ANOMALY DETECTION WITH FRAME-GROUP ATTENTION IN SURVEILLANCE VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

The paper proposes an end-to-end abnormal behavior detection network to detect strenuous movements in slow moving crowds, such as running, bicycling, throwing from a height. The algorithm forms continuous video frames into a frame group and uses the frame-group feature extractor to obtain the spatio-temporal information. The implicit vector based attention mechanism will work on the extracted frame-group features to highlight the important features. We use fully connected layers to transform the space and reduce the computation. Finally, the group-pooling maps the processed frame-group features to the abnormal scores. The network input is flexible to cope with the form of video streams, and the network output is the abnormal score. The designed compound loss function will help the model improve the classification performance. This paper arranges several commonly used anomaly detection datasets and tests the algorithms on the integrated dataset. The experimental results show that the proposed algorithm has significant advantages in many objective metrics compared with other anomaly detection algorithms.

1 INTRODUCTION

Nowadays anomaly detection is useful to maintain social security and conduct legal forensics. Due to the ambiguous definition of abnormal events, it increases the difficulty of detection. For example, the appearance of a vehicle on a road is normal, while it is abnormal when a vehicle is on the sidewalk. So we make the goal clear in order to carry out consequent measures. Abnormal behavior is defined as rapid movements in slow moving crowds, such as cycling, running, throwing from a height, etc. Based on this definition, the commonly used datasets for anomaly detection are integrated and relabeled. The new fusion dataset contains more scenes and anomaly types, so it is more challenging for the anomaly detection. Subsequent experiments are performed on this new dataset. The test results show that the algorithm proposed in this paper has great advantages in many metrics.

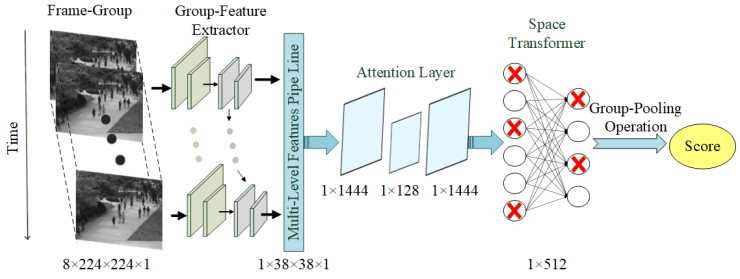


Figure 1: Proposed Anomaly Detection Network

The end-to-end anomaly detection network shown in Fig. 1 has the following contributions:

(1) The video group composed of consecutive multiple frames is the basic processing unit to extract expressive features with the designed group feature extractor. On the one hand, the spatial-temporal information could be retained comparing with single-image processing. On the other hand, the

abnormality score of a single frame can be easily obtained without the access of the whole video, which can copy with the situation where video streams are the input.

(2) The designed framework composed of group feature extractor and group score mapper can effectively obtain the abnormality score using the spatial-temporal information. Implicit vector-based attention mechanism is used to weight the frame-group features. The more important the feature is, the higher the weight is.

(3) The basic cross-entropy loss and the improved hinge loss are united to improve the performance of the network. The latter devote to make the score of the abnormal frames greater than that of normal frames.

The paper is organized as follows. Section 1 introduces the background of anomaly detection. Section 2 introduces the related work on anomaly detection. Section 3 mainly introduces the details of the proposed anomaly detection algorithm. Section 4 introduces the fusion dataset and then analyzes the experimental results. Section 5 gives the summary of the whole paper.

2 Related Work

The challenges of video semantic analysis lie in the extraction and representation of video features. The video contains complex spatial texture information and time information. Multidimensional data provides more information but meanwhile it contains a lot of redundant information. How to extract low-redundant, comprehensive and representative video features is one of the research focuses. The manual extraction method of video features focuses on the extraction and analysis of low-level visual features, such as guided gradient histograms Xiao et al. (2014), optical flow maps Reddy et al. (2011), spatio-temporal points of interest Dollár et al. (2005), texture models Xiao et al. (2018), filtering models Zhang et al. (2018), etc. After obtaining the statistical information, the visual dictionary Roshtkhari & Levine (2013) and other methods will be used to save the normal distribution, and then calculate the similarity criterion to determine whether the target is abnormal. Luo & Wang (2019) explores the method of multi-stream manual features for video representation. It constructs a three-dimensional video representation structure composed of spatio-temporal vector and positional vector, and improves the encoding method to make the extracted video representation structure more powerful in representation.

With the rapid development of deep learning, automatic feature extraction using neural networks has become a research hotspot. Zhou et al. (2016) uses 3D convolutional networks to detect abnormal events in surveillance video. In Sabokrou et al. (2017), the video frame is divided into several small areas, and each small area is feed into a 3D self-encoding network combined with 3D convolutional neural network to extract features and detect anomaly. Medel (2016) proposes a long-short term memory network based on convolution, which simultaneously extracts spatial and temporal information. Xu et al. (2015) first uses stacked auto-encoders to learn and fuse the appearance and motion characteristics of abnormal individuals, and then trains multiple single-classifiers to calculate the abnormal score. Ionescu et al. (2019) uses one-to-many classifiers instead of the single-classifiers after obtaining multiple pseudo-anomaly classes from the trained normal behavior pattern. Hinami et al. (2017) multiple attributes of the same target to extract features.

Due to the lack of anomalous videos and various types of anomaly, it is difficult to find a general model that covers all anomalous events. The auto-encoding network Yuan et al. (2019) performs anomaly detection based on the reconstruction error. Hasan et al. (2016) uses a convolutional neural network to implement video anomaly detection. Since the convolutional layer operates on a two-dimensional structure, time information will be lost. Chong & Tay (2017) designs a spatio-temporal autoencoder that encodes video sequence with spatial convolution and Convolutional Long-Short Term Memory (ConvLSTM) Shi et al. (2015) structure, and then uses the symmetric structure called the decoder, convert the video encoding into the image sequence. The abnormal score can be obtained from the calculation of the Euclidean distance between the decoded images and the original images. An & Cho (2015) proposes a variational autoencoder (Variational Autoencoder, VAE), which uses the results of video encoding to fit a distribution function. Jing & Yujin (2017) supplements a gradient difference constraint on the basis of the sparse denoising auto-encoding network, which is helpful to make the model more effective in detecting abnormal behavior.

For specific tasks, the model trained according to the specific task can often get better results. In Sultani et al. (2018), the video containing abnormal events is divided into several video segments to form multiple video instances. A fully connected network is designed to map the video features extracted by the C3D network Tran et al. (2015) into abnormal scores. The score of the abnormal instance is higher than the score of the instances with only normal frames. The experiments in this paper show that the network has achieved good results. But the detection process is disconnected. In order to minimize the interference to the input data and obtain the final score directly and timely, we designs an end-to-end network, which uses both positive and negative samples to make the model more targeted.

3 Proposed Method

The paper proposes an end-to-end anomaly detection network, which operates on the frame-group formed with consecutive frames to obtain the abnormality score. The whole framework is composed of group feature extractor and group score mapper. The former performs on the raw frame-group to obtain the spatial-temporal group-feature. The latter performs on the group-feature to obtain the abnormality score.

3.1 The Group-Feature Extractor

The frame-group is defined as a structure consisting of consecutive τ frames, which contains rich spatial texture information and temporal change information. Fig. 2 shows the details of the group-feature extractor, where $\mathbf{In}_t^{(0,Spa,Tem)}$ and $\mathbf{Out}_t^{(0,Spa,Tem)}$ represent the input and output matrixes in time t respectively. Spa and Tem mean the input and output feature map in the spatial feature extractor and the temporal feature extractor respectively.

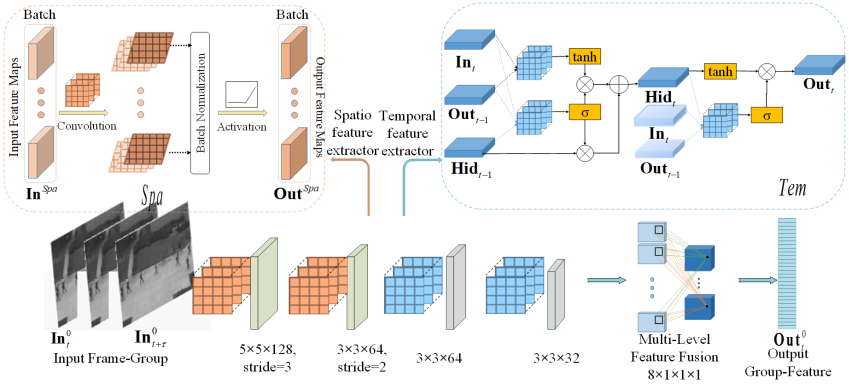


Figure 2: The Group-Feature Extractor

Since the frame-group contains multiple continuous frames, in order not to destroy the time information within consecutive frames, we use a trainable convolution kernel to extract spatial information of each single-frame in the time dimension. Then batch normalization Ioffe & Szegedy (2015) is used to prevent the gradient dispersion and accelerate the model convergence. Activation function helps to make the model nonlinear. Two sets of "convolution-regular-activation" structures are used in the experiments. The first group uses 128 convolution kernels with size of 5×5 and step size of 3, and the second one uses 64 convolution kernels with size of 3×3 and step size of 2.

After obtaining the spatial feature maps, ConvLSTM will further extract the spatial and temporal information using gate structures. The implementation of hidden features is shown in Equation (1). Where the hidden feature $\mathbf{Hid}_t^{Tem}$ is used to record the accumulated state information up to time t. The symbol $\bigotimes$ represents the Hadamard product. σ and $\tanh$ represents the sigmoid and tanh nonlinear activation functions respectively. $Conv$ is the convolution operation. This paper uses two cascading ConvLSTM layers to extract time flow information. The first layer uses 64 convolution kernels with size of 3×3 and the second one uses 32 same kernels. We use the "same padding" to keep the size of input and output feature maps same.

$$\mathbf{Hid}_t^{Tem} = \sigma(Conv(\mathbf{Out}_{t-1}^{Tem}, \mathbf{In}_t^{Tem}, \mathbf{Hid}_{t-1}^{Tem})) \bigotimes \mathbf{Hid}_{t-1}^{Tem}$$
$$+ \sigma(Conv(\mathbf{Out}_{t-1}^{Tem}, \mathbf{In}_t^{Tem}, \mathbf{Hid}_{t-1}^{Tem}) \bigotimes \tanh(Conv(\mathbf{Out}_{t-1}, \mathbf{In}_t^{Tem})) \quad (1)$$

While the final output is shown in Equation (2):

$$\mathbf{Out}_t^{Tem} = \sigma(Conv(\mathbf{Out}_{t-1}^{Tem}, \mathbf{In}_t^{Tem}, \mathbf{Hid}_t^{Tem})) \bigotimes \tanh(\mathbf{Hid}_t^{Tem}) \quad (2)$$

In order to better extract the spatial and temporal information as well as high-level and low-level features of frame-group, we use a multi-level feature fusion structure to merge multi-level features of the frame-group. The output result of the multi-level feature fusion structure is regarded as the final feature representation of the frame-group. In the experiment, the structure was implemented using a 3D convolution Zhou et al. (2018) layer with size of $8 \times 1 \times 1$.

3.2 The Group-Score Mapper

In the group-score mapper, the attention mechanism Ilse et al. (2018) is used to increase the decisive influence of useful features and weaken the effects of irrelevant features on the results. The fully connected network is used to make the encoding of the group-feature more expressive while reducing feature dimensions. The group-level pooling is applied to map the refined group-feature to the abnormality score of the video group. The specific process is shown in Fig. 3.

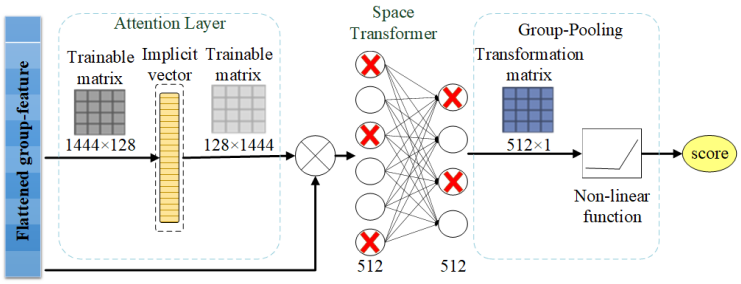


Figure 3: The map from group-features to group-scores

The implicit vector method based attention mechanism assigns different weights to different features in order that the key features have a more important impact on the result and the interference of noise can be suppressed. The trainable transformation matrix will be used to project the original group feature into the implicit space, and then the weight vector will be obtained by an inverse transformation matrix. Different from the attention mechanism aiming at multiple instances in Ilse et al. (2018), this paper focus on the attention of the group-feature. To be specific, the implicit vector is used to generate a weight vector of the original group, which has a dimension of 128 in the experiment. $\mathbf{F}_{grp}$ is the flattened $\mathbf{Out}_t^0$ with length =1444. T represents transposition operation. $\mathbf{V}$ and $\mathbf{W}$ are the feature space transformation matrices. Ψ_{nl} is a non-linear transformation function. Therefore we can define the coefficient χ_k of the k^{th} element $\mathbf{F}_{grp}^{(k)}$ as Equation (3).

$$\chi_k = \frac{\exp((\mathbf{W}^T \Psi_{nl}(\mathbf{V}\mathbf{F}_{grp}^T))_k)}{\sum_{i=1}^{K} \exp((\mathbf{W}^T \Psi_{nl}(\mathbf{V}\mathbf{F}_{grp}^T))_i)}, k \in [1, K] \quad (3)$$

The weighted group feature $\tilde{\mathbf{F}}_{grp}$ is as Equation (4).

$$\tilde{\mathbf{F}}_{grp} = (\chi_1 \mathbf{F}_{grp}^{(1)}, \cdots, \chi_k \mathbf{F}_{grp}^{(k)}), k \in [1, K] \quad (4)$$

The weighted group-feature next passes through two fully connected layers with Dropout operation to reduce the feature dimensions and the computation and enhance the feature expressive ability. The output dimensions of the two fully connected layers are 512, and the $Relu$ activation function is used. The Dropout parameter is set to 0.5, that is, 50% of the neural units are discarded randomly and not participated in each iteration of training, which can reduce the risk of neural network overfitting.

Group-pooling is used to get the final group score. The trainable weight matrix followed by the sigmoid function is used to map the refined group feature to the abnormality score. The positive samples marked as 1 are the frame-groups containing anomalies, while negative samples contain only normal. As the trainable weight matrix of the refined group feature $\tilde{\mathbf{F}}_{grp}$ is marked $\mathbf{\Phi}$ and the sigmoid is marked σ the predicted group-score $\hat{\mathbf{B}}^{(t)}$ corresponding to the input $(\mathbf{In}_t^0, \cdots, \mathbf{In}_{t+\tau}^0)$ is defined as Equation (5).

$$\hat{\mathbf{B}}^{(t)} = \sigma(\tilde{\mathbf{F}}_{grp} \cdot \mathbf{\Phi}) \tag{5}$$

3.3 The Loss Function

The cross-entropy loss is the most commonly used loss in general logical classification problems, which considers the maximum likelihood estimation $\hat{\mathbf{B}}^{(i)}$ of the i^{th} true group-score $\mathbf{B}^{(i)}$ from the perspective of the information entropy. The two-class cross-entropy loss $\mathbf{Loss}_{bc}^{(i)}$ for a single sample is shown in Equation (6).

$$\mathbf{Loss}_{bc}^{(i)} = -\mathbf{B}^{(i)} \log \hat{\mathbf{B}}^{(i)} - (1 - \mathbf{B}^{(i)}) \log(1 - \hat{\mathbf{B}}^{(i)}) \tag{6}$$

The aim of the proposed algorithm is to predict high abnormality score when the video group contains at least one abnormal frame. Therefore we hope that the abnormality score of video group containing abnormal frames is higher than that of group with normal frames only, that is to say that the loss function should preferably increase the category gap so the network has a certain fault tolerance ability when dealing with a small amount of noise input. Thus we propose the improved hinge loss. The used two-class hinge loss function for a single-sample is shown in Equation (7).

$$\mathbf{Loss}_{bh}^{(i)} = \max(0, 1 - \mathbf{B}^{(i)} \cdot \hat{\mathbf{B}}^{(i)} + (1 - \mathbf{B}^{(i)}) \cdot \hat{\mathbf{B}}^{(i)}) \tag{7}$$

With N as the total number of samples, the final loss is:

$$Loss = \frac{1}{N}(\sum_{i=1}^{N} \mathbf{Loss}_{bc}^{(i)} + \sum_{i=1}^{N} \mathbf{Loss}_{bc}^{(i)}) \tag{8}$$

4 Experiments

4.1 The datasets

The proposed anomaly detection network uses the videos taken with the fixed-angle surveillant cameras, to detect anomalous events, including running, cycling, throwing objects, etc. The commonly used anomaly detection datasets are Avenue, UMN and UCSD.

The Avenue dataset Lu et al. (2013) contains the video segments captured on CUHK Campus Avenue, and the anomaly is when at least one runs, walks in the opposite direction, or hangs out. The UMN dataset Mehran et al. (2009) contains different indoor and outdoor scenes with crowd burst. In indoor scenes, the overall sight is dark due to insufficient light. The UCSD dataset Mahadevan et al. (2010) has the anomaly include cycling, carts, and turf crossings. The viewing angle of Ped1 is perpendicular to the camera plane, while the two planes are parallel in Ped2.

Above datasets only contain abnormal videos in testing set, while both positive and negative samples are required for the proposed network training. Thus we merge and filter these testing videos to choose 40 videos with single abnormal event in each one as the new training set, and the remains as the new testing set. It is worth mentioning that the anomalies, such as walking in the wrong direction and turf crossings, are treated as normal since they are not belong to the strenuous motion

in the crowd. The details of the new dataset are shown in Table 1. Compared with the original dataset, the new dataset encounters more challenges: (1) More scenarios. There are six different indoor and outdoor scenes, with different camera positions and angles. (2) More types of abnormal events. The new dataset contains single-person and group abnormal events. Subsequent experiments and comparisons were performed on this dataset.

Table 1: The distribution of the new dataset

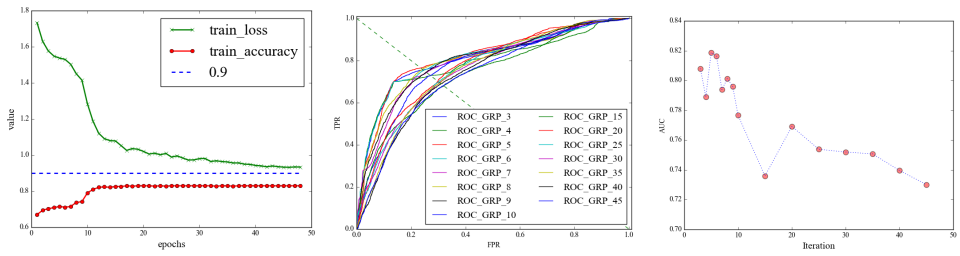
		Avenue	UMN	UCSD
Train Set	Videos	8	6	26
	Total Frames	6528	4360	5050
	Abnorm. Frames	727	721	2915
Test Set	Videos	8	5	20
	Total Frames	6087	3379	3560
	Abnorm. Frames	1103	658	2750

4.2 EXPERIMENTS

The training and testing codes are run under the centos7 system, using Intel i5-8600K @ 3.60GHz six-core CPU. The network is built using keras framework with tensorflow as the backend, supposed by python, opencv, h5py, etc.

4.2.1 THE MODELS IN TRAINING STAGE

The proposed algorithm uses sequential $\tau = 8$ frames to generate a single video group, and the sliding stride is 1. The output group score is taken as the abnormal score of the first frame of the group. Due to the discontinuity of different video files, the last $\tau - 1$ frames of each video file were discarded when the training set was generated, so the resulting training set contained $15,618$ video frame-groups. The Adam optimizer was used during training, and the learning rate was set to 0.0001; the batch-size was set to 32; each epoch contained $15,618$ sets of inputs during training. Fig. 4 shows the performance of models in training stage.



(a) Loss and Accuracy (b) ROC (c) AUC

Figure 4: The models in training stage

Fig. 4 (a) shows changes of parameters during training stage. The train_loss represents the value of the compound loss function introduced in Section 3.3. The train_accuracy is the ratio of the number of correctly classified samples to the total number of samples. We end the training stage when the train_loss stays basically unchanged in 5 iterations. The accuracy rate basically does not change after the 12^{th} epoch, so the best model occur after the 12^{th} epoch.

The ROC (Receiver operating characteristic) curve and its corresponding area (AUC) Mahadevan et al. (2010) Saligrama & Chen (2012) are the two commonly used metrics in classification algorithms. We select models every 5 epochs for performance testing and then narrow the scope according to the test results to find the model that performs best on the testing set. Thus we draw the Fig. 4 (b) and (c). TPR indicates the probability of correct prediction in all positive samples, and FPR indicates the probability of incorrect prediction in negative samples. Therefore the EER can be defined as the average of the incorrect prediction in both positive and negative samples. The ideal

target is TPR = 1 and FPR = 0, so the closer the ROC curve is to the (0, 1) point, the better. Table 2 recorded the AUC and EER of the models to make the metrics more specific and easier to compare the performance.

Table 2: The AUC and EER of different models

Iteration of the model	3	4	5	6	7	8	9	10
AUC	0.808	0.789	0.819	0.816	0.794	0.801	0.796	0.777
EER	0.244	0.271	0.235	0.246	0.257	0.242	0.255	0.270
Iteration of the model	15	20	25	30	35	40	45	
AUC	0.736	0.769	0.754	0.752	0.751	0.740	0.730	
EER	0.321	0.298	0.301	0.305	0.307	0.310	0.320	

Though the models before 10^{th} epoch have higher AUC values, they are in a large fluctuation state from Fig. 4 (c). The extreme point after 12^{th} epoch is at 20^{th} epoch. Thus the model at 20^{th} epoch is used as the representative model of the proposed algorithm in this paper to compare with other algorithms.

4.2.2 THE COMPARISON WITH OTHER ALGORITHMS

(1) Metrics in AUC and EER

We test the AUC and EER of the spatio-temporal autoencoder (ENC), variational encoder (VAE), multiple instance ranking framework (MIR) and the proposed algorithm (GRP) using the fused dataset introduced in Section 4.1, see Table 3. The GRP has the best performance with the highest AUC and the lowest EER, followed by the ENC. While the VAE based on implicit distribution prediction performs the worst due to the introduction of too much noise.

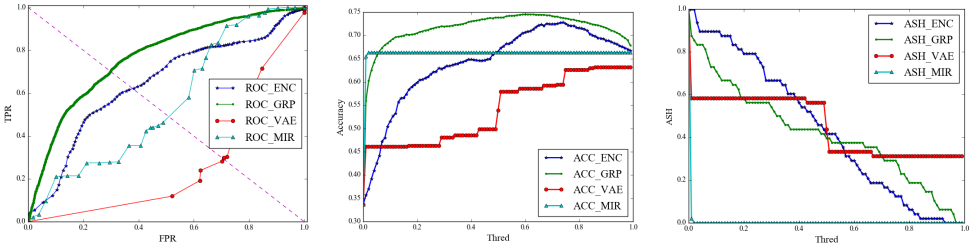
Table 3: The AUC and EER of different models

algorithms	AUC	EER
ENC Chong & Tay (2017)	0.645	0.380
VAE An & Cho (2015)	0.269	0.706
MIR Sultani et al. (2018)	0.555	0.513
GRP w.o/Atten(proposed)	0.754	0.292
GRP w/Atten(proposed)	0.769	0.298

In order to make the comparison more intuitive, we draw the ROC curves of each algorithm, as shown in Fig. 5. VAE generates fewer and more concentrated abnormal scores, the candidate values of the category segmentation threshold are relatively less and the data has poor separability, resulting in a higher binary classification error rate and a concave state of the ROC curve. In contrast, the GRP considers more scene information and reflects it on the abnormal score, so the obtained abnormal score set performs better under the established rules.

(2)Metrics in frame-level and segment-level

The AUC is used in the situation where the testing set includes both positive and negative samples, which may not be satisfied in real scenes. Thus we refer to the metric $accuracy$, which indicates the ability of the algorithm to classify the samples. The $accuracy$ is the proportion of correct detection in all the samples. Due to the similarity of the appearance under different actions, it is difficult to judge the border of the abnormal event, which will greatly affect the frame-level metrics. Therefore, the segment-level metric is supplemented, labeled as abnormal segment hit rate (ASH). For a segment containing abnormal events named as abnormal segment, a frame-level detection rate of not less than 60% is judged as detecting the abnormal segment. Define the abnormal segment according to the following two principles. (1) Two anomalous frames are counted as two abnormal segments when the distance between two anomalous frames exceeds 20 frames. (2) The division of an abnormal segment is not based on semantics, and consecutive adjacent anomalous frames belong to the same abnormal segment, even if two types of abnormal events occur in the two frames.



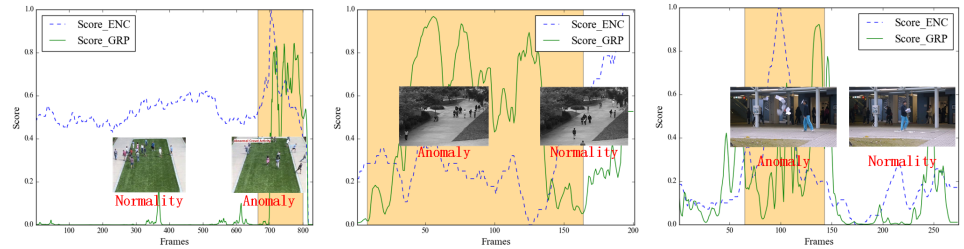
(a) ROC (b) Accuracy (c) ASH

Figure 5: The ROC curve frame-level and segment-level metrics of different algorithms

Fig. 5 shows the metrics of different algorithms on the testing set using the threshold selected every 0.01 between 0 and 1. $accuracy$ is the proportion of samples that are detected correctly. The maximum $accuracy$ of the GRP is the best of all algorithms, followed by the ENC. While when the threshold is low, the ASH of ENC is indeed higher than GRP, but it should be noted that each network model has a different focus on abnormality scores, that is the threshold to classify a score as anomaly is different. It can be found that the optimal thresholds of the two networks are basically in the high threshold area, where the ASH of GRP is greater than that of ENC, which is the reason why GRP is superior in other metrics. It is worth noting that the ASH of VAE is ladder-shaped and is relatively less affected by the threshold, which means that the score set of VAE is relatively sparse.

(3)Metrics in the detection effect

From the above metrics, the proposed algorithm and the spatio-temporal autoencoder are the two algorithms with the best performance in the comparison algorithms. Therefore, only the effects of these two algorithms are listed in the following video detection effect examples. Fig. 6 shows the specific score curve of the video. Fig. 6(a) is from the UMN dataset with a sudden burst of the crowds. The scores of the GRP are lower than that of the ENC in normal segments and higher in abnormal segments. Fig. 6(b) is from the UCSD dataset with a bumper car running. Although the GRP generates a high abnormal score at the end of the normal segment, it can detect most of the abnormal frames in the abnormal segment. So the overall effect is better than that of the ENC. Fig. 6(c) is from the Avenue dataset with a parabolic at high altitude. The GRP and the ENC both show a peak when the abnormal event occurs, but the peak time of GRP is slightly delayed. As a whole, the trend of the score curve of GRP and ENC is roughly the same.



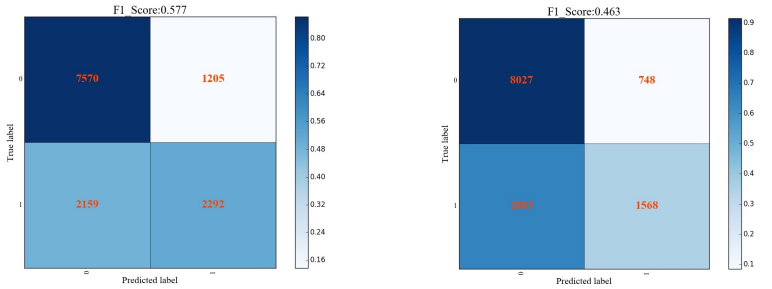
(a) UMN (b) UCSD (c) Avenue

Figure 6: The detection result of different algorithms

The threshold corresponding to the maximum point of the $accuracy$ in Fig. 5 is used as the hyper-parameter of the algorithm to determine whether an abnormal event occurs in each frame. The threshold of spatio-temporal autoencoder is 0.68, and its corresponding $accuracy$ is 0.725; the threshold of our algorithm is 0.6, and its corresponding $accuracy$ is 0.746. We use F_β score to evaluate the performance of the algorithms under the specified threshold. The definition of F_β score is shown in Equation (9).

$$F_\beta = (1+\beta^2) \cdot \frac{pre \cdot rec}{\beta^2 \cdot pre + rec},\ \beta \in N \tag{9}$$

Where pre represents the proportion of correctly detected samples in all the detected positive samples, and rec represents the proportion of correctly detected positive samples in all positive samples. F_β score is the harmonic value of pre and rec with the weight of β. When $\beta = 1$, namely F_1 score, it means that precision and recall are equally important. Fig. 7 calculates the confusion matrix and F_1 score of the two algorithms at their respective thresholds.



(a) The proposed algorithm (b) The spatio-temporal autoencoder

Figure 7: The confusion matrix of different algorithms

The numbers in Fig. 7 indicate the number of frames in the corresponding case. It can be seen from Fig. 7 that the GRP detects more abnormal frames with a higher number of both correct and wrong detection than that of the ENC. Hence, the calculation of F_1 score is necessary. The F_1 score of the GRP is 0.577, while that of the ENC is 0.463. Therefore, when $precision$ and $recall$ are considered together with a same weight, the performance of the GRP is slightly better than that of the ENC.

5 Conclusion

This paper proposes an end-to-end anomaly detection network, which can be used to detect strenuous movements in slow moving crowds, such as running, bicycling, throwing from a height. We arrange the multiple images of continuous video frames in a sequence called a frame group as the input of the proposed network to capture the spatio-temporal information. Then the attention mechanism is used to weight the extracted features to highlight the important information and weaken the interference of irrelevant information on the detection results, and fully connected layers to transform space and reduce the dimensions of the features. Finally, group-level pooling is used to map video group-level features to anomalous scores ranging from 0 to 1. The higher the score is, the more probability the frame has abnormal behaviors. More results and source code reported in this paper can be downloaded from website https://github.com/xiaojs18/Anomaly-Detection/tree/main/fgan.

References

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. Technical Report 2, SNU Data Mining Center, 2015. Special Lecture on IE.

RC Cai, WH Xie, ZF Hao, et al. Abnormal crowd detection based on multi-scale recurrent neural network. *Journal of Software*, 26(11):2884–2896, 2015.

Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pp. 189–196. Springer, 2017.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, Beijing, China, 2005. IEEE.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–742, 2016.

Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3619–3627, 2017.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136, 2018. arXiv:1802.04712.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015. arXiv:1502.03167.

Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851, Long Beach, CA, USA, 2019. IEEE.

Yuan Jing and Zhang Yujin. Application of sparse denoising auto encoder network with gradient difference information for abnormal action detection. *Acta Automatica Sinica*, 43(4):604–610, 2017.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the 2013 IEEE international conference on computer vision*, pp. 2720–2727, Sydney, NSW, Australia, 2013. IEEE.

Hui-Lan Luo and Chan-juan Wang. An improved vlad coding method based on fusion feature in action recognition. *Tien Tzu Hsueh Pao/Acta Electronica Sinica*, 47(1):49–58, 2019.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, San Francisco, CA, USA, 2010. IEEE.

Jefferson Ryan Medel. Anomaly detection using predictive convolutional long short-term memory units. Master's thesis, Rochester Institute of Technology, 2016. accessed from http://scholarworks.rit.edu/theses/9319.

Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 935–942. IEEE, 2009.

Vikas Reddy, Conrad Sanderson, and Brian C Lovell. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 55–61, Colorado Springs, CO, USA, 2011. IEEE.

Mehrsan Javan Roshtkhari and Martin D Levine. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision and Image Understanding*, 117(10):1436–1452, 2013.

Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017.

Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2112–2119. IEEE, 2012.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pp. 802–810, 2015.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, Salt Lake City, Utah, USA, 2018. IEEE.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE international conference on computer vision*, pp. 4489–4497, Santiago, Chile, 2015. IEEE.

Jinsheng Xiao, Hong Tian, Yongqin Zhang, Yongqiang Zhou, and Junfeng Lei. Blind video denoising via texture-aware noise estimation. *Computer Vision and Image Understanding*, 169:1–13, April 2018.

Tan Xiao, Chao Zhang, Hongbin Zha, and Fangyun Wei. Anomaly detection via local coordinate factorization and spatio-temporal pyramid. In *Proceedings of the 12th Asian Conference on Computer Vision*, pp. 66–82, Singapore, Singapore, 2014. Springer.

Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *Proceedings of the British Machine Vision Conference*, pp. 8.1–8.12, September 2015.

Fei-Niu Yuan, Lin Zhang, JT Shi, Xue Xia, and Gang Li. Theories and applications of auto-encoder neural networks: A literature survey. *Chinese Journal of Computers*, 42(1):203–230, 2019.

Yongqin Zhang, Jinsheng Xiao, Jinye Peng, Yu Ding, Jiaying Liu, Zongming Guo, and Xiaopeng Zong. Kernel wiener filtering model with low-rank approximation for image denoising. *Information Sciences*, 462:402–416, 2018.

Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 47:358–368, 2016.

Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 449–458, Salt Lake City, Utah, USA, 2018. IEEE.