Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2022.027182
Article

Tech Science Press

Multi Chunk Learning Based Auto Encoder for Video Anomaly Detection

Xiaosha Qi[1], Genlin Ji[2,*], Jie Zhang[2] and Bo Sheng[3]

[1]School of Mathematical Sciences, Nanjing Normal University, Nanjing, 210023, China
[2]School of Computer and Electronic Information/Artificial Intelligence, Nanjing Normal University, Nanjing, 210023, China
[3]Department of Computer Science, University of Massachusetts Boston, Boston, 02125, USA
*Corresponding Author: Genlin Ji. Email: glji@njnu.edu.cn
Received: 12 January 2022; Accepted: 24 February 2022

Abstract: Video anomaly detection is essential to distinguish abnormal events in large volumes of surveillance video and can benefit many fields such as traffic management, public security and failure detection. However, traditional video anomaly detection methods are unable to accurately detect and locate abnormal events in real scenarios, while existing deep learning methods are likely to omit important information when extracting features. In order to avoid omitting important features and improve the accuracy of abnormal event detection and localization, this paper proposes a novel method called Multi Chunk Learning based Skip Connected Convolutional Auto Encoder (MCSCAE). The proposed method improves the accuracy of video anomaly detection by obtaining more vital information. In the data sorting phase, non-uniform chunking is proposed to divide the video frame into several chunks of different sizes to avoid obtaining unnecessary information and omitting crucial information. In order to well reflect the abnormal motion of objects in the video, a new feature, the inter frame flow feature, which is obtained by merging inter frame difference and optical flow features, is proposed to extract motion feature. Moreover, in this paper, skip connection in the auto encoder is utilized during the training phase to reduce the reconstruction error between the original frames and the reconstruction frames, so that the reconstruction error can be used to detect abnormal events during testing. Experience on three public datasets verifies the effectiveness and accuracy of our proposed method. Experimental results show that the proposed method can detect and locate abnormal events outperforms other recent methods significantly.

Keywords: Auto encoder; non-uniform chunking; fusion feature; skip connection; anomaly detection

1 Introduction

With the rapid development of science and technology, the range of application for surveillance cameras continues to expand [1], which leads to an increasing amount of video that need to be detected. Traditional video anomaly detection and location methods cannot meet the growing social needs. Therefore, in the field of computer vision, many scholars are exploring how to innovate and improve video anomaly detection

methods. Various studies at this stage show that the difficulties [2] in video anomaly detection mainly lie in the variation of scenes and abnormal behaviors: 1) Scene dependence: the required detection methods for each video scene are different, but existing video anomaly detection methods are usually focus on only one or a few scenes, not all scenes; 2) Fuzzy definition: the diversity of scenarios leads to the diversity of abnormal events, resulting in no clear boundary between normal and abnormal; 3) Rarity: it is sometimes impossible to detect whether an object or behavior is abnormal or not since abnormal samples in real scenes are much smaller than normal samples.

In recent years, scholars usually propose corresponding video anomaly detection methods and establish models for specific anomalies [3]. For example, to identify whether pedestrians exhibit the same type of abnormal behavior or action, such as going over railings or falling, researchers usually use human trajectory or action recognition [4–6] methods to detect whether pedestrians are abnormal in the video. However, in non-open scenes and dense traffic, pedestrians or vehicles [7] can be obscured, which poses some troubles for researchers adopting both methods, allowing them to extract only non-occluded features. Lacking occluded features, complete and accurate features cannot be obtained. Thus, the accuracy of anomaly recognition is reduced. In view of the shortcomings of the above methods, this paper proposes a feature extraction fusion method [8] called inter frame flow method, which extracts motion features by fusing inter frame difference method and optical flow method, while using Histogram of Oriented Gradient (HOG) to extract texture information in the video and obtain apparent features. Motion features are very important in video anomaly detection. If only apparent features are used, the lack of motion features will lead to abnormal behaviors such as running and jumping of pedestrians, and abnormal events such as speeding and slowing of vehicles cannot be accurately detected and localized. The inter frame flow method proposed in this paper can extract relatively complete motion features and avoid information omissions caused by slow moving objects, while HOG can clearly display a variety of texture information contained in the video. By combining the two different features of inter frame flow and HOG can better obtain more complete video features, thus enhancing the accuracy of video anomaly detection model. The objective of this paper is to improve the accuracy of detecting abnormal events in videos and accurately locate these abnormal events, which will be achieved by the following three main contributions:

(1) To reduce the loss of information in video frames, non-uniform chunking is proposed instead of uniform chunking to extract enough features from a limited amount of video.
(2) In order to avoid information omission caused by too slow motion, a new fusion feature, the inter frame flow feature, is presented as a motion feature.
(3) A new deep learning based video anomaly detection method MCSCAE, which can detect and locate abnormal events well, is proposed in this paper.

The rest of the paper is organized as follows. Section 2 presents related work on two important steps: ***Feature Extraction*** and ***Model Establishment***. Section 3 details our proposed MCSCAE. Section 4 describes the experimental settings and results evaluation of MCSCAE on three publicly available datasets: University of California, San Diego Pedestrian1 (UCSD Ped1), University of California, San Diego Pedestrian2 (UCSD Ped2) [9] and The Chinese University of Hong Kong (CUHK) Avenue [10]. Section 5 summarizes and outlooks this work and possible future improvements.

2 Related Work

Video anomaly detection can usually be divided into four steps [11]. 1) Data sorting: the video set is divided into frame level images; 2) Feature extraction: the features of video frame are extracted by feature extraction methods; 3) Model establishment: update the model parameters through the training data set to obtain the trained model; 4) Testing: the accuracy of the trained model is judged by the test

set. Among them, the most important steps in existing video anomaly detection methods are: ***Feature Extraction***, which focuses on extracting useful information from video frames, and ***Model Establishment***, which focuses on setting up a model for the training and testing of features.

Feature Extraction: The completeness of extracted features is a critical indicator for accurate detection of abnormal events in video. According to literatures, researchers usually employ manual design or deep learning to extract video frame features [12,13]. Considering different feature clarity of different chunks in the video frame, Leyva et al. [14] extract video foreground occupancy features and optical flow features by Gaussian Mixture Model (GMM) and extract HOF features by dictionary model, finally the two features are fused to detect anomalies in online datasets. Since human skeleton features can clearly represent human postures, literature [15] obtains dynamic skeleton features by dividing skeleton motion into overall body motion and local body posture. In this way, human related video abnormal events can be effectively detected by skeleton motion sequences. Based on the dual flow framework, Li et al. [16] adopt 3-dimension (3D) gradient maps and optical flow maps to extract appearance features and motion features. After integrating these two features, the detection results can be more comprehensive. Literature [17] proposes two new features to jointly represent video features. The local motion based video features can extract the spatial distribution information of 3D local area of video well, while the Principle Component Analysis (PCA) based texture motion features can seamlessly combine optical flow and texture information. Combining these two feature methods, the model can better detect and locate video abnormal events. Since single features are not sufficient to handle variations about video frames, Kumar et al. [18] propose multi-feature fusion to capture high level relationships between features and diminish low level relationships. Multi-feature fusion can prevent events from being misidentified during background clutters, occlusion, and fast motion, thus improving the accuracy of detection.

Model Establishment: Similar to feature extraction, model establishment can be based on traditional statistical methods, such as Markov and Gaussian methods. There are also deep learning [19] based methods such as Generative Adversarial Networks (GAN) [20] and Convolutional Neural Networks (CNN) [21]. To solve the impact of background noise on the accuracy of video anomaly detection models, Luo et al. [22] propose a prediction network based on spatial temporal graph convolutional networks for skeleton-based video anomaly detection. This paper applies graph convolutional networks on skeleton-based video anomaly detection for the first time and describes the graphical connectivity of joints in normal data by building a normal graph, where abnormal events will be judged by the abnormal values of this graph to determine their existence. Literature [23] proposes a weakly labelled deep multiple instance learning (MIL) based model, which can well avoid the time-consuming process when labeling abnormal videos, reduce the time spent on model training and testing. Experiments in this literature show the superiority of the model by introducing large-scale videos consisting of various real anomalies. Considering the advantages of traditional methods and deep learning methods in video anomaly detection, literature [24] combines the two and proposes a depth probability model for video anomaly detection. In this model, video anomaly detection is regarded as an unsupervised outlier detection task. However, this model cannot solve the problem regarding the impact of target detector accuracy. Deepak et al. [25] propose two novel methods based on a multi-view representation learning framework. The first approach is hybrid multi-view representation learning, which combines deep features extracted from a 3D spatiotemporal auto encoder (3D-STAE) with robust handcrafted features based on gradient-based spatiotemporal autocorrelation. The second approach is a deep multi-view representation learning, which combines deep features extracted from dual-stream STAEs to detect anomalies. The results on the datasets show that the proposed multi-view representation modeled with one-class Support Vector Machine (SVM) performs significantly better than recent state-of-the-art methods. Literature [26] presents an intelligent video anomaly detection and classification using a faster Region-CNN (RCNN) with deep reinforcement learning model. The presented model consists of two components: Faster RCNN and Deep

Reinforcement Learning (DRL). Experimental results show that the model performs better than other methods. Feng et al. [27] develop a multiple instance self-training framework (MIST) to efficiently refine task-specific discriminative representations using only video-level annotations. MIST is composed of a multiple instance pseudo label generator and a self-guided attention boosted feature encoder. After self-training to optimize these two components, a task-specific feature encoder can be obtained, and experiments on two public datasets demonstrate the efficacy of MIST.

In short, previously proposed methods usually miss some valid information or are too complex. Our proposed method enhances the accuracy of the model by reducing information omission and increasing the simplicity and understandability of the model.

3 Proposed Method

In this section, firstly, the video frames are unevenly segmented into multiple chunks to extract appearance and motion features. Secondly, in the model training phase, corresponding models are established on different feature sets. Finally, multiple models are integrated into MCSCAE to detect abnormal events.

3.1 Overall Structure and Processing Flow

To address common problems in video anomaly detection, such as degradation of video anomaly detection accuracy due to missing video features, this paper divides video frames, fuses features and improves the convolutional auto encoder to reduce the average reconstruction error of video frames. In this way, the accuracy of video anomaly detection is enhanced. The main steps of the proposed method are as follows (shown in the Fig. 1):

(1) Data sorting: After splitting the original video set into multiple frame level sequences $\{frame_1, frame_2, \ldots, frame_n\}$, the video frames in the frame level sequences [28] are unevenly partitioned to obtain multiple chunks of size $x_n * y_n$ which are non-overlapping with each other

(2) Motion feature extraction: As single optical flow feature extraction often suffers from information omission, the inter frame difference method is used to obtain difference images by subtracting the previous frame from the next frame, and then the obtained images are binarized to obtain the motion target positions. Similarly, the same location chunks corresponding to adjacent video frames are input into Flownet2 [29] to obtain the optical flow features of each chunk in the video frame. By fusing the inter frame difference feature and optical flow feature, the inter frame flow feature is obtained, which can avoid the great omission of information due to the slow speed of moving objects.

(3) Apparent feature extraction: In video frames, apparent features are the representation and shape of local targets. HOG [30] can well describe the representation and shape of the directional density distribution of gradients or edges in video frames. Apparent features can be obtained by HOG.

(4) Anomaly detection model: Chunks of features extracted from training and testing samples are fed into MCSCAE for training and testing. During the training phase, MCSCAE learns normal modes through inter frame flow features and HOG features extracted from normal training samples. In this way, the test sample features can be reconstructed more accurately during testing, and the sample chunks can be determined to be abnormal or not based on the obtained reconstruction errors. Therefore, the trained MCSCAE is capable of reconstructing the test sample features more accurately and judging whether the sample chunks are abnormal or not.

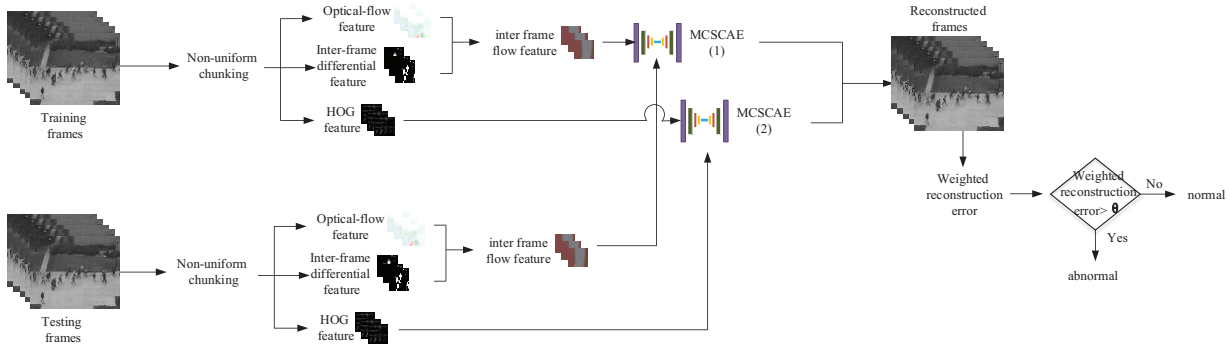

Figure 1: Structure diagram of video anomaly detection method

3.2 Feature Extraction

Although uniform chunking improves the accuracy of anomaly location, it is still a possibility of extracting unnecessary information and omitting important information in feature extraction. To reduce the probability of this happening, non-uniform chunking is utilized to divide the video frames. Non-uniform chunking mainly divides the video frames from far to near and from small to large, in which the video frame width is Y, the chunk width is $\{y_0, y_1, \ldots, y_n\}$, the frame length is X, and the chunk length is $\{x_0, x_1, \ldots, x_n\}$. The width y_n of each chunk can be calculated by the growth rate λ:

$$Y = \sum_{k=0}^{n-1} \lambda^k * y_0 \tag{1}$$

Similarly, the neural network characteristic stipulates that the length x_n and width y_n of the input features must be equal, which means $y_n = x_n$. Thus, the number of chunks with different lengths is calculated as follows:

$$m = X/x_n \tag{2}$$

where n = 6. When x_n cannot be divisible, round up.

Then, apparent features are extracted by HOG, while motion features are extracted from a limited number of video volumes based on inter frame difference features and optical flow features. The dimensions of each frame chunk are $b_x * b_y * b_t$, where b_x and b_y respectively correspond to the horizontal and vertical dimensions of that chunk, and b_t denotes the number of consecutive frames.

Inter frame difference feature and optical flow feature can properly describe the motion anomalies like crowd panic, fights and other sudden variations. These two features are fused together in this paper to obtain a new fusion feature called inter frame flow feature. It is easy to see from Fig. 2 that the inter frame flow feature can reflect the abnormal motion of objects well and reduce the detection error rate. The optical flow feature [31] and the inter frame difference feature are formulated as follows:

$$Opt(x,\ y,\ t) = \frac{1}{Countp} \sum_{cp=1}^{Countp} ||(v_x^{(cp)},\ v_y^{(cp)})||_2 \tag{3}$$

$$Ifd(x,\ y) = |frame_{fn}(x,\ y) - frame_{fn-1}(x,\ y)| \tag{4}$$

where $Countp$ is the total number of pixels in a chunk, and $v_x^{(cp)}$ and $v_y^{(cp)}$ respectively correspond to the horizontal and vertical components of the optical flow. Similarly, fn represents the number of frames.

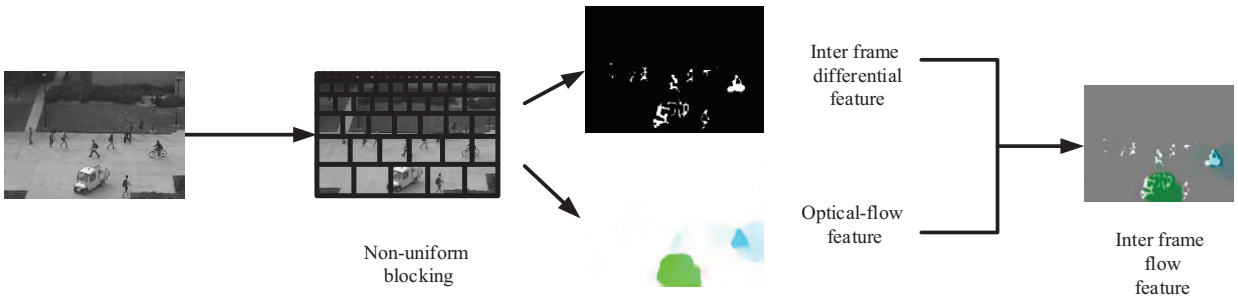

Figure 2: Fusion feature extraction

3.3 Multi Chunk Learning Based Skip Connected Convolutional Auto Encoder (MCSCAE)

Based on the auto encoder, MCSCAE is designed to learn the features of the training samples for detecting the presence of abnormal events in the testing phase. Full connection layer is used in the basic structure of traditional auto encoder, which causes the 2D images lose some spatial information. While MCSCAE utilizes convolution to convert the input chunk, which can effectively retain the required spatial information, and the use of skip connections in convolution and deconvolution also improves the accuracy (shown in Fig. 3). Unlike other literatures using traditional fully connected auto encoders, this paper obtains a trained MCSCAE by reconstructing the inter frame flow features and HOG features of the test samples, and this model can calculate the reconstruction error precisely. In this way, it is possible to judge whether an event is abnormal or not with low error and obtain the specific location of the abnormal event.

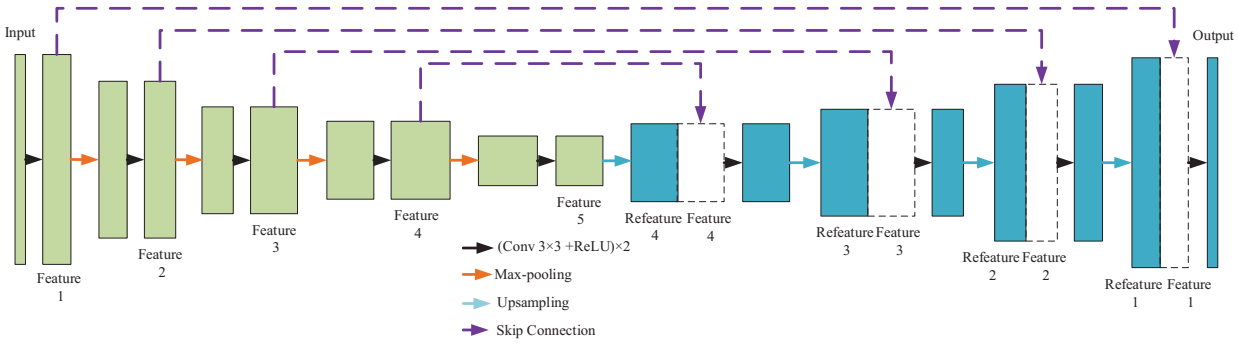

Figure 3: Overview of MCSCAE structure

MCSCAE is mainly composed of an encoder and a decoder [32]. The encoding part consists of two 3 × 3 convolutional layers and a 2 × 2 pooling layer repeatedly. With each down-sampling, the channels are doubled and the latent features are obtained. While a 2 × 2 up-sampling layer, the skip connection function, and two 3 × 3 convolutional layers are repeated to form the decoding part. This structure reconstructs the latent features and resize them to the same size as the input features before output. In this case, the size of input features is resized to 128 × 128 × 3.

The probability distribution of the output classification is obtained by processing the output of MCSCAE with the *softmax* [33] function and comparing it with the original classification probability distribution. Loss is calculated by cross entropy. *Softmax* function and cross entropy loss [34] function are expressed as follows:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{u} e^{z_i}} \tag{5}$$

$$Loss = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[z_i * \log(p_i) + (1 - z_i) * \log(1 - p_i)] \tag{6}$$

where u is the number of classifications, z_i is the model output, i is the input sample, p_i represents the probability of sample i being predicted as a positive class. From Eq. (6), loss of chunks can be obtained, and then the reconstruction error between the original chunks and output chunks can be obtained. In short, the smaller the reconstruction error during training, the higher the accuracy of the model obtained when testing.

3.4 Anomaly Detection

In the test phase, the inter frame flow features and HOG features extracted from the video chunks are respectively put into the trained MCSCAE to calculate the reconstruction [35] error. When the reconstruction error is large, the test sample is abnormal, otherwise, the test sample is normal. Based on the obtained HOG feature reconstruction error $Loss_{hog}$ and inter frame flow feature reconstruction error $Loss_{iff}$, the total reconstruction error is calculated. The formula of $Loss_{sum}$ is as follows:

$$Loss_{sum} = \alpha * Loss_{hog} + \beta * Loss_{iff}, \quad \alpha + \beta = 1 \tag{7}$$

where α is the reconstruction error weight of HOG features and β is the reconstruction error weight of inter frame flow features. Particle swarm optimization (PSO) [36] is used to optimize the weights α and β with the following formulas:

$$v_i = \omega * v_i + c_1 * rand * (pbest_i - x_i) + c_2 * rand * (gbest_i - x_i) \tag{8}$$

$$\omega = (\omega_{int} - \omega_{end}) * \frac{G_k - g}{G_k} + \omega_{end} \tag{9}$$

Among them, v_i is the speed of the particle, c_1 and c_2 are learning factors, and $c_1 = c_2 = 2$, *rand* is a random number between 0 and 1, $pbest_i$ and $gbest_i$ are the extreme values of particle update parameters, x_i is the current position of particles, ω is the inertia factor which can be calculated by Eq. (9), G_k is the maximum number of iterations, $\omega_{int} = 0.9$ is the initial inertia weight, and $\omega_{end} = 0.4$ is the inertia weight when iterating to the maximum generations.

Set θ as the reconstruction error threshold, which is the judgment standard for video abnormal events. When loss is greater than θ, the video chunk is abnormal, otherwise, no abnormal events occur. The judgment formula is as follows:

$$F = \begin{cases} 0, & Loss \le \theta \\ 1, & Loss > \theta \end{cases} \tag{10}$$

The video frame is divided into multiple chunks, then the abnormal detection model judges whether anomalies exist in a certain chunk, so that the location of abnormal events can be performed at the same time as detecting them. The training process of video abnormal detection model is shown in Algorithm 1.

Algorithm 1: MCSCAE

Input: Training video frames *FS*, the number of training video frames *FC*, chunks *P*, the number of chunks *PN*, the number of iterations *ES*

Output: Reconstruction error *Loss*, *reconIff*, *reconHOG*

```
for each f in FS do:
    divide f unevenly to obtain non-uniform chunks      //get non-uniform chunks
for i = 0 to FC do:
    for j = 0 to PN do:
        { Iff = inter frame flow feature
          Hog = HOG feature}
    for k = 0 to PN do:
        { for e = 0 to ES do:
            { connection and crop Iff
              reconIff = forward(MCSCAE (1), Iff )      //reconstruct inter frame flow feature
              bp(MCSCAE (1), reconIff)
              //back propagation and updates the parameters of MCSCAE (1)
              connection and crop Hog
              reconHog = forward(MCSCAE (2), Hog )   //reconstruct HOG feature
              bp(MCSCAE (2), reconHog)
              //back propagation and updates the parameters of MCSCAE (2)
              e = e + 1}}                  //get trained models
    Compare Iff and reconIff
    Compare HOG and reconHOG
return Loss, reconIff, reconHOG
```

4 Experiment

4.1 Experiment Settings

In order to verify the effectiveness and accuracy of MCSCAE, simulation experiments are implemented in Pycharm using the Tensorflow with NVIDIA GeForce GTX3080ti, and on three publicly available benchmark datasets, i.e., CUHK Avenue, UCSD Ped1 and UCSD Ped2. The scenarios for the UCSD dataset are sidewalks, where ped1 focuses on crowds moving towards and away from the camera, and ped2 is a scene in which pedestrians are moving parallel to the camera. The abnormal events in these scenarios include bicycles and wheelchairs on the sidewalk, running pedestrians and cars. Another scene for the CUHK dataset is Campus Avenue, which contains abnormal events such as pedestrians running, walking in the wrong direction, trucks, bicycles, suspicious items, etc. In these three datasets, the training sets contain only normal events, while the test sets contain both normal and abnormal events. Tab. 1 shows the details of the three datasets.

Table 1: Details of three publicly available benchmark datasets

Datasets	Scenarios	Anomalies	Resolution	Duration
CUHK avenue	Campus	Strange action, wrong direction, abnormal object	640 × 360	30 min
UCSD Ped1	Sidewalk	Non pedestrian objects on sidewalks and	238 × 158	10 min
UCSD Ped2		abnormal pedestrian movement patterns	360 × 240	

Depending on the growth rate λ and initial vertical dimension y_0, the video frames are divided into several non-uniform chunks. For chunks of different size, iterations are set by the decreasing speed of loss.

4.2 Results Evaluation

The Receiver Operating Characteristic (ROC) curve is plotted to measure the detection accuracy. The ROC curve is a curve of True Positive Rate (TPR) *vs.* False Positive Rate (FPR), as shown below:

$$TPR = \frac{TP}{TP + FN} \tag{11}$$

$$FPR = \frac{FP}{TN + FP} \tag{12}$$

where TP represents true positive cases, FN denotes false counter cases, TN means true counter case, and FP indicates false positive case.

Based on the ROC curve, three values are calculated as quantitative indexes: 1) ***Area Under Curve (AUC)*:** Area under the ROC curve. 2) ***Equal Error Rate* (EER):** FPR value when the condition $TPR + FPR = 1$ is satisfied. 3) Reconstruction error (Loss): The smaller the reconstruction error, the closer the input and output are. Notice that AUC and EER are similar performance evaluation metrics, specifically, $\boldsymbol{EER} \rightarrow 0$ when $\boldsymbol{AUC} \rightarrow 1$.

Fig. 4 shows the detected abnormal events in the test samples, which are marked with red masks. The performance of MCSCAE is evaluated against several state-of-the-art methods. The experimental results show that MCSCAE is more competitive in terms of detection accuracy compared to other methods.

CUHK Avenue UCSD Ped1 UCSD Ped2

Figure 4: Visualization of visual abnormal event detection results on the datasets

For the UCSD ped2 public dataset, the loss is shown in Fig. 5. When the literation times are less than 800, the loss decreases exponentially, otherwise it tends to be stable. As expected, our method attains a lower loss to achieve higher model detection accuracy. For the CUHK Avenue dataset, ROC curves are shown in Fig. 6, and the corresponding evaluated results of AUC and EER are shown in Tab. 2. From Tab. 2, our method achieves the highest AUC and the lowest EER.

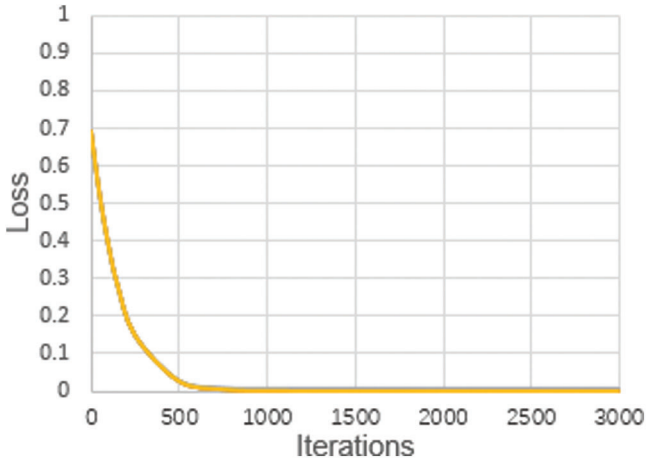

Figure 5: Loss for UCSDped2 datasets

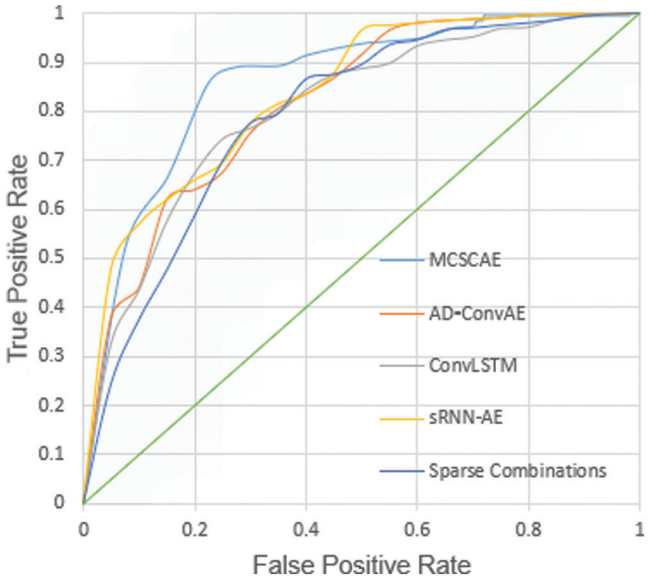

Figure 6: ROC curves for CUHK avenue

Fig. 7 shows ROC curves for the UCSD Ped1 dataset. The experimental results are evaluated in terms of AUC and EER in Tab. 3. As expected, our method attains lower AUC and higher EER than the other three methods compared. Another method achieves the lowest EER, while its AUC is lower than ours. Thus, our method achieves competitive detection accuracy and great performance.

Table 2: Comparison with the state-of-the-art methods for CUHK avenue dataset

Method	AUC (%)	EER (%)
AD-ConvAE [37]	81.65	27.7
ConvLSTM [38]	80.3	20.7
sRNN-AE [39]	83.48	-
Sparse combinations	80.9	-
MCSCAE	**84.34**	**21.1**

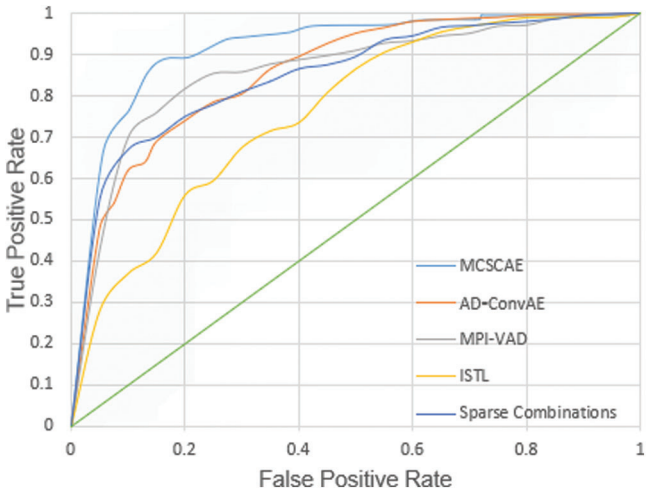

Figure 7: ROC curves for UCSD Ped1 dataset

Table 3: Comparison with the state-of-the-art methods for UCSD Ped1 dataset

Method	AUC (%)	EER (%)
AD-ConvAE	85.92	23.35
Spatial–temporal CNN [40]	85.0	24.0
ISTL [41]	75.2	29.8
MPI-VAD [42]	86.2	17.5
MCSCAE	**89.93**	**20.24**

Fig. 8 shows the ROC curves for the UCSD Ped2 dataset, and Tab. 4 evaluates the corresponding results in terms of AUC and EER. As can be seen from Tab. 4, our method attains the highest detection accuracy compared to other methods.

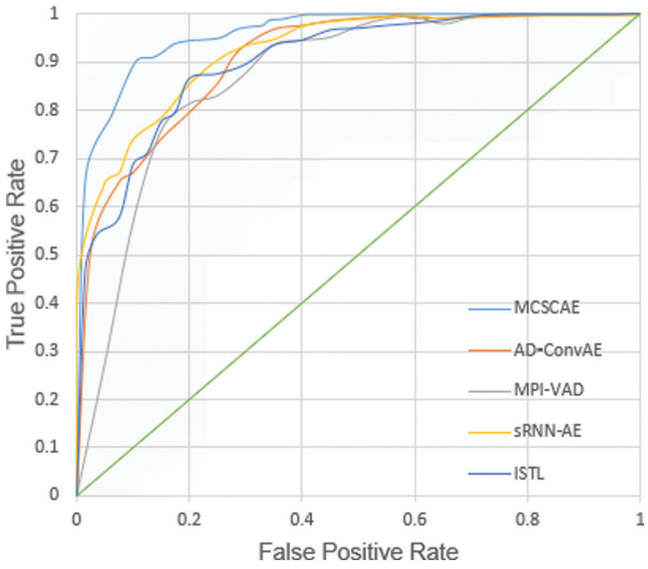

Figure 8: ROC curves for UCSD Ped2 dataset

Table 4: Comparison with the state-of-the-art methods for UCSD Ped2 dataset

Method	AUC (%)	EER (%)
AD-ConvAE	90.45	15.6
sRNN-AE	92.21	-
ISTL	91.1	8.9
MPI-VAD	87.5	16.8
MCSCAE	**94.26**	**10.3**

5 Conclusion

In this paper, a novel video anomaly detection and location method called MCSCAE is proposed. Our method employs a non-uniform chunk structure to extract appearance and motion features from a limited number of videos, in which motion features are fused by inter frame difference feature and HOG feature. Based on multi chunk learning, MCSCAE is designed to learn features of chunks at different positions in normal mode in the video. Experimental results on three publicly available datasets show that MCSCAE is superior to other methods in accurately detecting and locating abnormal events.

Apparently, MCSCAE can extract important features with almost no misses and reduce the reconstruction error between the original frames and the reconstruction frames by skip connection, thus this model can better detect abnormal events in videos with low probability of false detection. Compared to other methods, MCSCAE takes up much less storage space when using multi chunk learning. However, MCSCAE requires considerable time to train multiple chunks, therefore our future work will focus on reducing the required time.

Funding Statement: This work was supported by the National Science Foundation of China under Grant No. 41971343.

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

References

[1] Y. Z. Zhang, S. S. Zhang, Y. Li, J. Y. Zhang and S. Lam, "Surveillance video key frame extraction based on center offset," *Computers, Materials & Continua*, vol. 68, no. 3, pp. 4175–4190, 2021.

[2] T. Li, H. Chang and M. Wang, "Crowded scene analysis: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[3] R. Afzal and R. K. Murugesan, "Rule-based anomaly detection model with stateful correlation enhancing mobile network security," *Intelligent Automation & Soft Computing*, vol. 31, no. 3, pp. 1825–1841, 2022.

[4] E. Esene, A. A. Arabacim and M. Soysalm, "Fight detection in surveillance videos," in *Proc. of the 11th Int. Workshop on Content-Based Multimedia Indexing*, Veszprem, Hungary, pp. 131–135, 2013.

[5] A. Carneiros, P. Silvag and F. Guimaraess, "Fight detection in video sequences based on multi-stream convolutional neural networks," in *Proc. of the 32nd SIBGRAPI Conf. on Graphics, Patterns and Images*, Rio de Janeiro, Brazil, pp. 8–15, 2019.

[6] C. Zhu, Y. K. Wang, D. B. Pu, M. Qi, H. Sun *et al.,* "Multi-modality video representation for action recognition," *Journal on Big Data*, vol. 2, no. 3, pp. 95–104, 2020.

[7] X. J. Xue, Z. X. Wang, L. J. Ge, L. R. Deng, R. Song *et al.,* "Video recognition for analyzing the characteristics of vehicle–bicycle conflict," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 2779–2791, 2021.

[8] M. Y. Duan, J. Liu and S. Q. Lv, "Encoder-decoder based multi-feature fusion model for image caption generation," *Journal on Big Data*, vol. 3, no. 2, pp. 77–83, 2021.

[9] V. Mahadevan, W. Li and V. Bhalodia, "Anomaly detection in crowded scenes," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 1975–1981, 2010.

[10] C. Lu, J. Shi and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *IEEE Int. Conf. on Computer Vision*, Sydney, Australia, pp. 2720–2727, 2013.

[11] R. B. Kiran, M. D. Thomas and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 36, no. 4, pp. 1–25, 2018.

[12] A. Asokan, J. Anitha, B. Patrut, D. Danciulescu and D. J. Hemanth, "Deep feature extraction and feature fusion for bi-temporal satellite image classification," *Computers, Materials & Continua*, vol. 66, no. 1, pp. 373–388, 2021.

[13] J. Chen, Z. Zhou, Z. Pan and C. Yang, "Instance retrieval using region of interest based cnn features," *Journal of New Media*, vol. 1, no. 2, pp. 87–99, 2019.

[14] R. Leyva, V. Sanchez and C. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3463–3478, 2017.

[15] R. Morais, V. Le and T. Tran, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Los Angeles, USA, pp. 11996–12004, 2019.

[16] N. Li and F. Chang, "Video anomaly detection and localization via multivariate Gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, 2019.

[17] S. Wang, E. Zhu and J. Yin, "Video anomaly detection and localization by local motion based joint video representation and OCELM," *Neurocomputing*, vol. 277, pp. 161–175, 2020.

[18] A. Kumar, G. S. Walia and K. Sharma, "A novel approach for multi-cue feature fusion for robust object tracking," *Applied Intelligence*, vol. 50, pp. 3201–3218, 2020.

[19] L. Deng, X. Wang, F. Jiang, and R. Doss, "EEG-based emotion recognition via capsule network with channel-wise attention and LSTM models," *CCF Transactions on Pervasive Computing and Interaction*, vol. 3, no. 4, pp. 425–435, 2021.

[20] X. Duan, S. Ying, W. Yuan, H. Cheng and X. Yin, "A generative adversarial networks for log anomaly detection," *Computer Systems Science and Engineering*, vol. 37, no. 1, pp. 135–148, 2021.

[21] S. Rajendar and V. K. Kaliappan, "Sensor data based anomaly detection in autonomous vehicles using modified convolutional neural network," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 859–875, 2022.

[22] W. Luo, W. Liu and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021.

[23] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6479–6488, 2018.

[24] Y. Ouyang and V. Sanchez, "Video anomaly detection by estimating likelihood of representations," in *The 25th Int. Conf. on Pattern Recognition*, Milan, Italy, pp. 8984–8991, 2020.

[25] K. Deepak, G. Srivathsan and S. Roshan, "Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders," *Circuits, Systems, and Signal Processing*, vol. 40, pp. 1333–1349, 2021.

[26] R. F. Mansour, J. E. Gutierrez, M. Gamarra, J. A. Villanueva and N. Leal, "Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model," *Image and Vision Computing*, vol. 112, pp. 104229, 2021.

[27] J. C. Feng, F. T. Hong and W. S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*, On-line meeting, pp. 14009–14018, 2021.

[28] W. Fang, Y. P. Chen and Q. Y. Xue, "Survey on research of RNN-based spatio-temporal sequence prediction algorithms," *Journal on Big Data*, vol. 3, no. 3, pp. 97–110, 2021.

[29] E. Ilg and N. Mayer, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 1647–1655, 2017.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp. 886–893, 2005.

[31] M. Ponti, T. S. Nazare and J. Kittler, "Optical-flow features empirical mode decomposition for motion anomaly detection," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, pp. 1403–1407, 2017.

[32] C. Xia, J. Peng, Z. Ma and X. Li, "A multi-scale network with the encoder-decoder structure for cmr segmentation," *Journal of Information Hiding and Privacy Protection*, vol. 1, no. 3, pp. 109–117, 2019.

[33] X. Li, X. Li, D. Pan and D. Zhu, "On the learning property of logistic and softmax losses for deep neural networks," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Hilton Midtown, New York, USA, pp. 4739–4746, 2020.

[34] C. Zhang, Y. Hu and X. Zhu, "Anomaly detection for user behavior in wireless network based on cross entropy," in *IEEE 12th Int. Conf. on Ubiquitous Intelligence and Computing and IEEE 12th Int. Conf. on Autonomic and Trusted Computing and IEEE 15th Int. Conf. on Scalable Computing and Communications and Its Associated Workshops*, Beijing, China, pp. 1258–1263, 2015.

[35] L. Wang, T. K. Kim and K. J. Yoon, "EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8315–8325, 2020.

[36] X. Jin, D. Zhou and S. Yao, "Multi-focus image fusion method using S-PCNN optimized by particle swarm optimization," *Soft Computing*, vol. 22, pp. 6395–6407, 2018.

[37] X. Li, G. Ji and B. Zhao, "Convolutional auto encoder patch learning based video anomaly event detection and localization," *Journal of Data Acquisition and Processing*, vol. 36, no. 3, pp. 489–497, 2021.

[38] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. of the Int. Symp. in Neural Networks*, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, pp. 189–196, 2017.

[39] W. Luo, W. Liu and S. Gao, "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1070–1084, 2021.

[40] S. Zhou, W. Shen and D. Zeng, "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, 2016.

[41] R. Nawaratne, D. Alahakoon and S. Silvad, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2020.

[42] Z. Xu, X. Zeng, G. L. Ji and B. Sheng, "Improved anomaly detection in surveillance videos with multiple probabilistic models inference," *Intelligent Automation & Soft Computing*, vol. 31, no. 3, pp. 1703–1717, 2022.