Video Abnormal Event Detection Based on Optical Flow and 3DCNN

Zhenjiang Li*, Guangli Wu, Ye Liu, Yifan Shuai and Lei Wang
School of Cyberspace Security, Gansu University of Political Science and Law, Lanzhou, China

*Corresponding author e-mail: lzj6843@gsli.edu.cn

Abstract. With the advancement of technology, the detection of abnormal events in videos has extensive theoretical and practical significance. Most events in videos are normal events, but a small number of abnormal events contain a lot of information. How to identify individual abnormal events from massive videos is the focus of concern. The paper made full use of the spatio-temporal continuity of video images. Firstly, a space-time cube was constructed by integrating optical flow information; secondly, a multi-layer three-dimensional convolutional neural network model was constructed; finally, the entire process of video abnormal event detection was realized based on this model. The experiment on the UCSD dataset showed that the method in this paper can effectively detect abnormal events in videos.

Keywords: Optical Flow, 3DCNN, Video Detection, Abnormal Events

1. Introduction
During the implementation, deployment, operation and maintenance of the video monitoring system, managerial personnel should monitor the current system 24 hours a day. The cost investment for this part accounts for a considerable proportion of the cost for the entire system. To reduce the system cost and improve the monitoring efficiency, it is particularly necessary to use the automation method to deal with this process. Currently, with the development of relevant technologies, the application of intelligent algorithms to monitor video contents and timely alarms in positions that may be abnormal have certain practical conditions [1].

There is no strict definition of video abnormal events at present. This is because video abnormal events have a lot to do with specific application scenarios. The definition of anomaly is application-oriented rather than event-oriented. Therefore, the concept of video anomaly should be abstracted at a higher level during general research. It is generally believed that the monitored object has two different types of behavior patterns in a specific application scenario. One pattern has a high occurrence frequency and occupies the main content of the video; the other pattern has a low occurrence frequency but contains a large amount of information, which is the focus of the application, then the second pattern is called as an abnormal event [2]. The objective of the video abnormal event detection research is to identify a small number of low-frequency pattern events from a large number
of high-frequency pattern events, which is a valuable information extraction process from the perspective of information theory.

At present, many scholars have conducted various studies about how to detect abnormal events in videos. In terms of the granularity of research objects, studies can be divided into two categories, one is to detect the global anomaly of videos, and the other is to detect the local anomaly of videos [3]. In the detection of global video abnormal events, the research object is the video frame, in other words, it is necessary to identify the frame number of the abnormal behavior pattern [4]. In the detection of local video abnormal events, it not only needs to identify the frame number of the abnormal behavior pattern, but also needs to identify the specific pixel position of the abnormal behavior pattern on the frame, which is more difficult [5].

The learning algorithms can be divided into two types: supervised and unsupervised, both of which require massive data to support the algorithms accordingly. In the face of massive monitoring video data, it is difficult to annotate frame by frame. If the granularity of the annotation needs to be further reduced to the pixel level, it will be unrealistic to be realized on a slightly large dataset. Therefore, under realistic data conditions, the unsupervised method is the main way to realize video abnormal event detection [6]. However, the supervised method still has a high value in the research on video abnormal event detection while considering the generation of simulation data [7], transfer learning [8] and other problem solutions from different perspective. In addition, the characteristics of the abnormal event model itself determine that the data size of normal event samples is much higher than that of abnormal event samples. Therefore, how to reasonably solve the data size deviation and effectively augment the data is an important direction for the research on video abnormal event detection.

Based on the continuity of video frames, this paper designed a supervised detection model with the 3DCNN network as the core structure, and integrated the gray-scale features of video frames as well as the optical flow features to carry out research on video abnormal event detection. The chapters of the paper are arranged as follows: Chapter 1 is the introduction, which introduces the related work of the video abnormal event detection research; Chapter 2 introduces the 3DCNN model and the video optical flow model; Chapter 3 elaborates the research method, including model construction, feature extraction and data processing; Chapter 4 expounds related experiment and analysis of experimental results; Chapter 5 summarizes the work of the full text.

2. Related Works

2.1. 3DCNN

The classic CNN model is oriented to two-dimensional images. It can obtain the spatial connection between adjacent pixels in the two-dimensional space through multi-layer convolution operations, and then automatically learn and extract corresponding features based on the annotation information of the samples. When application scenarios turn to image sequences, namely, videos, the two-dimensional CNN model cannot effectively express the temporal continuity between adjacent frames. Although adjacent frame sequences can be regarded as multiple channels of a two-dimensional image, such an approximation still has certain shortcomings. On the one hand, when the image itself has multiple channels, it is impossible to properly handle the data association between different channels and frames; on the other hand, this approximation abandons the important feature of the great spatial correlation of image sequences in videos.

The 3DCNN model is an effective extension of the classic CNN model, and video processing is its natural application scenario [9]. In the video application of this model, the input sample is a 3D space-time cube with three dimensional attributes of height, width and time. Its convolution kernel also has three-dimensional characteristics. For a point \((x, y, z)\) in the model, the value in the \(j\)-th feature map of the \(i\)-th layer is as shown in Formula (1):
\[ v_{y}^{xyz} = \tanh \left( b_{y} + \sum_{m} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} w_{pqrs} \left( x+p \right) \left( y+q \right) \left( z+r \right) \right) \]  

(1)

Where \( R \) refers to the size of the convolution kernel along the time dimension, \( P \) and \( Q \) refer to the height and width of the convolution kernel in the space dimension, respectively. It can be seen from Formula (1) that this three-dimensional convolution can effectively include adjacent time dimension information in the convolution calculation to obtain richer video image features.

2.2. Video Optical Flow

In essence, a video is a projection sequence of an object in a two-dimensional space when it moves in a three-dimensional space. When it is moving, it will produce an instantaneous velocity on the plane, and the velocity field of multiple observation points constitutes the optical flow field [10]. The description of the video by the optical flow field model focuses on the motion state of the object in the video rather than the texture of the object. This method is widely used in various video behavior understanding applications. When the optical flow is calculated in pixels, the optical flow is called as dense optical flow, which reflects the instantaneous velocity and direction of each pixel between adjacent frames. After the spatially adjacent pixels with similar motion patterns are aggregated, the motion pattern of an object in the video can be obtained. In different application scenarios, some modes are considered as normal while some are considered as abnormal events.

During the calculation of the dense optical flow, the relationship between the motion vector \((u, v)\) of a pixel at any point on the image and the gray value \(I\) of the current point is shown in Formula (2):

\[
\begin{bmatrix}
I_x \\
I_y \\
u \\
v
\end{bmatrix}
= -I_t
\]

\[
I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y}, I_t = \frac{\partial I}{\partial t}
\]

(2)

Among them, \( I_t \) refers to the rate of change of the pixel in time, \( I_x \) and \( I_y \) refer to the rate of change of the gray value of the current pixel in the horizontal and vertical directions, respectively. The gray change in time can be obtained by subtracting adjacent frames, and the gray change in space needs to be solved by adding additional constraints.

3. Method of This Paper

The research framework of the thesis is shown in Figure 1. Firstly, several frame sequences were extracted from the original video as a single original sample. If the number of abnormal event frames in this sample was greater than the number of normal event frames, then this frame sequence sample was considered as abnormal, otherwise it was considered as normal. In this paper, such a single sample was called a space-time cube. Secondly, all the feature channels of the space-time cube were constructed. In addition to the original gray channel, two additional channels of optical flow intensity in the x and y directions were added as well. Finally, a 3DCNN model was constructed for training, which was a superposition of several “3DConvolution + 3DPooling+Batch Normalization” structures.
3.1. Space-Time Cube Construction
In the research, the basic unit of the recognition model input was the space-time cube, which is essentially a four-dimensional array whose dimensions include the width and height of the video frame, the length of the video frame sequence, and the feature depth of a single video frame. The first three dimensions can reflect the three-dimensional properties of the video in time and space. When the model is constructed, a three-dimensional convolution kernel will perform convolution operations on it. The last dimension is the characteristics of different channels, which will not participate in the convolution calculation of the three dimensions.

In the model, the width and height of the video are restricted by the physical properties of the original video, the performance of the image processing hardware, the granularity of detection and recognition, and the method of model pooling. The value is usually an integer multiple of 2. For example, 256, 240, etc. Under the condition that the aspect ratio of the original video frame was approximately unchanged, the original input video was scaled to obtain a new video to meet the length and width requirements of the model. In this paper, the value was set to 160*240.

For each frame of the input video, n adjacent frames were taken forward and backward respectively to obtain a video sequence with a length of 2n+1. If a video has m frames, then a sequence of m-2n frames will eventually be obtained. The sequence was reannotated based on the proportion of normal frames in the sequence. The annotation of a single frame was converted into the annotation of the frame sequence. In this paper, 3 frames were taken forward and backward respectively, and the final video sequence length was 7. To strengthen the extraction of moving target features in the sample, in addition to the original gray-scale features of the video frame, the optical flow features in both the horizontal and vertical directions were added. As the optical flow calculated from the current frame and the next frame was regarded as the feature map of the current frame, the resulting “gray-optical flow” sample was 1 frame less than that of the original video.

In this paper, a piece of video with m frames was processed through the above steps to obtain m-7 space-time cubes. The dimensions of the cube were 160*240*7*3. Each space-time cube was considered as the sample data corresponding to the intermediate frame and then passed into the subsequent model for training and prediction. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

3.2. Construction of 3DCNN Model
To detect abnormal events in videos, this paper constructed a three-dimensional convolutional neural network based on the processing of the original image frames in Section 3.1. The network structure is shown in Figure 2:
Figure 2. The network structure proposed in this paper

The entire network consisted of four modules: input, convolution, full connection, and output. The input size was the size of the space-time cube mentioned in Section 3.1. The convolution module was composed of three groups of similar structures, namely, “convolution-pooling-batch normalization”. To achieve the balance between feature visions and feature expression capabilities, the number of convolution kernels also increased while the size of the convolution kernel decreased with the increasing depth of the network. After each convolution was completed, the maximum pool was used to perform a pooling operation. The step size in the horizontal and vertical directions was 2, and the step size in the time axis direction was 1. Higher-level video features were extracted in a pyramid manner. After pooling, to accelerate training and improve the generalization ability of the model, a special batch normalization layer was added. In the fully connected module, first, the 3D global average pool was used to flatten the data, which also effectively reduced the number of parameters to be trained. Then, the final detection result was obtained through a fully connected layer with Dropout.

4. Experiment and Analysis

4.1. Dataset

To verify the effectiveness of the model proposed in the paper, related experiment was designed on the UCSD dataset. The data was divided into 2 subsets (p1 and p2) corresponding to different scenarios. The experiment in this paper only used annotated videos. After data preprocessing, 1930 samples were obtained in p1 and 1926 samples were obtained in p2. Some samples are shown in Figure 3, in which Figure 3a and Figure 3b are the video frames in P1, Figure 3c and Figure 3d are the video frames in P2, and the part surrounded by the red border is the abnormal area.

Figure 3. Some video abnormal samples on UCSD

4.2. Experimental Analysis

To compare and present the effectiveness of the method in this paper, three related tests were also designed in addition to the test using the model shown in Figure 2. The model structure of these three experiments is the same as that shown in Figure 2. There were some differences only in terms of input and convolution types. The configurations of all four experiments are shown in Table 1:
Table 1. The configurations of all experiments

| Experiment Name | The Type of convolution | Input size          | Video frame sequence length | Number of channels |
|-----------------|-------------------------|---------------------|----------------------------|--------------------|
| 2d-f            | Conv2D                  | (160, 240, 1)       | 1                          | 1                  |
| 2d-fs           | Conv2D                  | (160, 240, 7)       | 1                          | 7                  |
| 3d-fs           | Conv3D                  | (160, 240, 7, 1)    | 7                          | 1                  |
| 3d-fsof         | Conv3D                  | (160, 240, 7, 3)    | 7                          | 3                  |

The code implementation was performed on the experiment based on keras and tensorflow under python. The hardware environment for model training was 64G memory, I7CPU, and 1080Ti graphics card. During training, the samples were divided into the training set and the test set in accordance with the ratio of 4:1. The accuracy of the training set on p1 and p2 varies with the number of iterations, as shown in Figure 4:

Figure 4. The accuracy of the training set on p1 and p2

It can be seen from the figure that all models showed fast convergence speed and high accuracy. The performance of the two models based on 3DCNN was better than that of the two models based on 2DCNN. Since all models reached the best performance on the test set after 10 iterations, to further analyze the discriminative ability of the models, the accuracy distribution of all models on the verification set after 10 iterations was analyzed, as shown in Figure 5:

Figure 5. The accuracy distribution of the verification set on p1 and p2

It can be seen from the figure that after the first three models converged smoothly on the training set, the convergence on the validation set was not very satisfactory. The accuracy of the validation set always fluctuated greatly, showing that the models cannot accurately extract core distinctive features between normal events and abnormal events in the training process, resulting in overfitting during training. Although the accuracy of the model in this paper also fluctuated on the validation set, the fluctuation range was significantly smaller than that of the other three models. In particular, it had the best performance on p2, and the accuracy fluctuated at a high level.
5. Conclusion

A video is a collection of a set of image sequences continuous in time and similar in space. During the analysis and processing of a video, the continuity in time and space must be considered to extract the intrinsic characteristics. The optical flow reflects the motion characteristics of the objects in the video, while the three-dimensional convolution accumulates such characteristics in the space-time dimension at the same time. This paper used the idea of video three-dimensional convolution to construct a network model and studied the video abnormal event detection based on the gray scale and optical flow features. As shown by experimental results, the method in this paper can preliminarily detect video abnormal events. To further improve the detection performance and extend the application range of the model, the following work can be started from the introduction of semi-supervised methods, the abstraction of motion features, and the adjustment of the model. The research framework of the thesis is shown in Figure 1. Firstly, several frame sequences were extracted from the original video as a single original sample. If the number of abnormal event frames in this sample was greater than

Acknowledgments

This work is supported by the Project of Innovation Ability Improvement of Colleges in Gansu Province (2019A-091), Scientific Research project of Gansu Province(20JR10RA335), Scientific Research project of Gansu University of Political Science and Law (2017XQNLW14, GZF2020XZDA03), and Project of Innovation and entrepreneurship training program for College Students in Gansu Province (S202011406002, S202011406003).

References

[1] Tran D, Yuan J, Forsyth D. Video Event Detection: From Subvolume Localization to Spatiotemporal Path Search. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 36(2):404-416.
[2] Vu H, Nguyen T D, Travers A, et al. Energy-Based Localized Anomaly Detection in Video Surveillance// Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2017.
[3] Cheng K W, Chen Y T, Fang W H. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[4] Leyva R, Sanchez V, Li C T. Video Anomaly Detection with Compact Feature Sets for Online Performance. IEEE Transactions on Image Processing, 2017:1-1.
[5] Wang S, Zhu E, Yin J, et al. Video anomaly detection and localization by local motion based joint video representation and OCELM. Neurocomputing, 2017, 277(FEB.14):161-175.
[6] Deepak K, Chandrakala S, Mohan C K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal Image and Video Processing, 2020(4).
[7] Chu M, Xie Y, Mayer J, et al. Learning temporal coherence via self-supervision for GAN-based video generation. ACM Transactions on Graphics, 2020, 39(4).
[8] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1):221-231.
[9] Li Z., Yang W., Wu G., Liu L. Unsupervised Video Anomaly Detection Based on Sparse Reconstruction// Big Data Analytics for Cyber-Physical System in Smart City (BDCPS), 2020, pp 994-1001