The Journal of Supercomputing (2022) 78:3940–3954
https://doi.org/10.1007/s11227-021-04007-9

A novel two-stream structure for video anomaly detection in smart city management

Yuxuan Zhao[1] · Ka Lok Man[2,3,4,5,6] · Jeremy Smith[7] · Sheng-Uei Guan[1]

Accepted: 4 August 2021 / Published online: 17 August 2021

Abstract

Video anomaly detection is the problem of detecting unusual events in videos. The challenges of this task lie mainly in the following aspects: first, unusual events tend to make up only a very small portion of a video, which means a large amount of useless information needs to be culled. It further aggravates the test of algorithm performance and the computing ability of devices. Second, anomaly detection techniques are always used in the surveillance system, which contains massive video data. The analysis of such large video data is difficult. Last, the feature extraction ability of the algorithm appears a high performance since unusual video streams may lie close to normal video. Benefiting from the development of deep learning-based in computer vision fields, the accuracy and the efficiency of video anomaly detection has been improved a lot during recent years. In this paper, we present a newly developed two-stream deep learning model, which uses a 3D convolutional neural network (C3D) structure as the feature extraction part, to handle this task. Both the sequence of frames and the optical flow are required as the input of the model. Then, features of these two streams will be extracted from C3D and traditional convolutional neural network (CNN). Finally, a fusion layer will be used to fuse both results of streams and generate a final detection. Our experimental results on UCF-Crime video dataset outperform other benchmark results such as traditional deep CNN and long short-term memory (LSTM) in terms of area under curve (AUC). As the result, our proposed method achieves the AUC of 85.18%, which is 3% higher than the second highest method.

Keywords Anomaly detection · C3D · Deep learning · Computer vision

✉ Ka Lok Man
Ka.Man@xjtlu.edu.cn

Extended author information available on the last page of the article

1 Introduction

Video anomaly detection is the problem of detecting anomalies in videos. Anomaly refers to some unforeseeable events or emergency that deviates from what is standard, normal, or expected. Anomaly detection plays an important role in the smart city management, such as traffic control and criminal investigation. Unlike other anomaly detection tasks that can provide clear unusual signals [1], video anomaly detection requires the analytic of videos. Traditionally, we need professionals to monitor the video constantly to find out abnormal events. It always turns into a tough and time-consuming task. Therefore, research activities related to this task are of great practical significance since a feasible detection technique can reduce amount of human resources used for monitoring videos, especially for surveillance systems.

Anomaly detection in videos mainly faces the following challenges. Firstly, unusual events always happen with an extremely small probability. It makes relevant datasets difficult to be established. In addition, it causes the situation that the emphasis of related research activities can only be the features of normal videos. It affects the performance of classifiers in models, and makes the approach hard to provide correct detection result when the unusual video lies closely to normal video. Another factor exacerbates this phenomenon is that the differences between different anomalies may be huge, which makes it hard to extract general features from anomalies. Finally, video-based detection tasks are more complex than image-based tasks. Besides the spatial information, such as RGB data and grey-scale histogram, that both videos and images contain, methods used to handle videos should handle the temporal information as well. Particularly, anomaly detection techniques are always used to analyse massive video data in the surveillance system. To solve this important and challenging problem, methods are proposed and developed over the last decade. Traditional methods [2–7] focus on using clustering and classification approaches to judge if there is any abnormal event in videos. The kernel is to find anomaly from normal trajectories. Other methods focus on the deep learning-based models [8–13]. These methods always provide complex models, which are hard to explain and require powerful hardware settings. The models or results produced by traditional approaches have better interpretability. However, performances of them are not as good as deep-learning based methods. The past decade has seen a growth in computer hardware, which makes deep-learning-based models be the majority choices among research activities in recent years.

This paper proposes a model based on the two-stream structure to handle the anomaly detection problem. To adapt the traditional two-stream model to this specific task, plenty of improvements and changes have been introduced. For the structure of the model, the original two-stream model uses a spatial stream to extract spatial features in RGB frames of the video and uses a temporal stream to get temporal features from the optical flow of the video. Therefore, the temporal features of frames and the spatial features of the optical flow may be ignored by the model. To solve this drawback, we extract both spatial and temporal features from every stream. Considering that the combination of different methods may improve the

performance of the model for image and video processing [14], the structure combines outputs of long short-term memory (LSTM) [15, 16], CNN [17] and C3D [18] to replace the 2D convolutional models in the spatial stream. In addition, the structure of the DenseNet is in the stream of optical flow to enhance the connections among convolutional layers. Finally, the fusion layer is also improved to adapt the new model. For the video processing part, considering the massive video data, we cannot handle them frame by frame directly. It may lead to the problem of computing power. Our approach divides each video into clips and extracts C3D features for every clip to reduce the impact of large data input.

We carry out experiments on the UCF-Crime [19]. It is a large video dataset, which consists of long untrimmed surveillance videos which cover 13 real-world anomalies, including Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. It is a popular dataset that is used by plenty of research activities. We use AUC as the main evaluation standard.

2 Related work

Video anomaly detection approaches can be mainly divided into two directions. Traditional methods focus on the clustering-based detection [2–4, 20, 21] and low-level feature extraction [5–7, 22]. The principle of the clustering-based methods is the fact that an anomaly is always sudden and appears unusual features in a large range of videos [2]. Therefore, these methods can learn regular trajectories from normal events in a video stream. An unusual event will be detected if it cannot follow learned trajectories [20]. In order to improve the performance of the clustering, two models can be built to handle the spatial changes and movements in the video [4]. However, learning trajectories from normal events is complicated for traditional clustering methods. In addition, to solve the problem that clustering-based methods are too dependent on the moving objects, low-level feature extraction methods focus on low-level presentations in the video such as the change of grey scale, moving vectors [6] and textures [7].

Using reconstruction error is the most popular direction among these methods [8–11]. A model of normal videos is learned so that abnormal events will always show higher reconstruction errors than normal events during the testing part. Models in video anomaly detection task follow the basic structure of image-based model such as the convolutional neural network. To adapt this basic structure from the image to the video, methods for temporal features process are added such as LSTM, 3D convolutional neural network and two-stream model [23]. Besides the reconstruction error, future frame prediction chosen by some models, such as the reconstruction error, frame prediction techniques, also use autoencoders. They use autoencoders to generate anomalous frames [12, 13]. In addition, classifiers [5] and scoring methods [22, 24] are also used for video anomaly detection. The task can be considered as a binary classification problem, the classifiers are designed to produce accurate and robust features for both normal and unusual videos. Similarly, if we

consider this task as a regression problem, an anomaly score can be used to determine how likely the video is to be abnormal.

3 Proposed work

3.1 Overview

The proposed method is based on the two-stream structure for human behaviour classification [25]. Figure 1 shows the general structure of this new model. It follows the basic structure and inputs of the two-stream CNN model. For the stream of RGB frames, our model uses C3D to get both spatial and temporal features from clips. Then, an LSTM is used to get the video-based features. For the stream of optical flow, it uses a DenseNet structure to enhance the relationship between different CNN layers. Finally, there is a fusion layer to fuse the outputs of both these two streams and get the final detection result. In the proposed model, the input video should be turned into RGB frames and optical flow first. Then, the frames will be processed by C3D and LSTM in the first stream. The optical flow will be processed by the CNN in dense structure. Both results from these two steams will be fused by the fusion layer. It will also output the final detection result. Several new strategies are proposed to improve its performance of detection and adapt the massive input data. The specific modifications include the following aspects.

- Changing the processing part of the video
- Modifying the feature extraction from 2D CNN to C3D
- Simplifying the fusion layer to improve the detection speed

3.2 Video processing

Since the dataset of video anomaly detection (UCF-Crime) is much larger than the human action video dataset (UCF-101), video processing part becomes more

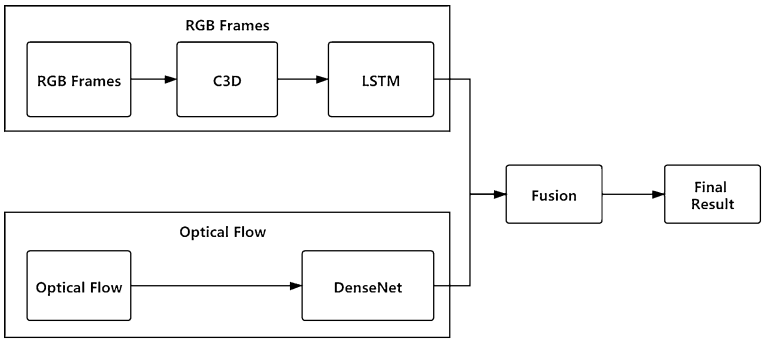

Fig. 1 General structure for the proposed method. It shows the two streams, which are designed for the RGB frames and the optical flow, and the fusion part

difficult. In addition, the usage of C3D increases the challenge of the task. Compared with traditional 2D CNN, the features of C3D are more complex. In addition, longer sequences need multiple LSTM layers or more parameters to extract the temporal information. As a consequence, more layers should be added or more parameters should be used in a single LSTM layer. All these factors will affect the time and difficulty of the training epoch. Finally, it may lead to the problem of computer power.

To solve this problem, the number of training samples should be reduced. Besides directly limiting the training samples by setting a specific parameter which would affect the detection performance, doing more preprocessing works for the input video should be a better solution. Therefore, reducing the input data becomes a serious problem. Some processing methods may cause the loss of video information. For example, picking one part of the video as the input, or setting a skip rate to pick one frame from every several frames. The decision of picking part needs the help of an attention mechanism, which may increase the complexity of the whole model. Selecting one frame from every several frames will definitely lead to the loss of the input features, especially for the optical flow. Finally, we decide to divide each video into clips. Similar clip-level feature extraction was used in some other research activities, such as Olympic events scoring [26]. In this video anomaly detection tasks, each video is divided into clips of 64 frames. Then, the model extracts C3D features for every clip. For the final clip, the model allows its number of frames less than 64. For example, if a video contains 816 frames, we can get 13 clips, while the final clip contains only 48 frames. Unlike the clip-level feature extraction before, we consider each clip a complete video and do the sampling for every frame in the clip. The general process is shown in Fig. 2.

For the stream of optical flow, we choose Horn–Schunck method [27] for the production of the optical flow in the format of vectors. Compared with Lucas–Kanade method [28] and Lucas–Kanade derivative of Gaussian (LKDoG) [29] method, the Horn–Schunck method can keep the balance between capturing enough motion vectors and reducing useless motion vectors. From Fig. 3, we can see that the Lucas–Kanade method is too sensitive that it records too many irrelevant and wrong features. Too many small motion vectors are recorded by this method. Usually, these vectors describe the vibrations of cameras, which may affect the further feature extractions. The LKDoG method has the opposite situation, it misses a lot of useful information of the subject in the frame. Figure 4 shows the part of the optical flow. Then, the sequence of optical images that contains the motion of x-channel and y-channel can be generated as the input of the model according to the optical flow vectors.

3.3 Feature extraction

The kernel challenge of video-based feature extraction is to get the spatiotemporal features. Two-dimensional ConvNets have been proven to be suitable for spatial features extraction. Furthermore, several approaches have been developed for temporal features. In the traditional two-stream model, it uses an optical flow as one of its

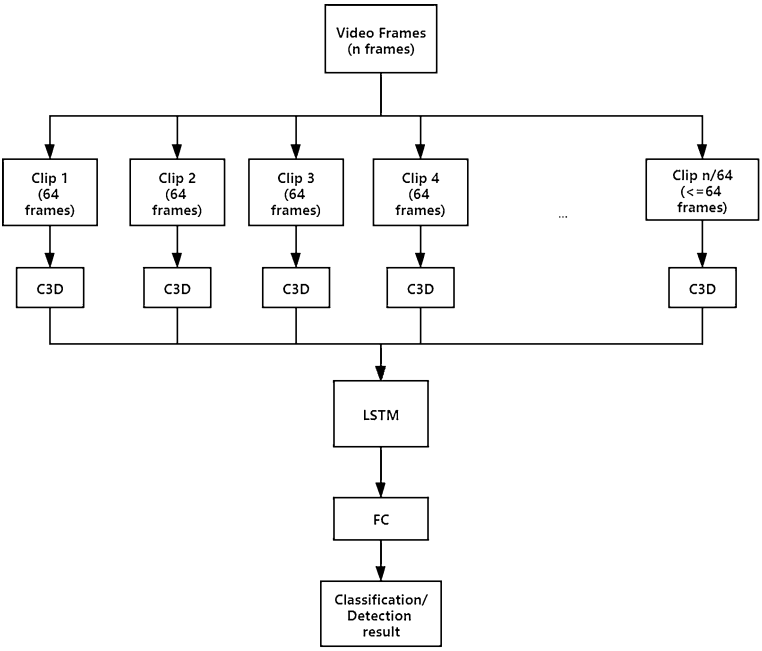

Fig. 2 Proposed method divides the input video into several clips. Except the final clip, every clip contains 64 frames. Each clip is processed by a C3D

inputs so that the temporal information in the optical flow can be captured. For the stream of RGB frames, an RNN-based structure, such as LSTM, can be used to process potential sequential information from CNN features.

To enhance the temporal features extraction, 3D ConvNet is used to the stream of RGB frames. According to the previous research [18], C3D can achieve a better performance than the traditional two-stream structure and LSTM. We do not use the whole network of C3D since it is outdated and not suitable for the video anomaly task. The 3D convolutional kernel is used for the spatiotemporal features extraction work. The basic structure of this part is also shown in Fig. 5. There are six convolutional layers that work for feature extraction. Four pooling layers are set after the first, second, fourth and last convolutional layers. Each convolutional kernel is set to be 3 x 3 x 3 to achieve the best performance. The whole structure is connected to the LSTM layer and fully connected layers for temporal feature aggression and produce the final output.

The LSTM built for this model is in a single directional structure to reduce the training time. It can filter the input information by three gates, which allow the layer to forget, enhance or output the features. This structure has been evaluated to be effective in video classification tasks [25]. It contains one LSTM layer and two fully connected layers. Each clip provides its clip-based features by C3D. Then, the LSTM further processes them into a video-based output. Therefore, it enhances the temporal features. For the streams of the optical flow, all parameters in the C3D layers, LSTM and the FC layer should be updated after every training epoch. Since the

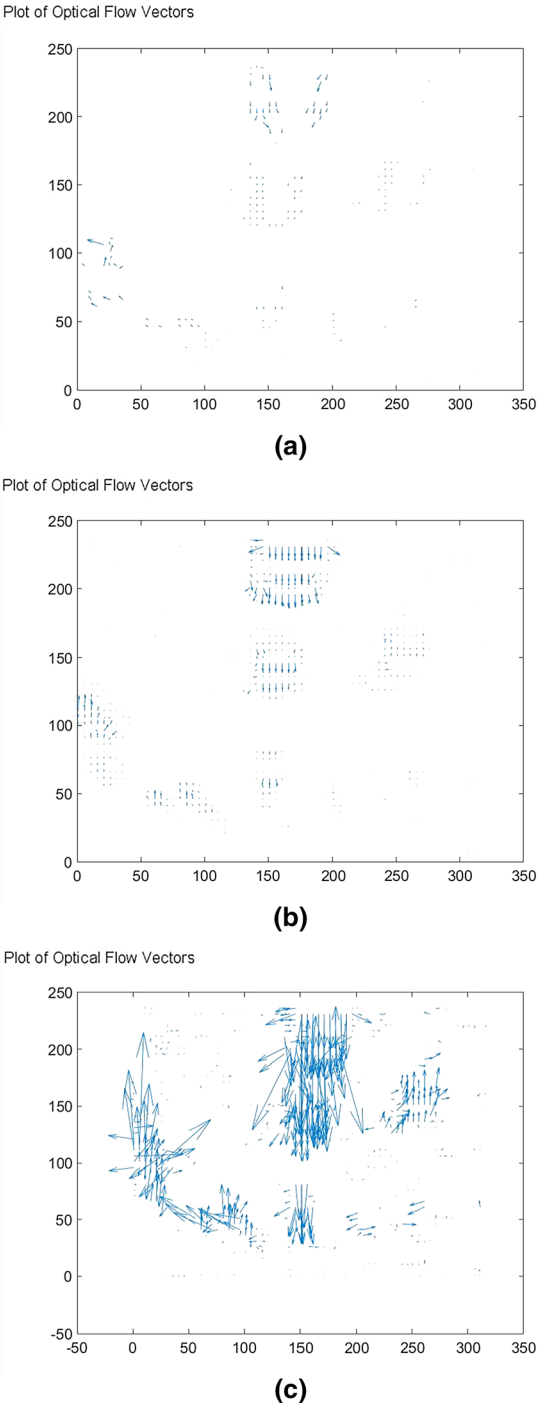

Fig. 3 Outputs of different approaches: **a** LKDoG method **b** Horn–Schunck method **c** Lukas–Kanade method

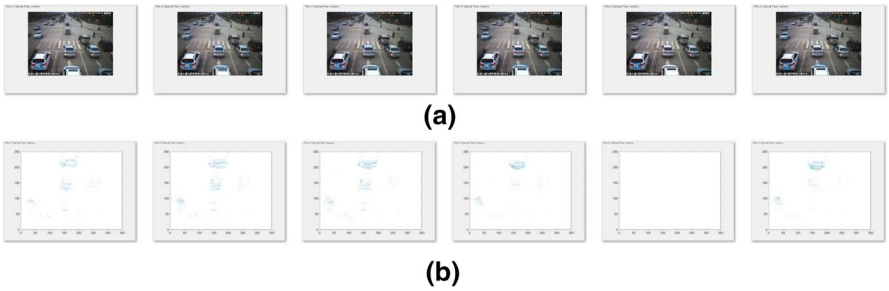

Fig. 4 **a** Original frames in the video **b** Corresponding optical flow

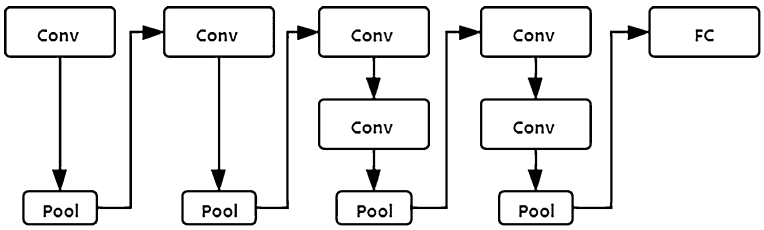

Fig. 5 Structrue of C3D in this method

original C3D is designed for behaviour classification, keeping its parameters may affect the performance.

3.4 Fusion layer

Since the streams of RGB frames and optical flows can provide their unique results, a fusion layer should be set after the two streams to get the final detection result. The traditional method uses an averaging method or a supported vector machine (SVM) [30], which uses the softmax scores as features to do the feature-based fusion. The detection task is not as complex as the classification task in this part, we simply use the averaging method in this model to reduce the training time. Both two streams will be processed by a softmax layer. Then, the fusion layer averages the scores from the softmax layers and gets the detection result according to the averaging score. In addition, the averaging method can reduce the training time since due to its simple structure.

4 Experiment

4.1 Dataset

We carry out experiments on the UCF-Crime dataset, which is a popular video dataset in the anomaly detection field. It consists of long untrimmed surveillance videos which cover 13 real-world anomalies. Table 1 presents some basic information for this dataset. From the table, we can see it is a large-scale video dataset that contains 1977 videos. UCF-crime is a benchmark dataset in the video anomaly detection field, and it could be easy to compare the proposed method to other anomaly detection methods (Fig. 6).

This dataset has some features that affect the efficacy of the training process and the performance of the algorithm. First, the sizes, lengths and frame rates of different videos vary a lot. Table 1 only shows the average data of these indexes. Second, some videos in the dataset are shot from a fixed view, while others contain clips from different views, which test the robustness of the algorithm. All these points make the detection method more challenging. For the evaluation part, we use the AUC as the standard.

4.2 Experiment environments and settings

This section shows both the hardware and software environment we use for the experiments. In addition, some details of the experiment settings are also described.

For the hardware, we use an Nvidia RTX Titan GPU for the experiments. The GPU is powered by the Turing GPU architecture. It has a 24 GB GDDR6 frame buffer, which can provide the ability for video related tasks. For the software, the PyTorch [31] is chosen to be the platform of the development. The whole training process is completed by PyCharm. Totally, 30 epochs are trained in this experiment, and we use Wandb to monitor the loss of every epoch.

4.3 Experiment results and discussion

In this section, the process of the experiment is analysed, and the AUC is compared with other methods.

The trend of the loss is shown in Fig. 7. The loss drops quickly from the first epoch. It is less than 0.1 on the 20th epoch. Then, the rate of decline gradually becomes lower and eventually fluctuates. Up to 30 epochs, the loss drops to 0.05, which is a very low value. After the 30th epoch, the value of the loss no longer shows a clear downward trend. Therefore, we choose the AUC in the 30th epoch as the final results of our experiment.

A receiver operating characteristic curve (ROC curve) [32] is a graphic plot that always be used to evaluate the performance of a model in classification tasks. It uses true positive rate (TPR) and false positive rate (FPR) as its parameters. The ROC curve of the proposed method is shown in Fig. 8. From the figure, we can see the

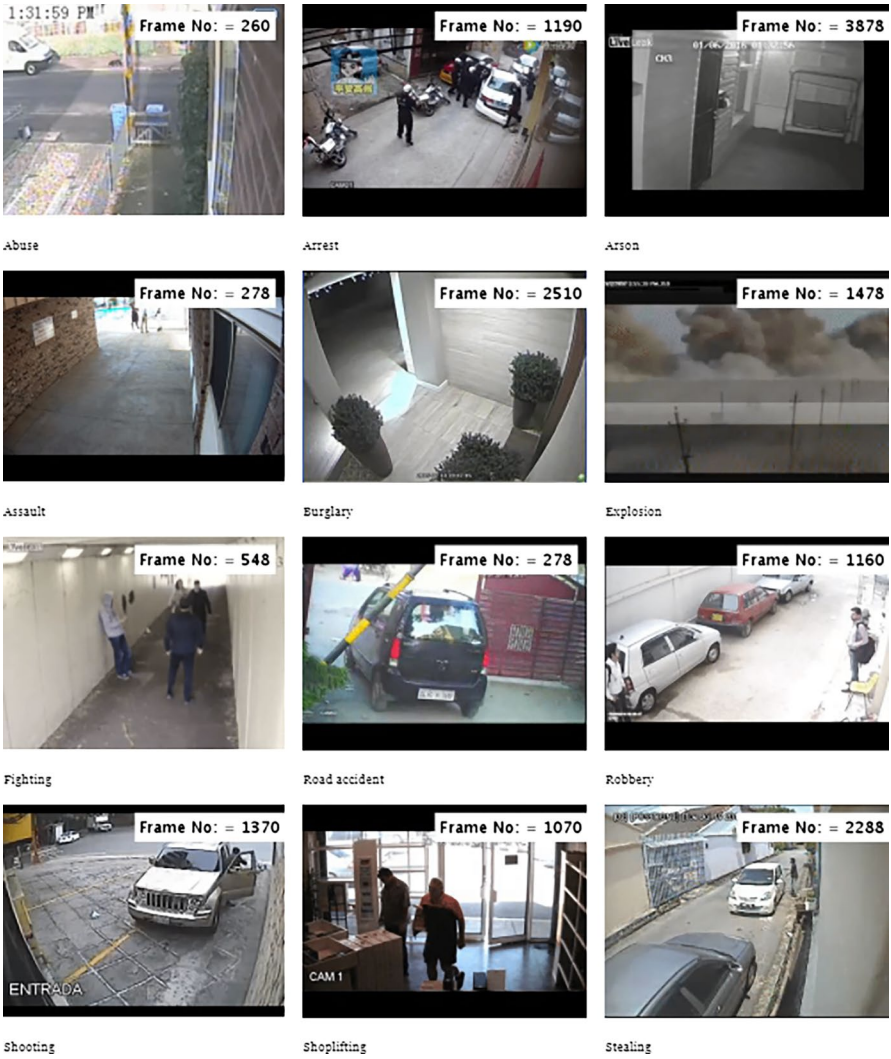

Fig. 6 Frames of videos in the UCF-Crime dataset

area under the curve is big, which means the performance of the model is good. In detail, according to the data in this figure, we finally get the AUC of our model for the experiments of UCF-Crime, which is 85.18%.

We compare our methods with other existing approaches for the UCF-crime. The result of AUC is shown in Table 2. Firstly, we choose some benchmarks and classic methods as an object for comparisons. They are C3D, combination of CNN (VGG19) [17] and LSTM, inception-V3 [33]. Our proposed method achieves an AUC of 85.18%. Compared with the traditional C3D structure, it gets a 4% improvement. In addition, it is 6% higher than the combination of VGG-19 and LSTM and 7% higher

Table 1 Basic information of UCF-Crime Dataset

Properties	Value
Number of videos	1977
Format	mp4
Average frames	7247
Frame rate	30 fps
Total length	128 hours
Labels	Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shooting, Stealing, Vandalism, Normal

Fig. 7 Trend of the training loss during 30 epochs. The horizontal axis presents the number of epochs. The vertical axis presents the loss of each epoch

than the inception-V3. Then, we compare our method with other complex methods. From Table 2, we can see that the AUC of the proposed method is 20% higher than the method of Lu et al [34]. Compared with the second highest AUC, which is achieved by Zhong et al [35], our result is still 3% higher than it. Other approaches in this table achieve AUC values of 63% [36] and 75.41% [19]. As a conclusion, our proposed method achieves the best performance among all these methods.

5 Conclusion

This paper proposes a novel model for the video anomaly detection. The model has two streams, which are used to receive RGB frames and optical flows of input videos. For the stream of frames, because the complexity of this stream is increased due

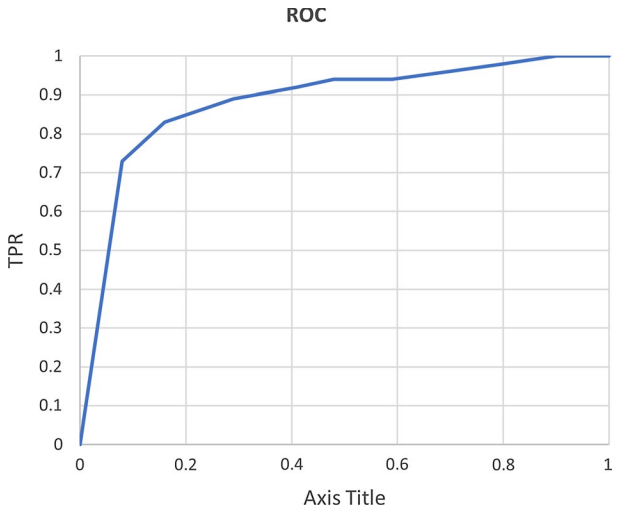

Fig. 8 ROC curve of the proposed method. The horizontal axis presents the false positive rate (FPR). The vertical axis presents the true positive rate (TPR)

Table 2 Experimental result

Methods	AUC (%)
Binary classifier	50
CNN (VGG19) + LSTM	80
C3D	81.08
Inception-V3 [33]	79
Lu et al [34]	65.51
Zhong et al [35]	82.12
Proposed Method	85.18

to the utilization of C3D and LSTM, we divide the frames into small clips. Then, the C3D can get clip-based spatiotemporal features of the video. Finally, an LSTM is used to enhance the temporal features and produce video-based features and detection results of the whole stream. Regarding the stream of the optical flow, since the input images have temporal information, the model uses traditional 2D CNN for the feature extraction part. In addition, the structure of DenseNet is used in this stream to enhance the relationship of different convolutional layers. Finally, to fuse the outputs of two streams and to avoid further increase the training time, we use the averaging method to get the final detection result. The model is evaluated by the UCF-crime video dataset and gets the highest AUC, which is 85.18%, among different methods. This model also has a good prospect in practical application. By the cooperation of different computer vision methods, it may be applied in different fields. For example, the proposed model can be the supplement of some existing traffic surveillance methods which use the vehicular cameras [37]. For human

behaviours, it can provide a basic event recognition result so that methods which use detailed behaviours as judgement factors can achieve a better performance [38] [39]. In the future, this new structure can be used in the monitoring system, which may improve the management of smart cities. The future work will focus on simplifying the model since the current model requires a large amount of computing power. The transformer structure [40] may be added into the method since it could reduce parameters for the model and would become popular in the video processing field from this year.

Acknowledgements This article is supported by Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China, with the Research Development Fund (RDF-15-01-01). Ka Lok Man wishes to thank the AI University Research Centre (AI-URC), Xi'an Jiaotong-Liverpool University, Suzhou, China, for supporting his related research contributions to this article through the XJTLU Key Programme Special Fund (KSF-E-65) and Suzhou-Leuven IoT & AI Cluster Fund.

References

1. Melvin AAR, Kathrine GJW, Ilango SS, Vimal S, Rho S, Xiong NN, Nam Y (2021) Dynamic malware attack dataset leveraging virtual machine monitor audit data for the detection of intrusions in cloud. Transactions on Emerging Telecommunications Technologies
2. Jiang F, Yuan J, Tsaftaris SA, Katsaggelos AK (2011) Anomalous video event detection using spatiotemporal context. Comput Vision Image Underst 115(3):323–333
3. Tung F, Zelek JS, Clausi DA (2011) Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. Image Vis Comput 29(4):230–240
4. Calderara S, Heinemann U, Prati A, Cucchiara R, Tishby N (2011) Detecting anomalies in people's trajectories using spectral graph analysis. Comput Vision Image Underst 115(8):1099–1111
5. Narasimhan MG, Kamath S (2018) Dynamic video anomaly detection and localization using sparse denoising autoencoders. Multimed Tools Appl 77(11):13173–13195
6. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell 30(3):555–560
7. Wang S, Zhu E, Yin J, Porikli F (2018) Video anomaly detection and localization by local motion based joint video representation and ocelm. Neurocomputing 277:161–175
8. Gong D, Liu L, Le V, Saha B, Mansour M. R, Venkatesh S, Hengel A. v. d (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1705–1714
9. Abati D, Porrello A, Calderara S, Cucchiara R (2019) Latent space autoregression for novelty detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 481–490
10. Chong Y. S, Tay Y. H (2017) Abnormal event detection in videos using spatiotemporal autoencoder. in International Symposium on Neural Networks, pp. 189–196, Springer
11. Zhou JT, Du J, Zhu H, Peng X, Liu Y, Goh RSM (2019) Anomalynet: an anomaly detection network for video surveillance. IEEE Trans Inf Forensics Secur 14(10):2537–2550
12. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection-a new baseline. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 6536–6545
13. Medel JR, Savakis A (2016) Anomaly detection in video using predictive convolutional long short-term memory networks," arXiv preprint arXiv:1612.00390
14. Jiang F, Chen Z, Nazir A, Shi W, Lim W, Liu S, Rho S (2021) Combining fields of experts (foe) and k-svd methods in pursuing natural image priors. Journal of Visual Communication and Image Representation 78:103142
15. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
16. Zhao Y, Zhang J, Man K. L (2020) Lstm-based model for unforeseeable event detection from video data," in CICET 2020. p. 41

17. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
18. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. in Proceedings of the IEEE International Conference on Computer Vision. pp 4489–4497
19. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 6479–6488
20. Suarez JJP, Naval Jr PC (2020) A survey on deep learning techniques for video anomaly detection. arXiv preprint arXiv:2009.14146
21. Maqsood M, Bukhari M, Ali Z, Gillani S, Mehmood I, Rho S, Jung Y (2021) A residual-learning-based multi-scale parallel-convolutions-assisted efficient cad system for liver tumor detection. Mathematics 9(10):1133
22. Pang G, Yan C, Shen C, A. v. d. Hengel, X. Bai, (2020) Self-trained deep ordinal regression for end-to-end video anomaly detection. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 12173–12182
23. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199
24. Del Giorno A, Bagnell JA, Hebert M (2016) A discriminative framework for anomaly detection in large videos. Springer. in European Conference on Computer Vision. pp. 334–349, Springer, 2016
25. Zhao Y, Man KL, Smith J, Siddique K, Guan S-U (2020) Improved two-stream model for human action recognition. EURASIP J Image Video Process 2020(1):1–9
26. Parmar P, Tran Morris B (2017) Learning to score olympic events. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp 20–28
27. Horn BK, Schunck BG (1981) Determining optical flow. Artif Intell 17(1–3):185–203
28. Lucas BD, Kanade T et al (1981) An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia
29. Baker S, Matthews I (2004) Lucas-kanade 20 years on: a unifying framework. Int J Comput Vision 56(3):221–255
30. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567
31. Ketkar N (2017) Introduction to pytorch. In Deep learning with python. pp. 195–208, Springer
32. Fawcett T (2006) An introduction to roc analysis. Pattern Recogn Lett 27(8):861–874
33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 2818–2826
34. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150. In Proceedings of the IEEE International Conference on Computer Vision. pp 2720–2727
35. Zhong J-X, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 1237–1246
36. Gianchandani U, Tirupattur P, Shah M (2019) Weakly-supervised spatiotemporal anomaly detection. University of Central Florida Center for Research in Computer Vision REU
37. Rathore MM, Paul A, Rho S, Khan M, Vimal S, Shah SA (2021) Smart traffic control: identifying driving-violations using fog devices with vehicular cameras in smart cities. Sustain Cities Soc 71:102986
38. Bukhari M, Bajwa KB, Gillani S, Maqsood M, Durrani MY, Mehmood I, Ugail H, Rho S (2020) An efficient gait recognition method for known and unknown covariate conditions. IEEE Access 9:6465–6477
39. Bilal M, Maqsood M, Yasmin S, Hasan NU, Rho S (2021) A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. J Supercomput pp 1–36
40. Dosovitskiy A, Beyer L, KolesnikovA, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Authors and Affiliations

Yuxuan Zhao[1] · Ka Lok Man[2,3,4,5,6] · Jeremy Smith[7] · Sheng-Uei Guan[1]

1 Department of Computing, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Ren'ai Road, Suzhou, China

2 Xi'an Jiaotong-Liverpool University, Suzhou, China

3 Swinburne University of Technology Sarawak, Kuching, Malaysia

4 imec-DistriNet, KU, Leuven, Belgium

5 Kazimieras Simonavicius University, Vilnius, Lithuania

6 Vytautas Magnus University, Kaunas, Lithuania

7 Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK