# A TWO-STREAM INFORMATION FUSION APPROACH TO ABNORMAL EVENT DETECTION IN VIDEO

*Yuxing Yang[1], Zeyu Fu[2] and Syed Mohsen Naqvi[1]*

[1] Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2] Department of Engineering Science, University of Oxford, UK

## ABSTRACT

Human abnormal activity detection for automatic surveillance systems is to detect abnormal objects and human behaviours in videos. In this paper, we propose to explicitly address different kinds of abnormal events by developing a two-stream fusion approach that integrates both geometry and image texture information. To be concrete, we firstly propose to utilize an object detector to divide the abnormal events into two catalogues: abnormal human behaviors and abnormal objects. For the detection of abnormal human behaviours, we exploit a spatial-temporal graph convolutional network (ST-GCN) which considers both spatial and temporal domains to capture the geometrical features from human pose graphs. The extracted geometric feature embeddings are further adapted with a clustering step to cluster the temporal graphs and output normality scores. For the detection of abnormal objects, the obtained from the object detector are reused to assist with generating normality scores of possible anomalies. Finally, a late fusion is performed to integrate normality scores from both screams for final decision. The experimental results on the datasets of UCSD PED2 and ShanghaiTech Campus demonstrate the effectiveness of our proposed approach and the improved performance compared to other state-of-the-art approaches.

***Index Terms*—** anomaly detection, object detection, pose tracking, graph convolutional neural network

## 1. INTRODUCTION

Abnormal event detection is an important yet challenging topic in computer vision and have applications in security, healthcare and entertainment industries [1, 2, 3]. The definition of abnormal events usually need to be considered with the scene context [4]. For example, a person walking on a sidewalk is defined as a normal event, while it becomes abnormal when walking on a lawn or a car lane. To address the problem of scene context dependency, most existing approaches [5, 6, 7, 8, 9] are based on generative models, formulating abnormal event detection in surveillance systems as a problem of outlier detection in a semi-supervised learning manner. Although these approaches can solve the lack of background information to a certain extent, they usually focus on either abnormal human behaviour or abnormal objects. For instance, if pedestrians walking on the sidewalk is considered as a normal event in the training data, one type of reconstruction model is required to find the abnormal behaviours such as cycling, chasing, running, jumping and falling in the testing set [6, 1]. And another type of reconstruction model used is to find the abnormal objects such as bicycle, skateboard, car and truck in the testing set [7, 8, 9]. However, if a monitoring video dataset has abnormal behaviours and objects at the same time, it may not be feasible for the aforementioned reconstruction models to capture the abnormal behaviours and objects together.

In this paper, we aim to explicitly address both kinds of abnormal event detection by introducing a two-stream fusion system as shown in Fig. 1. We firstly propose to classify the occurring events at the scene into two main categories: humans only and others via the pre-trained object classification model: YOLOv3 [11]. To detect abnormal human behaviors, we propose to analyze the human poses obtained by AlphaPose [12], which is motivated by skeleton-based human action recognition. And a spatial-temporal graph convolutional network (ST-GCN) [10] is exploited considering both spatial and temporal information to capture the features from human pose graphs. The nodes are the body landmarks and the edges are the connections of the body landmarks in spatial and temporal domains. After the graph convolutional network, a fully connected layer is appended to clustering the pedestrians' behaviours according to different normality scores. To detect abnormal objects, the class stream is developed to output normality score of other objects. Finally, a late fusion is performed to integrate normality score from both streams for final decision.

## 2. PROPOSED METHOD

### 2.1. Human-involved Anomaly Detection

#### *2.1.1. Pose Estimation*

Motivated by skeleton-based human action recognition [13] and human tracking[14], for our action stream system, we exploit human body joints information and use a ST-GCN

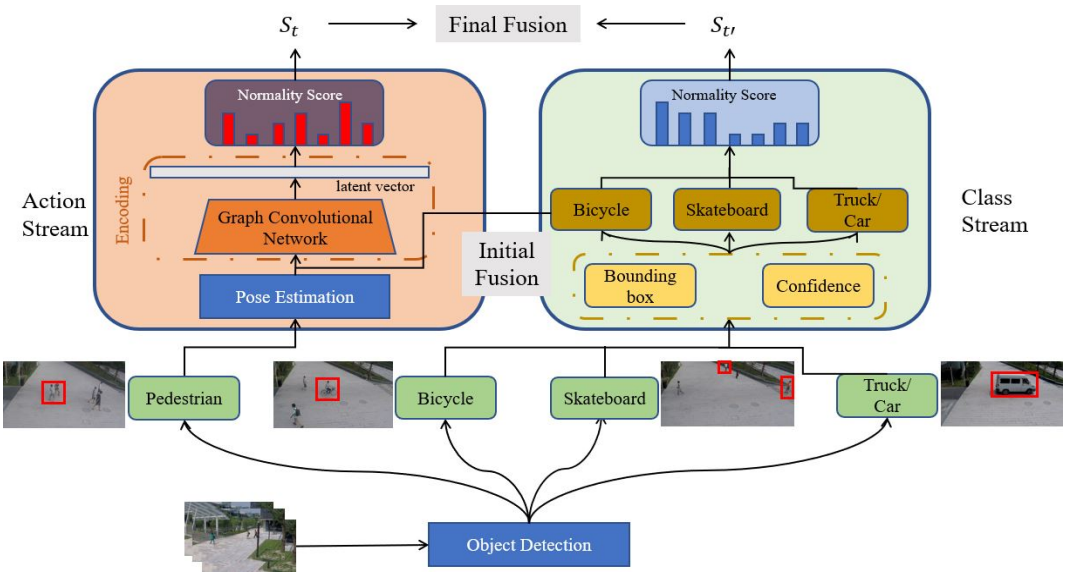


**Fig. 1**. The overall architecture of the proposed two-stream framework. The object detection does a primary classification to separate the detected objects into mainly four parts: pedestrian, bicycle, skateboard and car. For the action stream the frames with detected pedestrians are fed into the pose estimation step to obtain the corresponding body joints. The ST-GCN [10] is utilized to capture the spatial and temporal features from the normal body joints. The extracted features concatenated into a latent vector are further processed via a feature clustering step to output the normality scores. For the class stream, the information of detected bicycles and skateboards is initially fused into the pedestrians' normality scores according to their mixed bounding boxes. Finally, the results from both streams are fused at the decision level to output the final normality score. Color version is better to understand the figure.

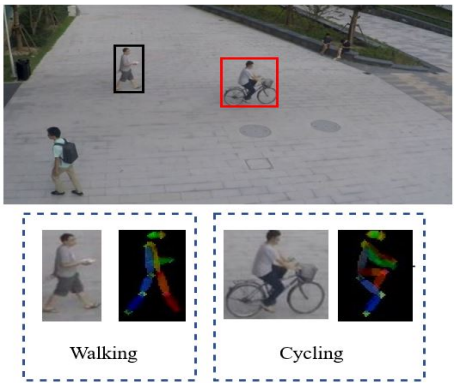


**Fig. 2**. The body joints for behaviours: walking and cycling. Color version is better to understand the figure.

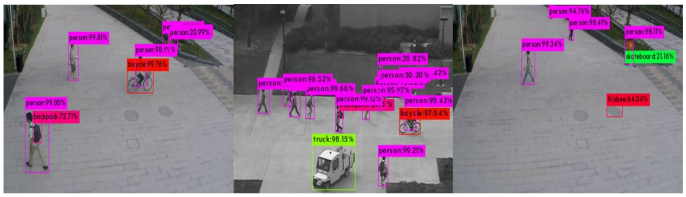

**Fig. 3**. Illustration of different types of abnormal objects: bicycle, truck and skateboard. Color version is better to understand the figure.

[10] to detect abnormal human behaviors, as it is found to be advantageous for analyzing the no-grid data (body joints). The pre-trained pose estimator is based on AlphaPose network [12] which can capture the 17 keypoints of the detected humans. In this stream, we aim to capture the latent features of different behaviours and achieve the relations between body keypoints. The results can decrease the influence of external factors related to the quality of videos such as background illumination, camera's viewpoints and pedestrians' appearances. All pedestrians are represented by pose graphs where the nodes mean the corresponding joints and the edges mean the relations between two nodes. The outputs of the detected video sequences are in 3-D lists which contain the identifications and localization of the detected pedestrians in each frame. Fig. 2 visualizes the body joints of two detected pedestrians with different behaviours: walking and cycling. With this multi-target pose estimator, the pose information extracted from a video sequence is shown as pose graphs in the time domain which contains the joints localization information in temporal sequences. The time domain adjacency can be defined by the vectors which are connected by neighbouring joints in continuous frames.

### *2.1.2. Spatio-Temporal Graph Convolutional Networks*

After capturing the body keypoints information with the pre-trained pose estimator on the dataset, a spatio-temporal graph convolutional network is applied to map the latent space of the normal data. For the spatial domain, it configures the physical connecting joints. The temporal domain manages the neighbouring connecting joints in consecutive frames. In the *i*-th layer, the graph contains $M$ joints and is represented as $\Gamma^i = \{N^i, E^i\}$ , where $N^i = \{\mu_j^i \in S^D | j = 1, ..., M\}$

is the set of joint nodes. $S$ means all joint nodes set and $D$ equals to 2 or 3 when the poses are in 2 dimensions or three dimensions. $E^i = \{\epsilon_j^i\}$ is the set of edges describing the connection joints. $D$ is the dimension of the joints. Meanwhile, two neighbouring nodes are represented as $C(\mu_j^i)$. Then the formula of the graph convolutional network is [6]:

$$p(\mu_j^i) = \sigma\left(\mu_j^i W_0^i + \sum_{\mu_k^i \in C(\mu_j^i)} \frac{1}{\gamma_{j,k}} \mu_k^i W_1^i\right) \quad (1)$$

where $\sigma$ is a nonlinear activation function. $W_0^i, W_1$ are two trainable weights, and $\gamma_{j,k}$ is a normalization variable.

There is also an adjacent matrix $A_{p,q}$ which means the weights of the connection between $\mu_p^i$ and $\mu_q^i$. We chose a fixed $A$ for all layers and the model is implemented with the following formula [6]:

$$GCN(N^i) = \Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}} N^i W^i \quad (2)$$

where $\Lambda$ is the degree matrix, $I$ is the identity matrix representing self-connections, and $N^i$ is the set of joint nodes at layer $i$. The equations above only consider spatial domain in GCNs. At time $t$, the $j$-th joint node $\mu_{j,t}^i$ at layer $i$ not only connects the neighbour joint nodes in the same frame, but also connects the same body joints at time $t'$. The captured latent vectors would make a final classification in the next clustering model.

#### *2.1.3. Feature clustering*

Given an input sample $S_i$, after extracting features $V_i$ by an encoder, a clustering layer is added to compute the features and assigns them to different clusters $y_i$. If the parameters in the clustering layer are $\Theta$ and there are $K$ clusters, the probability $p_i^k$ is given as [15]:

$$p_i^k = P(y_i = k) = \frac{exp(\Theta_k S_i)}{\sum_{k'=1}^{K} exp(\Theta_{k'} S_i)} \quad (3)$$

Since the proposed network is trained in a semi-supervised manner, we can get a normal probability distribution $P$ from the training stage. The optimization of the training is to minimize the Kullback–Leibler (KL) divergence between the clustering probability distribution $P$ and the detected objects distribution $Q$ [15],

$$Loss_c = KL(Q\|P) = \sum_a \sum_b q_{ab} log\left(\frac{q_{ab}}{p_{ab}}\right) \quad (4)$$

where $a, b$ are represented as the $a$-th sample assigned to $b$-th cluster. $p_{ab}$ is the probability of the prediction in the clustering model, and $q_{ab}$ is the target probability.

### 2.2. Object-based Anomaly Detection and Fusion

The proposed class stream is to directly capture the objects' classes, bounding boxes and confidence according to the pre-trained detector of pose estimation step in the action stream. The output matrix of detection is:

$$\mathbf{I}_j^k = [(F, x_1, x_2, y_1, y_2, S)_{1,j}, ..., (F, x_1, x_2, y_1, y_2, S)_{k,j}] \quad (5)$$

where $F$ represent confidence scores, $x_1, x_2, y_1, y_2$ represent the localization information and $S$ mean classes for the $k$-th object in the $j$-th frame, respectively. Besides, The results of different types of objects can strengthen the results of the action stream. If $F^k \in [bicycle, skateboard, car, truck]$ and $S^k > \tau$, the equation for the normality scores in frame $j$:

$$S_j = \frac{\sum_{k=1}^{K}(y_2^k - y_1^k)(x_2^k - x_1^k)S^k}{A} \quad (6)$$

where $A$ represents the area of the frames. The first abnormal class is bicycle which can be easily detected by the action stream. The second abnormal class is skateboard which is hard to be recognized due to its small size. The third abnormal classes are cars and trucks. Since these vehicles occlude the drivers, the action stream can not detect this type of abnormal event. If the normality score for frame $j$ in action stream is $P_j$, the final fusion of the action stream and class stream is:

$$D_j = P_j(1 + S_j) \quad (7)$$

the classification is as follows:

$$Frame(j) = \begin{cases} Normal & if\ D_j > \zeta \\ Abnormal & otherwise \end{cases} \quad (8)$$

where $\zeta$ is a variable to distinguish the abnormal events from video sequences.

## 3. EXPERIMENTS

### 3.1. Datasets and Evaluation Metrics

We consider the benchmark datasets on UCSD PED2 [16] and ShanghaiTech Campus [17] for our evaluations. They contain the abnormal events comprised of abnormal human behaviours and objects under the simple and complex scenes. For UCSD PED2 [16], there are 16 training and 12 testing video sequences with a fixed viewpoint. The main abnormal events on this dataset are abnormal objects, such as trucks, bicycles and skateboard. For ShanghaiTech Campus [17] , there are 13 scenes, 330 training video sequences and 107 testing video sequences with complex illuminations and different viewpoints. The main abnormal events on this dataset include abnormal human behaviors such as chasing, running, jumping and fighting; and abnormal objects include bicycles, skateboards, wheelchairs, cars and trucks.

In the action stream, there is a normality score for each frame. Thus, we can use the regularity scores to analyse

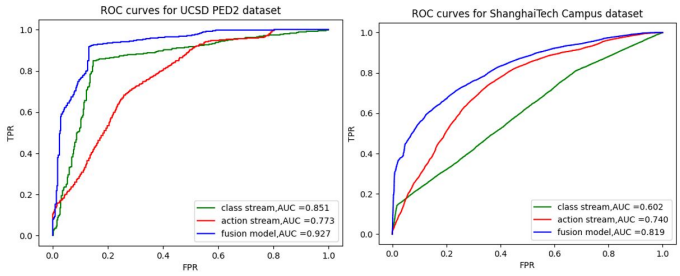


**Fig. 4**. The ROC curves for UCSD PED2 & ShanghaiTech Campus datasets.

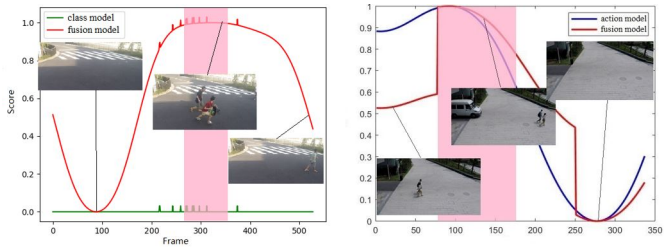


**Fig. 5**. Illustration of regularity scores for abnormal chasing and truck events.

the corresponding video sequences. The area under the curve (AUC) can also be used to evaluate the action stream. The AUC is calculated in Receiver Operating Characteristic (ROC) curve which has a false positive rate (FPR) at X-axis and a true positive rate (TPR) at Y-axis. Higher AUC means better performance.

### 3.2. Ablation Study

Table. 1 and Fig. 4 detail the performance of our three models on different individual abnormal events of UCSD PED2 dataset and ShanghaiTech Campus dataset. The class stream can detect the large-scale objects such as bicycles and cars. However, it can not be used to detect abnormal actions and is not sensitive to tiny objects such as skateboards. The action stream is robust to different abnormal events. The AUC performance for different abnormal events are similar. But the action stream must be based on human poses and can not capture the occluded abnormal events.

**Table 1**. AUC performance on different anomalies on ShanghaiTech Campus & UCSD PED2 datasets.

| | class stream | action stream | Fusion model |
|---|---|---|---|
| Strange actions | -/- | 0.720/- | 0.718/- |
| Bicycles | 0.820/0.861 | 0.776/0.810 | 0.886/0.930 |
| Skateboards | 0.763/0.792 | 0.729/0.826 | 0.788/0.928 |
| Cars | 0.843/0.980 | -/- | 0.867/0.999 |
| Total | 0.602/0.851 | 0.740/0.773 | **0.819/0.927** |

The fusion model can efficiently strengthen the classification of the action stream in abnormal objects such as bicycles and skateboards. Meanwhile, the fusion model can accurately capture the occluded abnormal vehicles such as trucks and cars. On UCSD PED2 dataset, the AUC performance of class stream is better than action stream which is contrary to ShanghaiTech Campus dataset. This is because the class stream is sensitive to high-resolution datasets. Fig. 5 quantitatively shows the regularity scores for abnormal chasing and truck event. It also describes the advantages of fusion model that it only care about the prior object information, but also focus on the moving pedestrians' behaviours. Compared with the results in class or action stream, the abnormal events frames are apparently separated from the video sequences in the fusion model.

### 3.3. Comparison with State-of-the-art

Table. 2 shows that the AUC of the proposed fusion model outperforms the state-of-the-art on ShanghaiTech Campus and UCSD PED2 datasets. For ShanghaiTech Campus dataset, we achieved the highest AUC performance. The normal solutions for abnormal event problem is generative model such as ConvAE [18] and ConvLSTM-AE [19]. These methods construct the normal event probability model, and detect the outlier event in testing. However, most of the time, the abnormal events are happened in small region which is hard to detect in frame level. The other solutions such as Markovitz1 *et al.* [15], it only focuses on the human-related abnormal events and can not handle the covered abnormal events which is not suitable for crowded environment. On UCSD PED2 dataset, the AUC performance of our proposed model is not the highest, which is the limitation of the low-quality dataset for the detection model. Motion feature can be extracted to help the class classification.

**Table 2**. AUC Performance for abnormal event detection compared with the state-of-the-art methods.

| | ShanghaiTech Campus | UCSD PED2 |
|---|---|---|
| Luo *et al.*[20] | 0.680 | 0.922 |
| Conv-AE [21] | 0.609 | 0.811 |
| MDT [18] | - | 0.829 |
| ConvLSTM-AE [19] | - | 0.881 |
| Abati *et al.*[22] | 0.725 | **0.954** |
| Markovitz1 *et al.* [15] | 0.761 | - |
| Proposed | **0.819** | 0.927 |

## 4. CONCLUSIONS

In this paper, we fused the information from the action-based model and the class-based model to improve the anomaly classification performance for video anomaly detection. We also analysed the accuracy performance for different types of anomalies in different models. The experimental results show that our proposed fusion framework has the best AUC performance compared with state-of-the-art on UCSD PED2 and ShanghaiTech Campus dataset. The proposed framework is suitable for video anomaly detection which contains abnormal behaviours and abnormal objects.

## 5. REFERENCES


[1] F. Angelini, J. Yan, and S. M. Naqvi, "Privacy-preserving online human behaviour anomaly detection based on body movements and objects positions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[2] J. Yan, F. Angelini, and S. M. Naqvi, "Image segmentation based privacy-preserving human action recognition for anomaly detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[3] V. Shashanka, K.-C.-Peng, S. R. Vikram, and M. Abhijit, "Attention guided anomaly localization in images," *European Conference on Computer Vision (ECCV)*, 2020.

[4] Z. Chen, Y. Tian, W. Zeng, and T. Huang, "Detecting abnormal behaviors in surveillance videos based on fuzzy clustering and multiple auto-encoders," *IEEE International Conference on Multimedia and Expo (ICME)*, 2015.

[5] Z. Fang, T. Zhou, Y. Xiao, Y. Li, and F. Yang, "Multi-encoder towards effective anomaly detection in videos," *IEEE Transactions on Multimedia*, doi: 10.1109/TMM.2020.3037538, 2020.

[6] W. Luo, W. Liu, and S. Gao, "Graph convolutional neural network for skeleton-based video abnormal behavior detection," *Generalization with Deep Learning*, pp. 139–155.

[7] M.Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," *IEEE International Conference on Image Processing (ICIP)*, 2017.

[8] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Y. Yang, Z. Fu, and S. M. Naqvi, "Enhanced adversarial learning based video anomaly detection with object confidence and position," *13th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2019.

[10] S.Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition.," *AAAI Conference on Artificial Intelligence*, 2018.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[12] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," *British Machine Vision Conference (BMVC)*, 2018.

[13] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2D pose-based real-time human action recognition with occlusion-handling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, 2020.

[14] Z. Fu, F. Angelini, J.A. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2277–2291, 2019.

[15] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30(5)*, pp. 909–926, 2008.

[17] W. Liu, D. Lian W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[19] W. Luo, W. Liu, , and S. Gao, "Remembering history with convolutional lstm for anomaly detection," *IEEE International Conference on Multimedia and Expo*, 2017.

[20] W. Luo, W. Liu, and S.Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[21] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[22] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.