution of the original model to a 3D basic convolution operation, and it adds a time dimension to the convolution kernel. The specific extension method repeats the 2D filter and its weight the same number of times along the time dimension and then divides the result by the number of repetitions to normalize. The main components of Inception-V1 are convolutional layers, pooling layers, and inception layers.

In the convolutional layer, 3D convolution (3D Conv) is better able to capture spatial and temporal information in videos than 2D convolution. For example, suppose the time dimension of 3D convolution is N; in that case, 3D Conv will carry out a convolution operation on N consecutive frames of video and connect the feature map of each frame with adjacent consecutive frames to obtain motion information.

The 3D convergence layer is similar to 3D convolution except that it obtains only the maximum or average value of the current window as the pixel value of the new image; it slides along in steps to obtain multiple feature maps.

The Inception layer is a subroutine module in the Inflated Inception-V1 model. Convolution kernels of 1×1×1 and 3×3×3 are applied simultaneously, which improves the adaptability of the network to objects of different scales. In addition, the 1×1×1 convolution has a unique effect: it reduces the number of weights and the size of the feature map. In addition, because a 1×1×1 convolution has only one parameter, it is equivalent to scaling the original eigenvalue, which improves the recognition accuracy. The model is shown in Fig. 2.

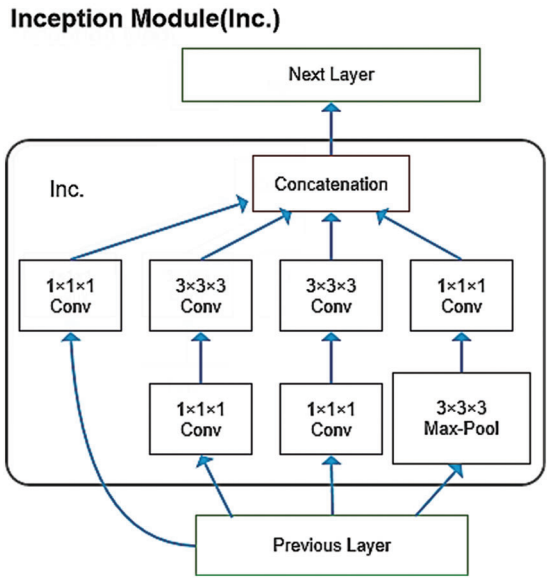


Figure 2: Network structure of an Inception module

At the end of the model, a softmax layer classifies the output. An architecture diagram of the entire model is shown in Fig. 3.

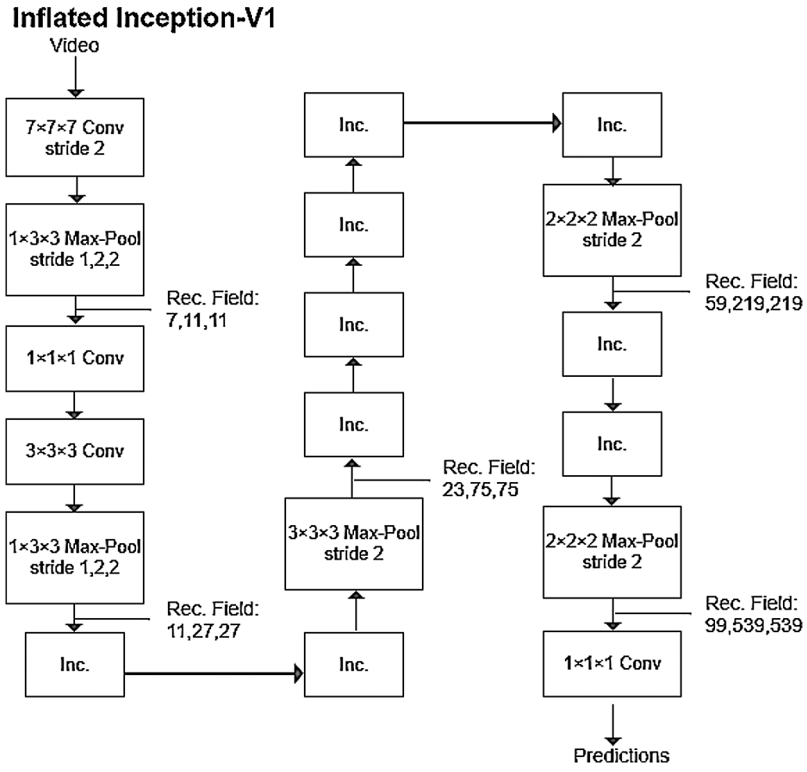


Figure 3: Network structure for I3D

In abnormal behavior recognition fields, network models based on deep learning have been shown to be superior to traditional methods in terms of recognition speed, recognition accuracy and system robustness. When selecting a deep learning method, researchers have mostly used C3D and its improved methods to process videos. However, C3D and its multiple simple network models often make it impossible to process the spatiotemporal video features in depth. In more complex situations, traditional 3D convolutional neural networks tend to miss some motion features, thus affecting the detection results. Therefore, a two-stream I3D network is adopted to perform feature extraction in our system, which deals with various complicated situations.

2.2 GRNN Regression Neural Network Model Principle

The generalized regression neural network is a radial basis function neural network [17,18]. Compared with a traditional regression neural network such as the deep neural network (DNN), a GRNN possesses an extra summation layer and removes the weight connection between the hidden and the output layers (the least square superposition of Gaussian weights), which results in higher accuracy. Simultaneously, as a forward propagation neural network, a GRNN does not require backpropagation to determine model parameters; thus, it converges rapidly. A GRNN has robust nonlinear learning ability and learning speed. The network eventually converges to an optimized regression with many samples.

As shown in Fig. 4, a GRNN has a relatively simple four-layer network structure consisting of an input layer, a model or pattern layer, a summation layer, and an output layer.

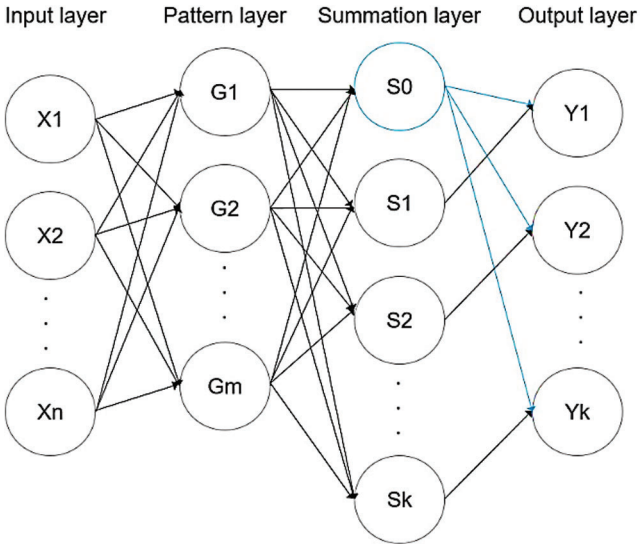


Figure 4: Network structure of a GRNN

The input to the input layer is a test sample whose nodes depend on the size of the sample's features. The pattern layer calculates the value of a Gaussian function in each sample, and the number of nodes is the number of training samples. The summation layer uses two calculation methods. One node performs an arithmetic summation on the output of the pattern layer, while other nodes perform weighted summations on the neurons of the pattern layer. The number of nodes reflects the dimensionality of the output sample plus one. The number of nodes in the output layer is equal to the dimensionality of the sample output vector, and the output is the corresponding summation divided by the output of the first node of the summation layer.

The first type of summation formula found in the summation layer is as follows:

$$\sum_{i=1}^{n} exp\left[-\frac{(X-X_i)(X-X_i)^T}{2\sigma^2}\right] \tag{1}$$

where X is the output of the pattern layer, X_i is the observed value of the random variable and X, σ is the smoothing coefficient. The second type of summation formula found that the summation layer is

$$\sum_{i=1}^{n} Y_i exp\left[-\frac{(X-X_i)(X-X_i)^{\mathrm{T}}}{2\sigma^2}\right] \tag{2}$$

where X_i is the observed value of the random variable X, Y_i is the observed value of the random variable Y, and σ is the smoothing coefficient.

2.3 Abnormal Detection

Overall, the algorithm proposed in this paper is composed of two main parts: a two-stream-I3D feature extractor and a GRNN classifier.

Video information can be divided into spatial information and temporal information. Spatial information refers to the surface information in a frame, while temporal information refers to the relationships between frames. For I3D, the time dimension needs to be set to medium speed. If the speed is set to be too fast, the

model can mix the edge information of different objects together and be unable to capture the dynamic scene accurately.

To extract video information better, two-stream I3D consists of two I3D networks. The first subnet extracts RGB features from the video, while the second subnet extracts the optical flow features. The two I3D networks are trained separately. One is trained on the RGB stream; its input is RGB video frame information extracted from every 16 frames. The other undergoes optical flow training. First, the horizontal and vertical optical flow frames in the video are extracted. Multiple optical flow diagrams are formed into an optical flow group, and this optical flow group with optimized optical flow information is used as input. The TV-L1 algorithm is applied to calculate the optical flow in this second I3D model. As a recursive algorithm, it makes extracting motion features from the video more efficient than a single model.

After several layers of convolution and aggregation, the features extracted by training the RGB stream binary and the features obtained by interleaving with the training optical flow are fused. The specific fusion method used is the weighted average of two adjacent outputs. Further visual features are demonstrated below. To avoid overfitting, L2 regularization is utilized on each convolutional layer, as shown in Fig. 5.

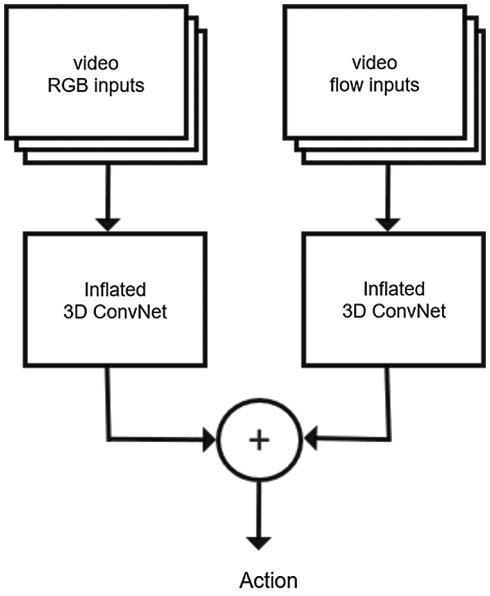


Figure 5: Network structure for the two-stream-I3D mode

Second, the GRNN classifier is adopted to replace the softmax classification layer of the two-stream-I3D model. The features extracted by the two-stream-I3D network are used as the input to the GRNN classifier. The goal of abnormal probability regression on visual features is eventually achieved.

Third, the overall model is shown in Fig. 6. The overall flow chart is as follows: 1) video represents the input data. 2) Two-stream I3D is mainly used for feature extraction of input data. 3) GRNN is used for feature classification. 4) Prediction represents the judgment of the model on the probability of abnormal behavior of the frame.

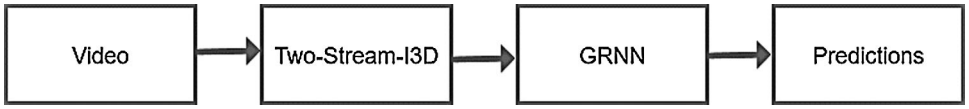


Figure 6: Network structure of the I3D-GRNN model

2.4 Optimization Function

During training, the model uses the AdaGrad stochastic gradient descent function as the model optimization function [19,20]. Unlike other optimization functions, AdaGrad adjusts its learning rate and adaptively assigns a different learning rate to each parameter. When the gradient is larger, the learning rate decays faster; when the gradient is smaller, the rate decays more slowly.

$$\varepsilon_n = \frac{\varepsilon}{\delta + \sqrt{\sum_{i=1}^{n-1} g_i \odot g_i}} \tag{3}$$

where δ is the constant coefficient, g is the gradient value, and ϵ is the global learning rate.

3 Experimental Results

In this section, the results of competitive experiments on the dataset used for abnormal behavior detection are presented.

3.1 UCF-Crime Dataset

The UCF-Crime dataset [21] was developed by the University of Florida. The complete dataset contains multiple types of abnormal events, including up to 13 classes of abnormal behaviors, including arrest, explosion, car accident, and shop robbery. The duration and number of videos in this dataset are more abundant than those in previous datasets.

In addition, the dataset also contains normal activities. The video set can be roughly divided into two categories: videos with abnormal behaviors and videos without abnormal behaviors. We labeled the videos with normal behavior as 1 and videos containing abnormal behavior as 0 and trained the model with supervision. Since the video duration of the training set video is short after preprocessing, we can try to label all frames of the video with abnormal behavior as 0. After the model was trained, we selected 50 videos as our test set. Fig. 7 shows four screen captures from the videos.

(a) (b) (c) (d)

Figure 7: Sample image from the UCF-Crime dataset: (a) Fighting (abnormal); (b) Robbery (abnormal); (c) Explosion (abnormal); (d) Traffic intersection (normal)

3.2 Experiments

According to the definition of abnormal behaviors in Section 1.2, behavior videos can be classified into normal behavior videos and abnormal behavior videos. However, due to various objective factors, abnormal

behavior videos can be decomposed into different types of abnormal behavior videos. In this paper, abnormal behavior videos with less activity range in our data set are defined as hidden abnormal behavior videos, such as theft. At the same time, we defined abnormal behavior videos with low resolution as high-distortion abnormal behavior videos, such as an explosion in a very distant place. It is obvious that the detection of hidden and highly distorted abnormal behavior videos can bring great challenges to the algorithm mode.

The above datasets were used to train the model, and fifty representative videos were selected as a test set to test the model. The videos are divided into four categories: (1) abnormal behavior videos, (2) hidden abnormal behavior videos, (3) high-distortion abnormal behavior videos, and (4) no abnormal behavior videos. The distribution of these categories in the test dataset is approximately 42:9:5:8 (hidden abnormal behavior videos and high-distortion abnormal behavior videos both belong to abnormal behavior videos). First, the input videos were re-encoded to MP4 format at a frame rate of 30 fps and a resolution of 240×320. The model was trained and tested using an RTX2080ti GPU. The rectified linear unit (ReLU) activation function was used to activate the fully connected (FC) layer in the first layer, and the sigmoid activation function in the last layer was optimized with AdaGrad. Sixty percent dropout regularization was adopted among the fully connected layers, and the initial learning rate was set to 0.001. Four representative video samples of different categories are selected for this analysis.

Fig. 8 shows an explosion video, while the lower part shows the system detection result. In the results, the abscissa represents the number of frames, and the ordinate represents the probability of the system detecting an anomaly, which is a value ranging from 0 to 1. As shown in the figure, when no abnormal behavior is present, the ordinate of the prediction curve is infinitely close to zero, with almost no change. When abnormal behavior occurs, the curve changes significantly, and the ordinate can rise to 0.7 or even to 1.

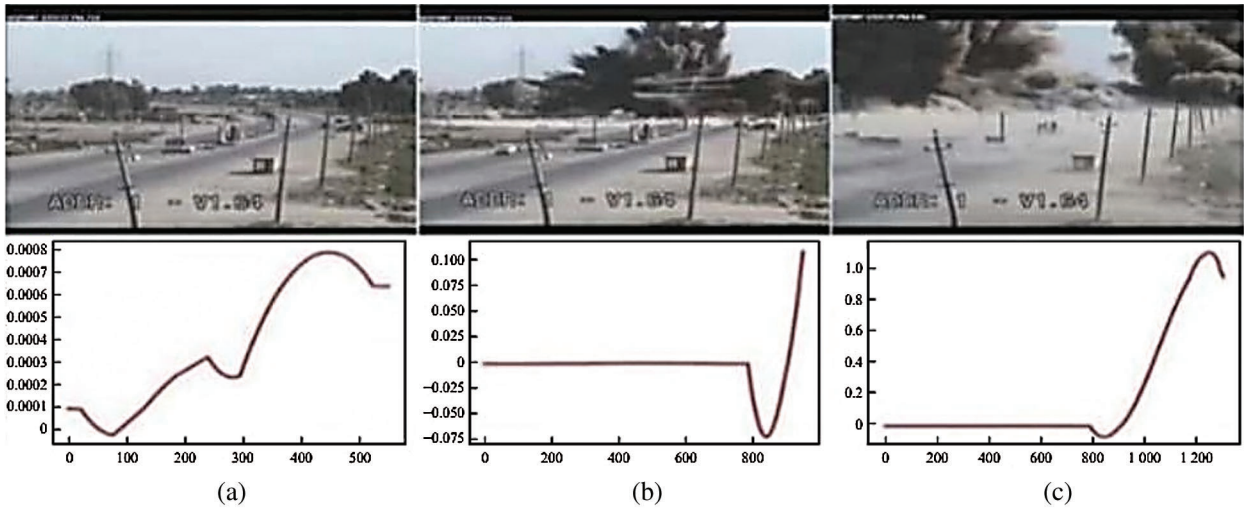
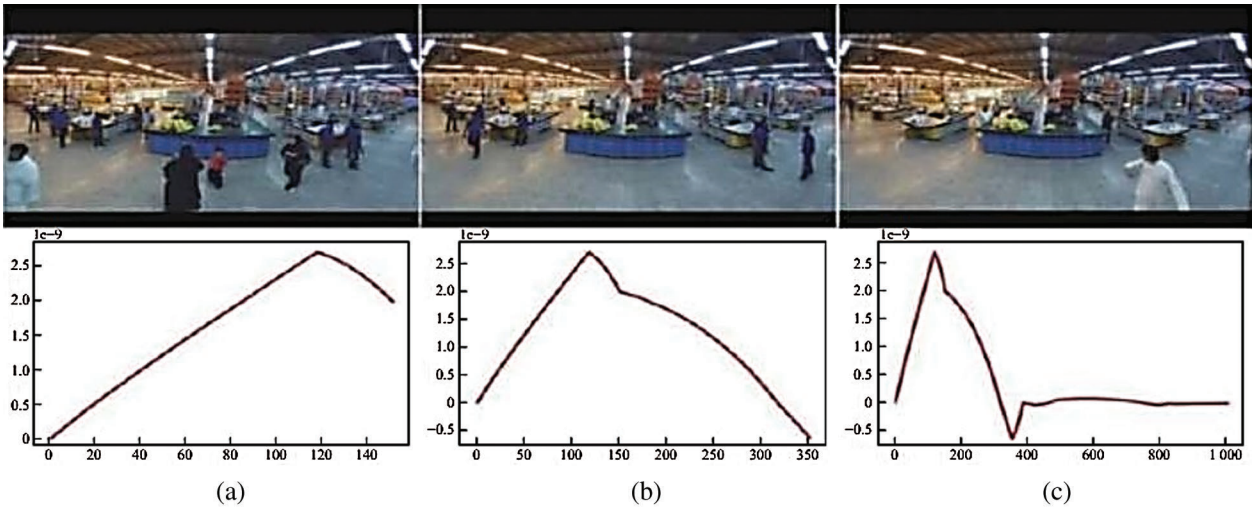


Figure 8: Video detection of abnormal behavior: (a) image with no abnormal content, (b) explosion image; (c) postexplosion image

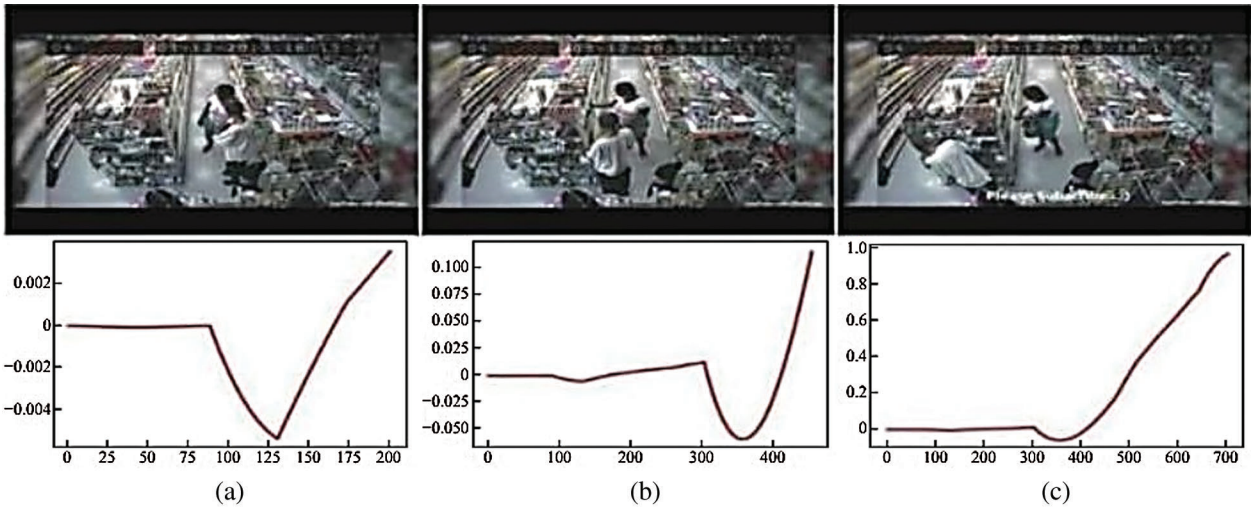
Fig. 9 shows the video detection results without abnormal behavior. The experimental analysis shows that because no abnormal behavior occurs in the video image, the predicted ordinate of the curve is approximately 10 to the -9th power. After zooming in and displaying the ordinate, the curve does change up and down. In fact, in the original coordinate system, the curves would fluctuate only slightly and would still approach zero indefinitely.



(a) (b) (c)

Figure 9: Video detection of normal behavior: (a, b, c) all show normal images

Fig. 10 shows a video scene where a woman steals a glass of water from a supermarket. The scope of this criminal action is minimal and is mainly concentrated in the hands; the overall situation of the image does not change significantly. The experimental results show that even when the behavior is hidden, the system can detect anomalies. When abnormal behavior occurs, the prediction curve presents a large jump, and the ordinate immediately reaches above 0.8.



(a) (b) (c)

Figure 10: Video detection of hidden abnormal behavior: (a) not an abnormal image; (b, c) theft images

Fig. 11 shows the results of detecting a video with high distortion and abnormal behavior. The video shows pictures of explosions in the distance. Only explosive smoke appears in the upper left of the picture. The experimental results show that even when the video distortion is high, the system can still detect abnormal behavior. However, its detection strength is not as good as that of other abnormal videos. When abnormal behavior occurs, the prediction curve also changes significantly. However, after the

ordinate reaches 0.4, there is a downward trend. Although the predicted value never exceeds 0.7, the curve varies considerably compared with the nonabnormal video.

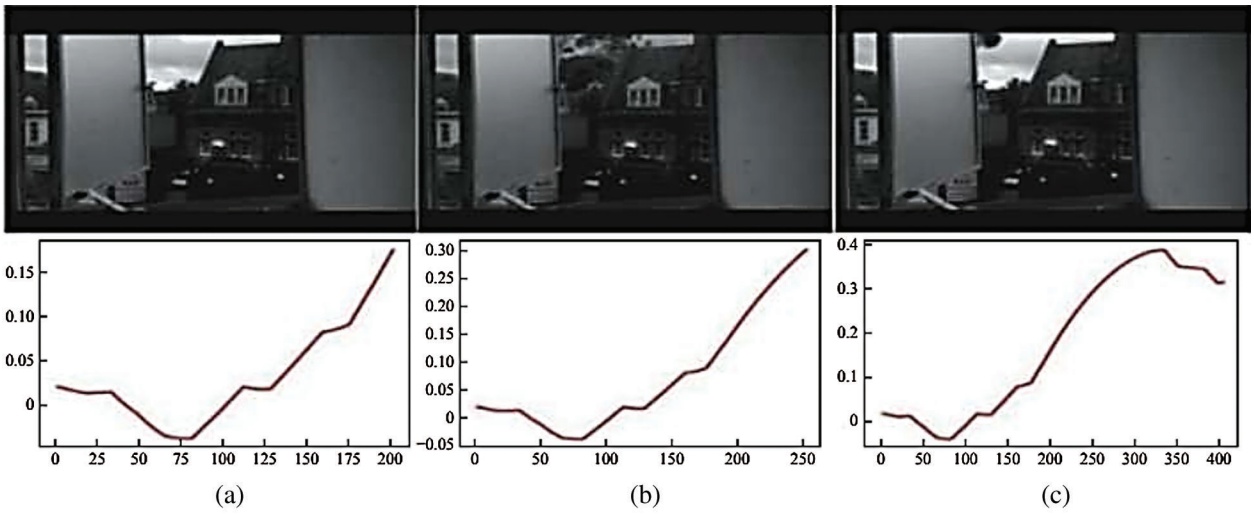


Figure 11: Video detection of high-distortion video: (a) not an abnormal image; (b) explosion image; (c) postexplosion image

4 Discussion

This paper uses two 3D ConvNet networks with inflated traffic and deep learning networks. The probability regression technique is used to identify daily abnormal behaviors to improve accident responses. Compared with existing methods, the two-stream I3D feature extraction technology can extract better visual features while retaining both the optical flow and RGB information in the video. Probabilistic regression technology is better able to quantify the data and provide a more intuitive visual experience. On the UCF-101 dataset, our model achieves an average accuracy of 93.7%, and it requires less detection time than the other models (see Tab. 4).

In a video, images with the present abnormal probabilities often have only minuscule durations. In contrast, unnecessary information occupies most of the video, which makes learning abnormal features quite difficult. Suto et al. [22] believed that data preprocessing has an obvious effect on neural network effects. Xiang et al. [23] proposed a method to select the best element as the final element. Determining a method to solve this problem is worthwhile. Therefore, in our approach, two-stream-I3D visual feature extraction is applied every 16 frames to reduce the amount of network input and the number of required calculations and to remove redundant information from the experimental results.

To detect anomalies, systems that use only static pictures are insufficiently accurate. Connecting static images and increasing the time dimension reflects this connection. Compared with the C3D network, the I3D network has a more suitable network structure. Compared with ordinary two-stream networks, the I3D network used in two-stream-I3D expands the 2D basic convolution to 3D basic convolution and adds a time dimension to the convolution kernel and pooling layer. The two-stream-I3D model has a deeper network structure. It involves training two separate I3D networks—an RGB stream network and an optical flow network; then, it combines the two I3D outputs at the end of the model, allowing more accurate predictions to be obtained.

Feature classification is another challenging problem. Yang et al. [24] used variational autoencoder steganography (VAE-Stega) to learn the overall statistical distribution characteristics. A GRNN is a relatively simple four-layer network structure that includes an input layer, a pattern layer, a summation layer, and an output layer. Compared with other classifiers, a GRNN can better fit the sample data and has a fast-training speed and good classification ability. In particular, a GRNN can obtain favorable classification results when contending with unstable or small-sample data.

Tab. 1 shows test data of some parts of the UCF-Crime dataset. We regard abnormal behaviors as positive and normal behaviors as negative. By calculating their precision and recall, the precision and the recall of the model was as high as 97% and 88%, indicating that the model had high reliability for abnormal behavior detection. Under four different types of videos, the detection accuracy of hidden abnormal behavior video is the worst. Nearly half of the detected videos are normal and contain no abnormal behavior. During further study of the dataset, it was found that in hidden abnormal behavior videos, most of the abnormal behaviors occur only using a particular part of the body; in fact, some of these movements are challenging to observe with the naked eye. Moreover, in the test set, the number of videos belonging to this type is small, and the possibility of errors is greater (see Fig. 12).

Table 1: Test samples and related data

Video category	Total number of videos	Number of abnormal behaviors	Accuracy (%)	Precision	Recall
No abnormal behavior video	8	1	87.5	97%	88%
abnormal behavior video	42	37	88.4		
Hidden abnormal behavior video	9	5	55.6	–	–
High-distortion abnormal behavior video	5	4	80	–	–

Figure 12: A video clip labeled “stealing”

To verify the superiority of the algorithm in this paper, 5 methods are selected, namely, a support vector machine, random forest, 3D convolutional network (C3D)-GRNN, k-nearest neighbors (KNN), and an ordinary C3D model. These models are all applied to recognize and classify abnormal behaviors in videos. As shown by the results in Tab. 2, compared with the standard C3D model, the C3D-GRNN model and other typical methods, the algorithm proposed in this paper improves the efficiency of identifying sample frames; its detection time is similar to the time used by the best current algorithms (support vector machine (SVM) and KNN).

Table 2: Time consumption of different classification method algorithms

Classification method	The time required to test for an exception in a frame
C3D	0.220
KNN	0.020
Random Forest	0.017
SVM	0.014
C3D-GRNN	0.134
Our model	**0.056**

Furthermore, we compare the performance of the five architectures (long short-term memory (LSTM), 3D-ConvNet, two-stream, 3D-fused, and traditional two-stream I3D) with our improved two-stream I3D architecture for the UCF-101 dataset. We test on the split 1 test sets of UCF-101. Specifically, during testing, we consider only the output on the last frame. Input video frames are subsampled by keeping one out of every 16 from an original 25 frames-per-second stream. The parameters and temporal input sizes of all models are given in Tab. 3. As shown in Tab. 3, our improved I3D model can obtain more out of the flow stream and RGB steam than the other models. Tab. 4 shows the classification accuracy when training and testing on UCF-101. As shown in Tab. 4, our new I3D models perform best in all datasets, with either RGB, flow, or fusion of RGB and flow modalities. More interestingly, our improved model slightly outperforms the traditional two-stream inflated 3D ConvNet for all the modalities. This may be due to the following improvements. (1) We use max pooling instead of average-pooling layers in the new model. Although max pooling and average pooling both downsample the data, max pooling is more similar to feature selection than average pooling; it selects features that enable better classification recognition, effectively reducing the deviation of the estimated mean caused by convolutional layer parameter error. (2) In contrast to the original model output, we process the processed video feature output as a two-dimensional matrix because that approach enables better classification. (3) At the end of the model, we use the GRNN classifier instead of the original softmax classification layer to classify and regress the output features; this helps the subsequent anomaly recognition and provides a probability estimation. The GRNN can adapt to a small number of datasets during training and achieve better classification results.

Table 3: Number of parameters and temporal input sizes of the models

Method	Params (M)	Training		Testing	
		Input Frames	Temporal Footprint (s)	Input Frames	Temporal Footprint (s)
LSTM	9	25 RGB	5	50 RGB	10
3D-ConvNet	79	16 RGB	0.64	240 RGB	9.6
Two-Stream	12	1 RGB, 10 flow	0.4	25 RGB, 25 flow	10
3D-Fused	39	5 RGB, 50 flow	2	25 RGB, 25 flow	10
Two-Stream I3D	25	64 RGB, 50 flow	2.56	250 RGB, 250 flow	10
Our model	36	30 RGB, 64 flow	4	250 RGB, 250 flow	10

Table 4: Architecture comparison: training and testing on split 1 of UCF-101

Architecture	UCF-101 dataset		
	RGB	Flow	RGB + Flow
LSTM	81.0	–	–
3D-ConvNet	51.6	–	–
Two-Stream	83.6	85.6	91.2
3D-fused	83.2	85.8	89.3
Two-Stream I3D	84.5	90.6	93.4
Our model	**85.3**	**91.1**	**93.7**

5 Conclusion

An abnormal behavior recognition algorithm based on the I3D-GRNN model is proposed in this research. The algorithm can identify a wide range of abnormal behaviors in various scenarios and achieves high detection accuracy on a small range of test sets. The model is used to improve video understanding, and the GRNN network is applied to perform regression. Based on the image information in the video, the two-stream-I3D model used by the algorithm extracts optical flow information through the dual subnet to improve video feature understanding and then performs feature fusion in the inception layer to achieve competitive accuracy. The GRNN is used at the end to determine the network output. With the added features in the probability regression, it is beneficial to visualize the abnormal probability data [25]. The experimental results show that our model outperforms the state-of-the-art methods. Based on the above experimental results, we will continue to optimize the model in the future, including reducing the amount of calculation and improving the regression ability of the model, to improve the ability of the model to detect hidden abnormal behaviors. To achieve intelligent monitoring and management contributions.

Author Contributions: Conceptualization, L. L. and J. P.; methodology, L. L.; investigation, M. L., H. L and L. Z.; experiment, M. L., C. Z. and L. L.; writing-original draft preparation, W. P., H. L. and M. L.; writing—review and editing, J. P., F. W. and S. L.; supervision, J. P.

Funding Statement: This project was funded by the Guangzhou Science and Technology Plan Project Key Field R&D Project (202007030005), the National Natural Science Foundation of China (61876067), and the Guangdong Natural Science Foundation of China (2019A1515011375).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. M. S. Islam, S. Rahman, M. M. Rahman, E. K. Dey and M. Shoyaib, "Application of deep learning to computer vision: A comprehensive study," in *2016 5th Int. Conf. on Informatics, Electronics and Vision (ICIEV)*, Dhaka, Bangladesh, pp. 592–597, 2016.

[2] J. You, P. Shi and X. Bao, "Multistream i3d network for fine-grained action recognition," in *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, pp. 611–614, 2018.

[3] D. Koshti, S. Kamoji, N. Kalnad, S. Sreekumar and S. Bhujbal, "Video anomaly detection using inflated 3D convolution network," in *2020 Int. Conf. on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp. 729–733, 2020.

[4] A. Datta, M. Shah and N. D. V. Lobo, "Person-on-person violence detection in video data," *Object recognition supported by user interaction for service robots*. vol. 1, pp. 433–438, 2002.

[5] J. F. Kooij, M. Liem, J. D. Krijnders, T. C. Andringa and D. M. Gavrila, "Multimodal human aggression detection," *Computer Vision and Image Understanding*, vol. 144, no. 3, pp. 106–120, 2016.

[6] P. Zhou, Q. Ding, H. Luo and X. Hou, "Violence detection in surveillance video using low-level features," *PLOS One*, vol. 13, no. 10, pp. e0203668, 2018.

[7] L. Xiang, G. Guo, Q. Li, C. Zhu and H. Ma, "Spam detection in reviews using LSTM-based multientity temporal features," *Intelligent Automation and Soft Computing*, vol. 26, no. 4, pp. 1375–1390, 2020.

[8] S. Mohammadi, A. Perina, H. Kiani and V. Murino, "Angry crowds: Detecting violent events in videos," in *European Conf. on Computer Vision(ECCV)*, Amsterdam, Netherlands, pp. 3–18, 2016.

[9] D. Dawei, Q. Honggang, H. Qingming, Z. Wei and Z. Changhua, "Abnormal event detection in crowded scenes based on structural multiscale motion interrelated patterns," in *2013 IEEE Int. Conf. on Multimedia and Expo (ICME)*, San Jose, CA, USA, pp. 1–6, 2013.

[10] G. Zhang, H. Sun, Y. Zheng, G. Xia, L. Feng *et al.,* "Optimal discriminative projection for sparse representation-based classification via bilevel optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1065–1077, 2020.

[11] D. Chen, P. Wang, L. Yue, Y. Zhang and T. Jia, "Anomaly detection in surveillance video based on bidirectional prediction," *Image and Vision Computing*, vol. 98, pp. 103915, 2020.

[12] Y. Luo, J. Qin, X. Xiang and Y. Tan, "Coverless image steganography based on multiobject Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1, 2020.

[13] Z. Xia, C. Yuan, R. Lv, X. Sun, N. N. Xiong *et al.,* "A novel weber local binary descriptor for fingerprint liveness detection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 4, pp. 1526–1536, 2020.

[14] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "Realtime multiperson 2D pose estimation using part affinity fields," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, USA, pp. 1302–1310, 2017.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, USA, pp. 4724–4733, 2017.

[16] S. Park and D. Kim, "Study on 3D action recognition based on deep neural network," in *2019 Int. Conf. on Electronics, Information, and Communication (ICEIC)*, Auckland, NewZealand, pp. 1–3, 2019.

[17] L. Qiao, Z. Wang and J. Zhu, "Application of improved GRNN model to predict interlamellar spacing and mechanical properties of hypereutectoid steel," *Materials Science and Engineering: A*, vol. 792, no. 1, pp. 139845, 2020.

[18] L. Li, "Analysis and data mining of intellectual property using GRNN and SVM," *Personal and Ubiquitous Computing*, vol. 24, no. 1, pp. 139–150, 2020.

[19] H. Saito and M. Kato, "Machine learning technique to find quantum many-body ground states of bosons on a lattice," *Journal of the Physical Society of Japan*, vol. 87, no. 1, pp. 14001, 2018.

[20] X. Ran, Z. Shan, Y. Fang and C. Lin, "An LSTM-based method with attention mechanism for travel time prediction," *Sensors*, vol. 19, no. 4, pp. 861, 2019.

[21] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 6479–6488, 2018.

[22] J. Suto and S. Oniga, "Efficiency investigation of artificial neural networks in human activity recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 4, pp. 1049–1060, 2018.

[23] L. Xiang, S. Yang, Y. Liu, Q. Li and C. Zhu, "Novel linguistic steganography based on character-level text generation," *Mathematics*, vol. 8, no. 9, pp. 1558, 2020.

[24] Z. Yang, S. Zhang, Y. Hu, Z. W. Hu and Y. Huang, "VAE-Stega: Linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2021.

[25] T. Lima, B. Fernandes and P. Barros, "Human action recognition with 3D convolutional neural network," in *2017 IEEE Latin American Conf. on Computational Intelligence (LA-CCI)*, Arequipa, Peru, pp. 1–6, 2017.