Review for Anomaly Detection in Video Surveillance System Based on Deep Learning

Yuchang Si[1]

Software College, Shenyang Normal University
Shenyang 110034 China
siyuchang@163.com

Abstract. In this paper, abnormal target detection and location in video surveillance system are studied. In recent years, with the rapid development of network information technology, video surveillance technology has been widely used, artificial anomaly detection methods have no way to meet the effective growth of video surveillance data, with 3D technology, face recognition technology, etc., also promote the development of the field of computer vision, for the rapid analysis of a large number of video data to provide effective support. At present, abnormal target detection methods in video surveillance system mainly include the following two methods: One is to extract two-dimensional data features from video surveillance data, and effectively express video targets according to the extracted features. The information expressed mainly includes time information and spatial information. The second is to directly learn 3D space-time features for the module with motion information to detect the location of the abnormal target. Finally, the paper summarizes the full text and looks forward to the future development direction of video anomaly detection from three aspects: data set, method and evaluation index.

Keywords: Anomaly detection, Video surveillance, Deep learning.

1. Introduction

As the number of indoor and outdoor monitoring is increasing, the drawbacks of traditional methods of artificial detection of video anomalies are gradually exposed, such as the negligence of the staff themselves or the complexity of video information, resulting in low efficiency of video surveillance detection tasks. Therefore, adopting intelligent surveillance video to automatically detect abnormal behaviors plays an extremely important role in maintaining public safety and social order [1,2]. Starting from the needs of intelligent monitoring system, foreign scholar summarized the basic framework of automatic anomaly detection monitoring system, and many western developed countries attach great importance to the public transportation pedestrian monitoring project of system structure to improve the efficiency of video surveillance anomaly detection, such as CROMATICA and PRISMATICA. Chinese scholars have also initiated the real-time intelligent video surveillance early warning system, which has been successfully applied in the subway, which is conducive to improving the efficiency of video surveillance in detecting abnormal behaviors, and effectively reducing the crime rate [3-5].

In order to study the abnormal target detection and positioning of the video surveillance system, it is necessary to clarify the meaning of the anomaly, and different videos have different definitions of the anomaly. Therefore, the definition of abnormal criteria depends on the video surveillance data itself, in general, the small probability of events in the video is called abnormal behavior. Secondly, abnormal behavior is divided into global anomaly and local anomaly. Global anomaly refers to the abnormal behavior of all people in the entire scene. Such abnormal behavior starts from a certain frame of video data and appears in the entire video scene [6,7]. For example, when there is a fire in the shopping mall, people are very scared and rush to the exit from all directions. Local anomaly refers to the individual's behavior in the video is different from that of other people around him, such as some people walking in the street to riding a bicycle or others are eating, and one person just sits and plays with his phone. Video anomaly detection is to detect the abnormal world from a large number of video data, so as to better maintain public safety and social order. Generally speaking, it mainly includes three steps. First, the foreground extraction detects the moving target, extracts the behavior characteristics of the target, and detects whether the behavior of the target is abnormal according to the criteria of abnormal behavior recognition and classification [8,9].

Video anomalies usually refer to abnormal appearance or motion attributes in a video, or normal appearance or motion attributes in an abnormal time or space [10]. Due to the scarcity and diversity of abnormal samples, video anomaly detection methods usually only model the normal sample distribution, and the video frames or video clips deviating from the normal sample distribution are regarded as abnormal during testing [11]. From the exception type, appearance exception usually refers to spatial exception, including local exception at pixel level and global exception at frame level. Motion exceptions usually refer to time exceptions, that is, contextual exceptions related to timing. The video anomaly detection task is to detect the time and space anomalies in the video [12]. Since the background of surveillance video in a specific scene is often fixed, surveillance video is a typical single scene video, and the research on video anomaly detection based on a single scene is also the focus of this review [13].

2. Overview of Video Anomaly Detection

For a given normal video data sample in a certain scene, first extract the motion and appearance features of the images in the video frame or video window, and build a model to learn the distribution of the normal samples. During the test, the extracted test sample features are input into the model, and the model determines the anomalies according to the reconstruction error, prediction error, anomaly fraction and peak signal-to-noise ratio.

2.1. Types of Video Anomaly Detection

The basic types of video anomaly detection can be divided into the following two types:

1. Local abnormal and global abnormal. Local anomaly usually refers to an individual's activity significantly deviates from that of its neighbors in a moderate or crowded environment. A global exception refers to a global exception in a particular scenario, perhaps the activity of a local individual may be normal [14].
2. Time anomaly and space anomaly. Time anomaly refers to the anomaly related to motion information, reflecting the change rule between video frames; Spatial anomalies refer to location-related anomalies that reflect the abnormal information inside the video frame [15].

2.2. Learning Paradigm of Video Anomaly Detection

There are four learning paradigms for video anomaly detection, namely supervised, unsupervised, weakly supervised and self-supervised [16-19].

1. Supervised learning. Supervised learning refers to the process of mapping all data samples to labels of different categories through model training by knowing data samples and corresponding labels. For video anomaly detection, it means using normal samples and abnormal samples and corresponding labels to train a binary classifier for anomaly detection. However, due to the scarcity of abnormal video, supervised learning based video anomaly detection methods are rare.
2. Unsupervised learning. Abnormal video detection based on unsupervised learning refers to the learning, clustering or distribution modeling of normal samples without relying on video annotation information and relying on the similarity between sample data. During the test, videos far away from normal samples are regarded as anomalies, so as to achieve anomaly detection. Video anomaly detection using unsupervised learning paradigm usually requires sufficient normal video data.
3. Weakly supervised learning. The video anomaly detection based on weak supervised learning refers to the modeling based on video-level annotation information, and the anomaly detection can be carried out frame-by-frame or video fragment during testing. The video anomaly detection method based on weak supervised learning greatly reduces the labeling dependence on data, no longer relies on frame-by-frame labeling, greatly reduces the workload of data labeling, facilitates the use of large-scale data sets, and further enhances the adaptability of the detection method to different scenarios and the detection performance of different anomaly types.
4. Self-supervised learning. Video anomaly detection based on self-supervised learning means that the model learns directly from unlabeled data without labeling data. Self-supervised learning no longer relies on labeling, but learns the relationship between various parts of the data to mine labels generated by its own supervised information from large-scale unlabeled data to guide its own training. Self-supervised video anomaly detection methods typically consider a more challenging experimental setup that does not rely on any training data.

2.3. Evaluation Method of Video Anomaly Detection

There are two main evaluation methods for video anomaly detection, namely precision first and efficiency first [20,21].

1. Precision first. Precision first video anomaly detection method requires higher accuracy and lower false alarm rate for anomaly detection and location. The purpose of this class of methods is to guarantee high accuracy by using all available training dataset videos, fixed model parameters, and predefined or fine-tuned exception thresholds, but it is difficult to guarantee real-time model performance.

2. Efficiency first. The efficiency-first video anomaly detection method is designed to detect and locate video anomalies with competitive accuracy at the fastest frame processing speed. The purpose of this method is to make the video anomaly detection method have faster real-time processing speed and meet the demand of online detection, so it is more suitable for practical application.

3. Foreground Extraction and Object Detection

Under normal circumstances, abnormal targets in video surveillance are moving objects and targets, but if there is a large area of background and stationary objects in the video, it will increase the difficulty of anomaly detection of the target to some extent, and there may be a variety of noises recorded in the surveillance video, which makes feature extraction and behavior representation more difficult. This reduces the efficiency of anomaly detection to some extent. Therefore, moving object detection is an important part of intelligent anomaly detection system. Traditional detection methods mainly include frame difference method, optical flow method, etc. The so-called detection method refers to the behavior of the moving target through the contrast change between adjacent frames. Optical flow method refers to the expression of a video frame on the imaging plane of a visual motion sensor, and the commonly used streamer methods mainly include HS and HK. At present, when detecting abnormal behaviors of surveillance videos, streamer method is needed to eliminate video background information to obtain two-dimensional images of moving targets and three-dimensional space-time interest blocks [22-24]. For example, Roberto et al conducted effective detection of surveillance videos to obtain relevant two-dimensional images, and then carried out feature extraction and behavior representation. Zhou Shifu et al. also used optical flow method to extract 3D space-time interest blocks of motion information and input them into 3D convolutional networks. This method can effectively reduce the influence of background information on behavior feature information, which is conducive to improving the efficiency and level of abnormal behavior detection.

4. Feature Extraction and Behavior Representation

In the study of abnormal behaviors in surveillance video, efficient behavior feature extraction and rapid detection of abnormal behaviors play an extremely important role in improving the detection and positioning efficiency of abnormal targets in video surveillance system. Therefore, many experts and scholars at home and abroad have also proposed various methods to carry out feature extraction and behavior representation. As far as the current situation is concerned, feature extraction is mainly divided into two categories: one is to extract features through manual design, including texture, spatial interest points, wide flow, etc.; the other is to carry out deep learning on original video frames to obtain deep learning characteristics of moving targets [25,26]. Both feature extraction methods are based on biological neural theory, and the difference is that feature extraction by manual design is to imitate human beings Visual framework to achieve, and deep rest feature extraction is to learn the rules of the data itself.

4.1. Artificially Designed Feature Behavior Representation

Artificial design features are mainly based on the sensitivity of human visual features to extract distinguishable features from images and clarify the relevant physical meaning. At present, the commonly used artificial design features for video anomaly detection mainly include texture features and spatial interest points. For example, Duan et al. [27] used dynamic texture mixing to model the behavior of normal people, and judged the behavior of moving targets according to the standard of abnormal behavior. If there was an outlier, it would be called an abnormal event. On this basis, Mousavi et al. [28] built a complete spatiotemporal texture model from the perspective of spatiotemporal video, and compared the feature space of all video surveillance data in crowd texture with the template behavior, so as to achieve the purpose of anomaly detection. Rabiee et al. [29] described the contrast and correlation of abnormal behaviors from the perspective of statistics, built a framework of abnormal behaviors on this basis, and then expressed abnormal behaviors in the crowd in time and space. From the perspective of significance, Chinese researchers put forward abnormal event detection methods. On the one hand, the spatial and temporal anomaly significance map is constructed by the feature points between continuous video frames, and on the other hand, the spatial anomaly significance map is constructed by color comparison, which is conducive to improving the accuracy and effectiveness of abnormal event detection. MoSIFT is an effective behavior feature description method, which can not only detect the interest points of abnormal targets [30], but also judge the motion intensity of behavior targets by the optical flow intensity of interest points. This method plays an extremely important role in abnormal behavior detection. For example, MoSIFT algorithm is used to extract features of surveillance video, and kernel density estimation is used to select features for MoSIFT, so as to better eliminate the influence of other information on the judgment results.

From the perspective of Harris, the spatio-temporal local structure was established for the local changes in surveillance video, so as to calculate spatio-temporal descriptors, and the concept of spatial interest was continuously expanded outward-so as to continuously improve the effect of abnormal behavior detection in surveillance video. In addition, Zhang et al. [31] analyzed the structure and texture of surveillance video images to obtain spatio-temporal points of interest and accurately and effectively describe the moving targets in the surveillance video in view of the problem of the description of motion behavior targets in color images. In addition, abnormal targets are accompanied by the rapid changes of moving targets. Optical flow method is widely used in abnormal target detection. The normal behavior and abnormal behavior of moving targets can be distinguished by optical flow on moving particles. The optical flow multi-scale histogram is used to extract the behavior characteristics of the target. MHOF can not only express the movement information of the target, but also the spatial information of the target, so as to distinguish the normal and abnormal events, and achieve good results in the global detection. In addition, in order to extract the local regional features of moving targets in surveillance video, the video frame is segmented to obtain optical flow information in each regional space, and based on this, the MRF model is established to detect normal and abnormal events in surveillance video. From the perspective of optical flow technology, researchers use SL-HOF and ULGP-OF to extract video features [32]. SL-HOF can capture space-time interest points, ULGP-OF contains 2D texture descriptors and optical flow algorithm, which is more accurate when locating video foreground information. Then OCELM is used to conduct deep learning for the two descriptors, and a general model of normal events is built on this basis.

In order to better carry out abnormal detection of surveillance video, Zhang et al. [33] used binary method to detect abnormal objects, segmented video frames in time to find spatiotemporal points of interest, then used binary wavelet difference to re-encode spatiotemporal points of interest, and used GMM to re-model foreground occupancy and Hamming distance to complete anomaly detection and positioning targets. In addition, the monitoring target would produce a trajectory during its movement, and the trajectory mainly included the length, position and degree of movement of the target. Through reclassifying the trajectory of different lengths and positions, all kinds of trajectory of each group were re-modeled, so as to distinguish normal events from abnormal events. The common anomaly detection method of sparse track reconstruction is to extract the approximate value of least squares cubic spline curve in surveillance video to complete the abnormal event detection. There are also certain limitations of the pixel detection method, there is no way to detect all abnormal behavior, only to measure the speed and direction of the anomaly. For example, the rest of the vehicles are going east, only one car is going west or the road speed limit of 80km/h, there is a car speed reached 90km/h, but it is difficult to detect human body movement abnormalities, such as thieves or terrorist-related movement abnormalities can not be detected. Therefore, this paper integrates the two detection methods to better express the movement trajectory of the target object, detect not only the speed and direction of the target, but also the local action of the target, which is conducive to improving the efficiency of anomaly detection and effectively reducing the calculation guidance of the algorithm. From the perspective of trajectory optimization, the object anomaly detection system is mainly divided into two parts. First, the trajectory information of moving objects is effectively handled with anomalies, and if abnormal behavior is detected, the alarm should be timely reported. Second, intensive video analysis algorithm is adopted to detect whether abnormal events are related to people [34].

4.2. Deep Behavior Feature Representation

The method of extracting artificial design features through manual design has many theoretical foundations, but it is seriously affected by human factors, and there is no way to objectively describe the behavior of moving objects. There is also a better degree of dependence on the database for this method of extracting features, but not all data can be collected into the database, and there is no way to compare the videos that are not stored in the database.

With the rapid development of deep learning theory and convolutional neural theory, it provides a new direction for the research of computer vision. Duan et al. [35] proposed that a parallel dual-flow network could perform feature learning and behavior judgment on the spatial information and optical flow diagram of RGB images, and effectively classified the discriminant results of the two networks. According to a large number of experimental research results, the dual-flow network has a good effect on feature learning and behavior judgment. Many researchers also improve two-stream network algorithms from various perspectives, such as convolutional two-stream network and temproal segment networks, etc., based on the two-stream network, they carry out two-dimensional feature learning for a single frame of surveillance video, and then use optical flow to express the relationship between frames, so as to make up for the shortage of space-time information. Chen et al. [36] proposed that the deep 3D convolutional neural network took continuous frames of video as input objects, so as to better obtain time domain information of video frames, which was conducive to solving the problem of classification of moving objects in surveillance video. Pei et al. [37] used 3D convolutional neural networks to effectively solve the problem of abnormal behavior detection and location, and learned the space-time interest blocks existing in surveillance videos directly as C3D inputs. At the same time, Wang et al. [38] adopt the method of cascade 3D neural network, and then complete the detection of abnormal targets and the determination of targets through the detection of C3D in the surveillance video by 3D autoencoder. In addition, there are many deep learning methods in the field of abnormal target detection, such as SSD, YOLO, etc., providing a new idea for abnormal target detection in video surveillance. Shang et al. [39] applied the idea of Faster-RCNN to the location of abnormal targets in video surveillance networks, combined with C3D networks to obtain R-C3D networks, so as to better detect and locate abnormal targets in surveillance videos. In addition, C3D network is also tested by CDC network. By applying convolution and deconvolution technology to the field of abnormal target detection in video surveillance, it can accurately predict the information of each frame of surveillance video during end-to-end learning, and achieve good abnormal target detection and positioning effect.

5. Video Anomaly Detection Method Based on Deep Learning

Different from most classification strategies based on learning paradigm and modeling process, this paper starts from the basic principle of distinguishing normal video from abnormal video, and divides deep learning-based video anomaly detection methods into reconstruction-based methods, prediction-based methods, classification-based methods and regression-based methods.

5.1. Reconstruction-based Method

The core idea of video anomaly detection method based on reconstruction is to obtain distribution representation of normal video data by training normal video data [40]. In the process of testing, the normal test sample will have a small reconstruction error, while the abnormal sample has a large reconstruction error, so as to realize the abnormal detection of video. If x represents a video clip or video frame, g represents the neural network that reconstructs x, $f()$ represents the function that calculates the reconstruction error between x and $g(x)$, and ε represents the reconstruction error. Reconstruction-based deep learning methods can be seen as reconstruction-based errors in minimization equation (1).

$$\varepsilon = f(x, g(x)). \tag{1}$$

A common reconstruction method is autoencoder. Danelljan et al. [41] proposed an appearance and motion deep net (AMDN). It could extract the appearance and motion information of the video at the same time, and use multiple single-class support vector machines (SVM) to predict the anomaly score of each input, and finally integrate the score for the final anomaly detection. Alom et al. [42] proposed two methods based on autoencoders. First, the traditional manual method was used to extract spatiotemporal features, and a fully connected autoencoder was learned on it. Secondly, a full-convolutional feed-forward autoencoder was built to learn local features and a classifier as an end-to-end learning framework, which could detect video anomalies with little or no supervision. However, due to the strong learning ability of deep neural networks, autoencoders could sometimes not only reconstruct normal samples better, but also make abnormal samples have small reconstruction errors. To solve this problem, Min et al. [43] proposed an improved AutoEncoder called MemAE (Memory-augmented AutoEncoder). When given an input, the method first took the encoding from the encoder and then uses it as a query to retrieve the most relevant memory item for reconstruction. Yan et al. [44] used a memory module with an update scheme to make the items of the recorded data prototype pattern constantly updated to better remember normal samples, and achieved anomaly detection results comparable to the most advanced methods at that time on the open benchmark data set.

Another commonly used method for video anomaly detection based on reconstruction is sparse coding, whose idea is to construct a set of dictionaries that can express normal videos, so that normal videos can be reconstructed well through this dictionary, while abnormal videos will become fuzzy or even impossible to reconstruct. Suppose the input video features $X = [x_1, x_2, \cdots, x_k]$, where each x_i represents a normal video frame feature. The goal of the sparse coding method is to find an optimal dictionary D, so that X can be reconstructed by the sparse coefficient $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_k]$. Where D and α are obtained by alternating iterative optimization. Its objective function is shown in equation (2):

$$\underset{\alpha}{min} = ||X - D\alpha||_2^2 + \lambda||\alpha||_1. \tag{2}$$

Chu et al. [45] proposed an unsupervised dynamic sparse encoding method to detect abnormal events in videos. The method first extracted spatiotemporal interest points from input video sequences, learned dictionary based on contextual video data, and determined anomalies according to whether query events could be reconstructed from the dictionary base in the test process. Considering that the size of the dictionary would affect the computational complexity of the model, Luo et al. [46] designed a dictionary selection method with sparse consistency constraint, and the purpose of dictionary compression was achieved by introducing sparse reconstruction cost (SRC). In addition, Wu et al. [47] pointed out that the anomaly detection method based on dictionary learning was very time-consuming in the sparse coefficient iterative optimization process. Therefore, a temporally-coherent sparse coding (TSC) network was proposed to encode adjacent frames with similar reconstruction coefficients, and a special type of stacked recurrent neural network (sRNN) was used to map TSC, so as to achieve accelerated parameter optimization. Narasimhan et al. [48] improved TSC and superimposed another layer on sRNN to reduce the calculation cost of alternating dictionary and sparse coefficient updates in the optimization process. Considering the importance of real-time video anomaly detection, Duives et al. [49] proposed a two-flow neural network to extract STFF (spatio-temporal fusion feature), and used FSCN (fast sparse coding network) to construct a dictionary for STFF. Compared with traditional networks, FSCN not only had a testing speed hundreds of times faster, but also reached an advanced level of accuracy.

5.2. Prediction-based Approach

Prediction-based video anomaly detection methods usually assume that a continuous normal video has some regular context connection, which can learn this dependency and better predict future frames, while abnormal videos often violate these dependencies, resulting in unpredictable future frames. When given t frames of continuous video $x_1, x_2, \cdots, x_t$, the goal of the prediction model is to predict the next frame $\hat{x}_{t+1}$ and make $\hat{x}_{t+1}$ as consistent as possible with the real x_{t+1}. In the process of testing, the error between the $\hat{x}_{t+1}$ predicted by the model and the real x_{t+1} is used to determine whether the video frame is abnormal. In particular, let h represent the prediction model, and the prediction-based approach can be represented as shown in equation (3):

$$\hat{x}_{t+1} = h(x_1, x_2, \cdots, x_t). \tag{3}$$

Ye et al. [50] proposed a framework for video anomaly detection based on predictive models, using U-shaped Network (U-Net) as a generator to generate future frames. The quality of future frames was constrained by intensity loss, gradient loss and optical flow loss, and the true or false frames were judged by a discriminator to strengthen the prediction ability of the model. Inspired by the processing of time sequence data by Long Short-Term Memory (LSTM), Luo et al. [51] proposed a Convolutional-LSTM (Conv-LSTM) to model the video sequence and, through constraints on the decoding process. It could reconstruct the past frame and predict the future frame, so as to realize the anomaly detection of video. Nayak et al. [52] combined VAE (Variational-AutoEncoder) with Conv-LSTM to propose a Conv-VRNN(Convolutional Variational Recurrent Neural Network) network structure to generate video future frames. Considering the fuzzy phenomenon of future frames caused by mean square error loss function in the prediction process, Slavic et al. [53] used a convolutional network to generate future frames through alternating convolution and corrected linear unit ReLU, and proposed a method to combine multi-scale structure, adversarial training and image gradient difference feature learning strategies to generate clear future frames. In addition, Yan et al. [54] proposed a novel deep predictive coding network (AnoPCN) to solve the anomaly detection problem. The network was composed of a predictive coding module (PCM) and an error refinement module (ERM). PCM was designed as a Conv-LSTM network structure for generating future frames. ERM reconstruction prediction error was introduced to realize anomaly detection by unifying reconstruction and prediction methods into an end-to-end framework.

5.3. Classification-based Approaches

Although the current mainstream models mainly rely on prediction methods based on reconstruction and future frames, there is still some research work that treats this problem as a classification problem. This classification method can be described by a general formula: let x represent the input video frame or video clip, $h()$ represents the mapping function obtained through network training, and y represents the corresponding category, as shown in equation (4):

$$y = h(x), y \in R. \tag{4}$$

Video anomaly detection methods based on classification are mainly divided into two types: single classification and multi-classification. The main idea of video anomaly detection based on single classification method is to train a single class classifier through normal video data. In the testing process, the classifier only needs to judge whether the given data belongs to the class. Inspired by the training of deep models in unsupervised and semi-supervised environments by Generative Adversarial networks (GAN), Xu et al. [55] proposed a video anomaly detection method based on single classification. Mansour et al. [56] proposed a deep single-classification neural network structure, which used stacked convolutional encoders to generate low-dimensional high-level representation information. By combining the adversarial mechanism and decoder, a compact single-class classifier could be trained on the premise of only given normal samples, so as to achieve anomaly detection. For the multi-classification video anomaly detection method, Sultani et al. [57] proposed a method using local features and global features. For local features, image similarity was used to represent time and space features on video cube blocks, and feature vectors of trained autoencoders were used to represent global features, and then the features were sent to Gaussian classifier for binary anomaly detection. Xiang et al. [58] transformed the anomaly detection problem into a binary classification problem of single pair residual classes, used clustering on the features generated by the convolutional autoencoder, and trained a single pair residual class classifier to distinguish clusters. During the test, if the highest classification score obtained by the classifier was negative, it indicated that the sample did not belong to any cluster, that is, it was marked as an exception. In addition to binary classification, Pawar et al. [59] proposed an adaptive intra-frame classification network (AICN) to transform the video anomaly detection task into a multi-classification problem. The network took the raw input, divided the extracted motion and appearance features into several subregions, and classified each subregion. During the testing process, if the test classification result of the subregion was different from the real classification, it was regarded as an exception.

5.4. Regression-based Approaches

In addition to the mentioned restructuring-based, prediction-based, and classification-based approaches, some researchers have also defined this problem as a regression problem [60-62]. The main idea is to take the abnormal score as an evaluation index and set an appropriate threshold. If the abnormal score is higher than the threshold, it will be regarded as abnormal, otherwise it will be normal. Let x represent the input video frame, $k()$ represent the function of the input x, and the real number z represent the outlier score, as shown in equation (5):

$$z = k(x), z \in R. \tag{5}$$

Li et al. [63] proposed a multi-example learning method mainly trained under weak supervision. First, each training video was divided into an equal number of fragments, which constituted a positive example package (containing only normal video frames) and a negative example package (containing at least one abnormal video frame), and used C3D (Convolutional 3D) to extract spatiotemporal features for each video segment. Then the features were input into the neural network for scoring, and the fragments with the highest scores were selected from the positive and negative example packages respectively for training model parameters. Finally, the hinge loss made the model output a high score for abnormal samples and a low score for normal samples, which was judged according to the abnormal score of the model output. On this basis, Liu et al. [64] pointed out that the hinge loss function used in the reference [65] was not smooth and may face the risk of gradient disappearance in the optimization process, and proposed a new loss function to make the model robust to the output anomaly score. Since the extraction of video features was crucial to the output anomaly score, Yu et al. [66] gave up the use of C3D and instead calculated the optical flow information, and then input the calculated optical flow information into the time-enhanced network to output anomaly score. This method significantly improved the performance of anomaly detection. Considering the complexity of manually labeling normal/abnormal video data, Hu et al. [67] designed an end-to-end trainable video anomaly detection method, which could perform representation learning and output anomaly scores without manually labeling normal/abnormal data, so as to realize video anomaly detection.

6. The Advantages and Disadvantages of Abnormal Behavior Classification Methods

The video anomaly detection method based on supervision is easier to operate and understand. The prior knowledge can be used to select training samples, and then the accuracy of abnormal target detection in surveillance video can be improved by repeated inspection. However, the supervision abnormal behavior classification method is greatly affected by subjective factors, and staff need to spend most of their time and energy to select and e-valuate training samples, and this method has relatively high capacity requirements. The calculation speed of semi-supervised abnormal behavior classification method is relatively fast, and it is easy to build a model. However, the model classification effect cannot automatically adjust the abnormal data for the sample, so if the sample data is replaced with a new scenario, it needs to be redesigned and detected. Classification methods based on semi-supervised abnormal behavior, such as sparse representation, are simpler to operate [68], but the calculation process is very complex, requiring a lot of memory, and sensitive to the data parameters in the detection device itself. In addition, this method is easy to judge the sample data not entered into the model as abnormal behavior.

The classification method based on unsupervised abnormal behavior does not require any prior knowledge and can be more convenient to calculate, but requires a large amount of analysis and processing to finally obtain the most reliable results. For example, GAN network can represent the normal behavior of surveillance video through unsupervised mode, but the final abnormal target detection and positioning still need to rely on the comparison with the data of normal video To get the final verdict.

7. Conclusion

To sum up, this paper mainly discusses foreground extraction and target detection, feature extraction and behavior representation, and abnormal behavior recognition and classification methods. With the development of abnormal target behavior detection and positioning technology in surveillance video, certain achievements have been achieved, but there are also certain limitations. For example, it is difficult to extract the features of abnormal moving targets in complex surveillance video backgrounds. The number of abnormal events is relatively small, and the recognition algorithm knowledge of many abnormal events is for a certain camera, which is seriously inconsistent with the actual video surveillance situation. Although there are experts and scholars to conduct in-depth analysis of the motion view captured by multiple cameras, the entire operation process is very complex. Also, whether a behavior is abnormal mainly depends on the specific scene, time and place, so when you change a scene, time and place, you need to re-train modeling. With the rapid development of network information technology, more scenes are modeled to increase the adaptability of scenes, which is conducive to improving the efficiency and level of abnormal behavior detection in surveillance video.

8. Conflict of Interest

All authors disclosed no relevant relationships.

Acknowledgments. None.

References

1. Patrikar D R, Parate M R. Anomaly detection using edge computing in video surveillance system[J]. International Journal of Multimedia Information Retrieval, 2022, 11(2): 85-110.
2. Berroukham A, Housni K, Lahraichi M, et al. Deep learning-based methods for anomaly detection in video surveillance: a review[J]. Bulletin of Electrical Engineering and Informatics, 2023, 12(1): 314-327.
3. Omarov B, Narynov S, Zhumanov Z, et al. State-of-the-art violence detection techniques in video surveillance security systems: a systematic review[J]. PeerJ Computer Science, 2022, 8: e920.
4. Liu T, Yin S. An improved unscented Kalman filter applied into GPS positioning system[J]. ICIC express letters. Part B, Applications: an international journal of research and surveys, 2015, 6(11): 2937-2942.
5. Ingle P Y, Kim Y G. Real-time abnormal object detection for video surveillance in smart cities[J]. Sensors, 2022, 22(10): 3862.
6. Wang L, Zhou Y, Li R, et al. A fusion of a deep neural network and a hidden Markov model to recognize the multiclass abnormal behavior of elderly people[J]. Knowledge-Based Systems, 2022, 252: 109351.
7. Verma K K, Singh B M, Dixit A. A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system[J]. International Journal of Information Technology, 2022, 14(1): 397-410.
8. Kuppusamy P, Bharathi V C. Human abnormal behavior detection using CNNs in crowded and uncrowded surveillanceCA survey[J]. Measurement: Sensors, 2022, 24: 100510.
9. Alairaji R M, Aljazaery I A, ALRikabi H T H S. Abnormal behavior detection of students in the examination hall from surveillance videos[C]//Advanced Computational Paradigms and Hybrid Intelligent Computing: Proceedings of ICACCP 2021. Springer Singapore, 2022: 113-125.
10. Chang C W, Chang C Y, Lin Y Y. A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection[J]. Multimedia Tools and Applications, 2022, 81(9): 11825-11843.
11. Liu T, Yin S. An improved neural network adaptive sliding mode control used in robot trajectory tracking control[J]. International Journal of Innovative Computing, Information and Control,11(5), pp: 1655-1666, October 1, 2015.
12. Wang H, Zhang S, Zhao S, et al. Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++[J]. Computers and Electronics in Agriculture, 2022, 192: 106512.
13. Li D, Wang G, Du L, et al. Recent advances in intelligent recognition methods for fish stress behavior[J]. Aquacultural Engineering, 2022, 96: 102222.
14. Patil N, Biswas P K. Global abnormal events detection in crowded scenes using context location and motion-rich spatio-temporal volumes[J]. IET Image Processing, 2018, 12(4): 596-604.

15. Izakian H, Pedrycz W. Anomaly detection and characterization in spatial time series data: A cluster-centric approach[J]. IEEE Transactions on Fuzzy Systems, 2014, 22(6): 1612-1624.
16. Tian Y, Pang G, Chen Y, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 4975-4986.
17. Barbalau A, Ionescu R T, Georgescu M I, et al. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection[J]. Computer Vision and Image Understanding, 2023, 229: 103656.
18. Li N, Zhong J X, Shu X, et al. Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning[J]. Neurocomputing, 2022, 481: 154-167.
19. Lv H, Yue Z, Sun Q, et al. Unbiased multiple instance learning for weakly supervised video anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 8022-8031.
20. Zhou J T, Du J, Zhu H, et al. Anomalynet: An anomaly detection network for video surveillance[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(10): 2537-2550.
21. Leyva R, Sanchez V, Li C T. Video anomaly detection with compact feature sets for online performance[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3463-3478.
22. Wang X, Che Z, Jiang B, et al. Robust unsupervised video anomaly detection by multipath frame prediction[J]. IEEE transactions on neural networks and learning systems, 2021, 33(6): 2301-2312.
23. Yin S, Zhang Y, Karim S. Large scale remote sensing image segmentation based on fuzzy region competition and Gaussian mixture model[J]. IEEE Access, 2018, 6: 26069-26080.
24. Li H, Teng L, Yin S. A New Bidirectional Research Chord Method Based on Bacterial Foraging Algorithm[J]. Journal of Computers, 2018, 29(3): 210-219.
25. Zhang Y, Zhang M, Cui Y, et al. Detection and tracking of human track and field motion targets based on deep learning[J]. Multimedia Tools and Applications, 2020, 79: 9543-9563.
26. Ding J, Wen L, Zhong C, et al. Video SAR moving target indication using deep neural network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(10): 7194-7204.
27. Duan S, Wang X, Yu X. Crowded abnormal detection based on mixture of kernel dynamic texture[C]//2014 International Conference on Audio, Language and Image Processing. IEEE, 2014: 931-936.
28. Mousavi H, Galoogahi H K, Perina A, et al. Detecting abnormal behavioral patterns in crowd scenarios[J]. Toward Robotic Socially Believable Behaving Systems-Volume II: Modeling Social Signals, 2016: 185-205.
29. Rabiee H, Mousavi H, Nabi M, et al. Detection and localization of crowd behavior using a novel tracklet-based model[J]. International Journal of Machine Learning and Cybernetics, 2018, 9: 1999-2010.
30. Hu Z, Zhang L, Li S, et al. Parallel spatial-temporal convolutional neural networks for anomaly detection and location in crowded scenes[J]. Journal of Visual Communication and Image Representation, 2020, 67: 102765.
31. Zhang T, Yang Z, Jia W, et al. A new method for violence detection in surveillance scenes[J]. Multimedia Tools and Applications, 2016, 75: 7327-7349.
32. Yao H, Hu X. A survey of video violence detection[J]. Cyber-Physical Systems, 2023, 9(1): 1-24.
33. Zhang T, Jia W, He X, et al. Discriminative dictionary learning with motion weber local descriptor for violence detection[J]. IEEE transactions on circuits and systems for video technology, 2016, 27(3): 696-709.
34. Yin S, Li H, Laghari A A, et al. An Anomaly Detection Model Based On Deep Auto-Encoder and Capsule Graph Convolution via Sparrow Search Algorithm in 6G Internet-of-Everything[J]. IEEE Internet of Things Journal, 2024.
35. Duan T, Liu Y, Li J, et al. DuFNet: Dual Flow Network of Real-Time Semantic Segmentation for Unmanned Driving Application of Internet of Things[J]. CMES-Computer Modeling in Engineering & Sciences, 2023, 136(1).
36. Chen L, Du L, Liu Q. A working human abnormal operation recognition approach based on the deep multi-instance sorting model[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2022, 236(4): 1986-1993.
37. Pei Y, Zhang X. Research on Human Action Recognition Model Based on Computer Laplacian Matrix and Convolutional Neural Network[C]//2023 IEEE 6th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2023: 1-6.
38. Wang X, Ding J, Zhang Z, et al. IPNet: Polarization-based Camouflaged Object Detection via dual-flow network[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107303.
39. Shang C, Wu F, Wang M L, et al. Cattle behavior recognition based on feature fusion under a dual attention mechanism[J]. Journal of Visual Communication and Image Representation, 2022, 85: 103524.
40. Wang Y, Qin C, Bai Y, et al. Making reconstruction-based method great again for video anomaly detection[C]//2022 IEEE International Conference on Data Mining (ICDM). IEEE, 2022: 1215-1220.
41. Danelljan M, Bhat G, Gladh S, et al. Deep motion and appearance cues for visual tracking[J]. Pattern Recognition Letters, 2019, 124: 74-81.
42. Alom M Z, Taha T M, Yakopcic C, et al. The history began from alexnet: A comprehensive survey on deep learning approaches[J]. arxiv preprint arxiv:1803.01164, 2018.
43. Min B, Yoo J, Kim S, et al. Network anomaly detection using memory-augmented deep autoencoder[J]. IEEE Access, 2021, 9: 104695-104706.
44. Yan H, Liu Z, Chen J, et al. Memory-augmented skip-connected autoencoder for unsupervised anomaly detection of rocket engines with multi-source fusion[J]. ISA transactions, 2023, 133: 53-65.
45. Chu W, Xue H, Yao C, et al. Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos[J]. IEEE Transactions on Multimedia, 2018, 21(1): 246-255.

46. Luo W, Liu W, Lian D, et al. Video anomaly detection with sparse coding inspired deep neural networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(3): 1070-1084.
47. Wu P, Liu J, Li M, et al. Fast sparse coding networks for anomaly detection in videos[J]. Pattern Recognition, 2020, 107: 107515.
48. Narasimhan M G, S S K. Dynamic video anomaly detection and localization using sparse denoising autoencoders[J]. Multimedia Tools and Applications, 2018, 77: 13173-13195.
49. Duives D C, van Oijen T, Hoogendoorn S P. Enhancing crowd monitoring system functionality through data fusion: Estimating flow rate from wi-fi traces and automated counting system data[J]. Sensors, 2020, 20(21): 6032.
50. Ye M, Peng X, Gan W, et al. Anopcn: Video anomaly detection via deep predictive coding network[C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 1805-1813.
51. Luo W, Liu W, Lian D, et al. Future frame prediction network for video anomaly detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(11): 7505-7520.
52. Nayak R, Pati U C, Das S K. A comprehensive review on deep learning-based methods for video anomaly detection[J]. Image and Vision Computing, 2021, 106: 104078.
53. Slavic G, Campo D, Baydoun M, et al. Anomaly detection in video data based on probabilistic latent space models[C]//2020 IEEE conference on evolving and adaptive intelligent systems (EAIS). IEEE, 2020: 1-8.
54. Yan C, Zhang S, Liu Y, et al. Feature prediction diffusion model for video anomaly detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 5527-5537.
55. Xu K, Sun T, Jiang X. Video anomaly detection and localization based on an adaptive intra-frame classification network[J]. IEEE Transactions on Multimedia, 2019, 22(2): 394-406.
56. Mansour R F, Escorcia-Gutierrez J, Gamarra M, et al. Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model[J]. Image and Vision Computing, 2021, 112: 104229.
57. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6479-6488.
58. Xiang T, Gong S. Video behavior profiling for anomaly detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(5): 893-908.
59. Pawar K, Attar V. Deep learning approaches for video-based anomalous activity detection[J]. World Wide Web, 2019, 22(2): 571-601.
60. Yin S. Object Detection Based on Deep Learning: A Brief Review[J]. IJLAI Transactions on Science and Engineering, 2023, 1(02): 1-6.
61. Jiang M, Yin S. Facial expression recognition based on convolutional block attention module and multi-feature fusion[J]. International Journal of Computational Vision and Robotics, 2023, 13(1): 21-37.
62. Jiang Y, Yin S. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment[J]. Computer Science and Information Systems, 2023, 20(4): 1869-1883.
63. Li Y, Ni P, Li G, et al. Effective piecewise CNN with attention mechanism for distant supervision on relation extraction task[C]//5th International Conference on Complexity, Future Information Systems and Risk. SciTePress, 2020: 53-62.
64. Liu Z, Tong J, Gu J, et al. A semi-automated entity relation extraction mechanism with weakly supervised learning for Chinese Medical webpages[C]//Smart Health: International Conference, ICSH 2016, Haikou, China, December 24-25, 2016, Revised Selected Papers. Springer International Publishing, 2017: 44-56.
65. Shen J, Li M, Zhang J. Temporal action detection methods based on deep learning[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2022, 36(03): 2252005.
66. Yu Z. Research on Video Detection Method of Mudslide based on Inflated 3D Convolutional Neural Network[J]. Frontiers in Computing and Intelligent Systems, 2023, 5(1): 103-106.
67. Hu K, Shen C, Wang T, et al. Overview of temporal action detection based on deep learning[J]. Artificial Intelligence Review, 2024, 57(2): 26.
68. Fan Y, Li H, Sun B. Cycle GAN-MF: A Cycle-consistent Generative Adversarial Network Based on Multifeature Fusion for Pedestrian Re-recognition: Cycle GAN-MF[J]. IJLAI Transactions on Science and Engineering, 2024, 2(1): 1-9.

Biography

Yuchang Si is with the Software College, Shenyang Normal University. His research direction is image processing, computer application and AI.