Journal of Big Data

RESEARCH Open Access

Deep anomaly detection through visual attention in surveillance videos

Nasaruddin Nasaruddin[1,2], Kahlil Muchtar[1,2,3*], Afdhal Afdhal[1,2] and Alvin Prayuda Juniarta Dwiyantoro[3]

*Correspondence:
kahlil@unsyiah.ac.id
[1] Department of Electrical and Computer Engineering, Syiah Kuala University, Aceh, CO 23111, Indonesia
Full list of author information is available at the end of the article

Abstract

This paper describes a method for learning anomaly behavior in the video by finding an attention region from spatiotemporal information, in contrast to the full-frame learning. In our proposed method, a robust background subtraction (BG) for extracting motion, indicating the location of attention regions is employed. The resulting regions are finally fed into a three-dimensional Convolutional Neural Network (3D CNN). Specifically, by taking advantage of C3D (Convolution 3-dimensional), to completely exploit spatiotemporal relation, a deep convolution network is developed to distinguish normal and anomalous events. Our system is trained and tested against a large-scale UCF-Crime anomaly dataset for validating its effectiveness. This dataset contains 1900 long and untrimmed real-world surveillance videos and splits into 950 anomaly events and 950 normal events, respectively. In total, there are approximately ~ 13 million frames are learned during the training and testing phase. As shown in the experiments section, in terms of accuracy, the proposed visual attention model can obtain 99.25 accuracies. From the industrial application point of view, the extraction of this attention region can assist the security officer on focusing on the corresponding anomaly region, instead of a wider, full-framed inspection.

Keywords: Visual attention approach, Convolutional neural network (CNN), Integrated surveillance system, Anomaly classification

Introduction

While monitoring of public violence for safety and security is becoming increasingly important, surveillance systems are now being widely deployed in public infrastructure and locations. The identification of anomalous incidents such as traffic accidents, robberies or illegal activities is a vital role of video surveillance. Most existing monitoring systems, however, also need human operators and manual inspection (prone to disturbances and tiredness). Therefore, smart computer vision algorithms for automated video anomaly / violence detection are increasingly needed today. A small step towards resolving detection of anomalies is to build algorithms to detect a particular anomalous occurrence, such as violence detector [1], fight action detection [2, 3], and traffic accident detector [4, 5].

Video action recognition has gained increased attention in recent years when achieving very promising performance by taking advantage of CNN's incredible robustness.

Recently, Sultani et al. [6] have introduced a broad dataset and a multiple-instance learning (MIL)-based solution [7, 8] for this computer vision challenge in order to bridge the gap between surveillance camera storage and the restricted number of human monitors. Different from [6], Landi et al. [9] and Xu et al. [10] introduce a localized detection instead of considering full-frame video processing. In particular, Landi et al. [9] proposes exploiting the inherent location of anomalies and investigating whether the use of spatiotemporal information [11] can help detect anomalies. They combine the model with a module for tube extraction, which helps the analysis to concentrate on a specific set of spatiotemporal coordinates. A downside of this approach is that authors prefer to choose the manual / in-hand annotation rather than automatically driven localization by computer vision techniques. It leads to a time-consuming effort and ineffective. In contrast, Xu et al. [10] automatically locate all potential attention regions where fighting actions may occur, extracting several activation boxes from a motion activation map that measures the level of activity at each position. Then, the authors cluster all localized proposals around the extracted attention regions based on the spatial relationship between each pair of human proposals and activation boxes. It is important to note that Xu et al. [10] only focus on localizing the fight event in public area, thus this approach is not applicable for a unified anomaly detection system.

In fact, occlusions, illumination changes, motion blur and other environmental variations [12] are still challenging tasks in untrimmed public video footage. Therefore, in this paper, we propose an automatic yet efficient attention region localization approach through background subtraction. First, attention/moving regions are located using a robust background subtraction method. Once the attention regions obtained, it will be fed into a 3D CNN action recognition. It is noteworthy that our model only uses the obtained attention region from each frame during training. Like [6], we also address the detection of anomalies as a regression problem and propose a model consisting of a video encoder followed by a fully trainable regression network. In summary, this paper makes the following contributions.

A hybrid approach incorporating background subtraction and bilateral filter to localize attention regions for efficient anomaly detection is proposed.

A novel localization idea for a deep learning network to learn anomaly scores for video segments is introduced.

This paper is organized as follows: In section II, we present related works. In section III, we introduce our proposed method, including extracting attention regions from spatiotemporal information, and the detailed implementation of localized anomaly detection. In section IV, we test our method and summarize our results. Finally, in section V, we conclude the paper.

Related works

Anomaly detection is one of computer vision's most difficult and ongoing issues. With the increasing demand for public safety and surveillance, vast numbers of cameras have been installed in many public spaces, including airports, plazas, subway stations, and train stations. These cameras generate huge amounts of video data, resulting in an inefficient and exhausting process for a human operator to find suspicious or unusual occurrences. Moreover, there is an urgent need for an automated device to increase

productivity and save energy. As a result, significant efforts have been made towards smart video surveillance, and many approaches have been proposed to allow significant progress to be made in the detection of video anomalies.

Various approaches to detecting abnormal behavior have been developed in the past [13–33]. In [34], the video and audio data is used to identify violent behavior in video surveillance. Mohammadi et al. [1] proposed a new behavior-based heuristic approach to classifying violent and non-violent videos. Different from previous works, authors in [14, 15] suggested to use tracking as an anomaly to model normal motion. Due to difficulties in obtaining reliable tracks, a number of approaches avoid tracking and learn about global motion patterns using histogram-based methods [16], social force models [30], mixture of dynamic texture models [20], Hidden Markov Model (HMM) [21], topic modeling [18], motion patterns [35] and context-driven method [19]. One of work from Mehran et al. [30] was trying to measure the interaction force of the scene by measuring the difference between desired and real velocities obtained by particle advection, which uses the social-force model. These approaches learn how to distribute normal motion patterns and detect low probable patterns as anomalies given the training videos of normal behaviors.

Recently, approaches based on deep learning have been presented. Xu et al. [13] use a machine learning framework to learn video features rather than to use hand-crafted features. Multiple Single Class Support Vector Machine (SVM) models are then used from the learned features to score the anomaly level for each input. The multiple SVM results are then combined for the final detection of anomalies. Hasan et al. [24] proposed a convolutional auto-encoder (Conv-AE) framework for the reconstruction of the scenes, and then computed the reconstruction costs for the identification of anomalies. Zhou et al. [36] proposed spatio-temporal CNNs to learn joint appearance and motion characteristics. Sultani et al. [6] combined deep neural network with multiple instance learning to classify real-world anomalies, such as accident, explosion, fighting, abuse, arson, etc. Similar to [6], our approach considers not only normal behaviors, but anomalous behaviors for detection of anomalies. In addition, our work introduces a visual attention idea in order to localize the region of interest (ROI).

Methods

Our algorithm incorporates the BG subtraction with the bilateral filter. The bilateral filter is used to alleviate the noise from the untrimmed public incoming frames. BG subtraction is used to retrieve the foreground (candidate attention regions to register). Finally, the extracted attention region of various anomaly events will be predicted through a deep learning pipeline. Figure 1 illustrates the overview of our proposed work, the details of which are discussed in detail in the following parts. We divide the section into two primary parts; visual attention detection and event action detection.

Visual attention detection

We use a bilateral BG subtraction approach based on texturing which effectively eliminates noise and retains the edges of observed areas [37, 38]. In our case, such a technique is capable of building a stable BG model in order to clearly see the area of the moving objects as a visual attention region and the rest as an uninterested area.

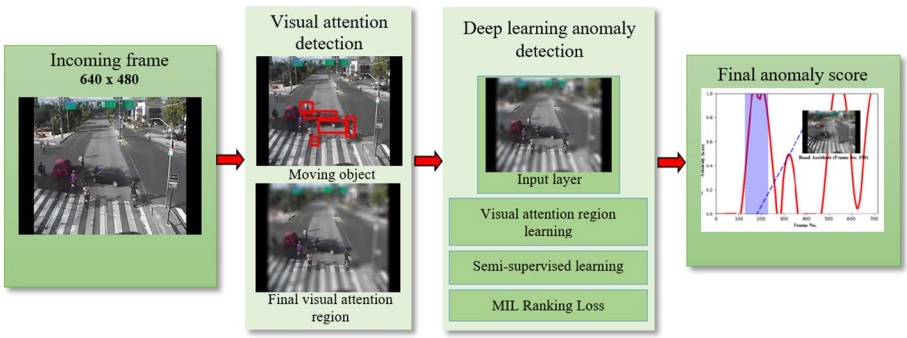


Fig. 1 System workflow of proposed work

As shown in Fig. 2, we visualize the comparison between the bilateral BG subtraction with two famous approaches, namely improved model of Mixture of Gaussians (MOG2) [39] and K-Nearest Neighbors (KNN) [40]. Both methods processed the image pixel by pixel and often regarded the noise as a candidate for moving pixels, such as shadow interference and intermittent motion. On the contrary, the bilateral texture-based approach was able to exclude the noise properly, and extract the correct region. To be specific, in Fig. 2, the noise/misclassified regions are highlighted through the circle in red color (mostly shadow is regarded as moving pixels on MOG2 and KNN approaches), while correct regions were drawn through the rectangle in green color (in the original image). Although the previous works are able to produce a more complete segmented foreground object, the misclassified region can affect the extraction of the region of interest (ROI). Therefore, in our proposed anomaly detection pipeline, the visual attention region can be obtained more accurately and efficiently through the bilateral BG subtraction method. The comparative evaluations (in Fig. 2) were conducted using the UCF-Crime [6] dataset that contains various classes of anomaly activity. This approach can achieve ~ 100 fps for 240×320 pixels input format in Graphical Processing Unit (GPU). Therefore, this approach is very efficient to localize the region before performing anomaly detection through deep-learning pipeline.

First, we use bilateral filtering [41] to an input frame I, and denoted the greyscale output image as $I_{\text{bilateral}}$. The $I_{\text{bilateral}}$ is used to generate a non-overlapping block-based texture. More specifically, the $I_{\text{bilateral}}$ is divided into blocks of sizes $n \times n$ pixels. The n setting is set to 4 in our system. Then, we calculate the mean of each block and construct a binary bitmap using it. The bitmap BM_{bil} is obtained by comparing the mean with each pixel value in a block. If the value of the pixel is below the mean, the binary value is 0, and vice versa. Finally, the BM_{bil} of each block is used to build the initial BG model BM_{mod}, and becoming a reference when the new incoming frame exists. Our current BG model update rule and its appropriate learning rate are similar to our previous method [42].

In theory, when a new frame arrives, we simply calculate a hamming distance for each block to decide if the observed block $BM_{\text{bil\_obs}}$ is regarded as BG block or attention region block. Note that, the b_{ij} indicates the corresponding bit value in i, j position of a block.

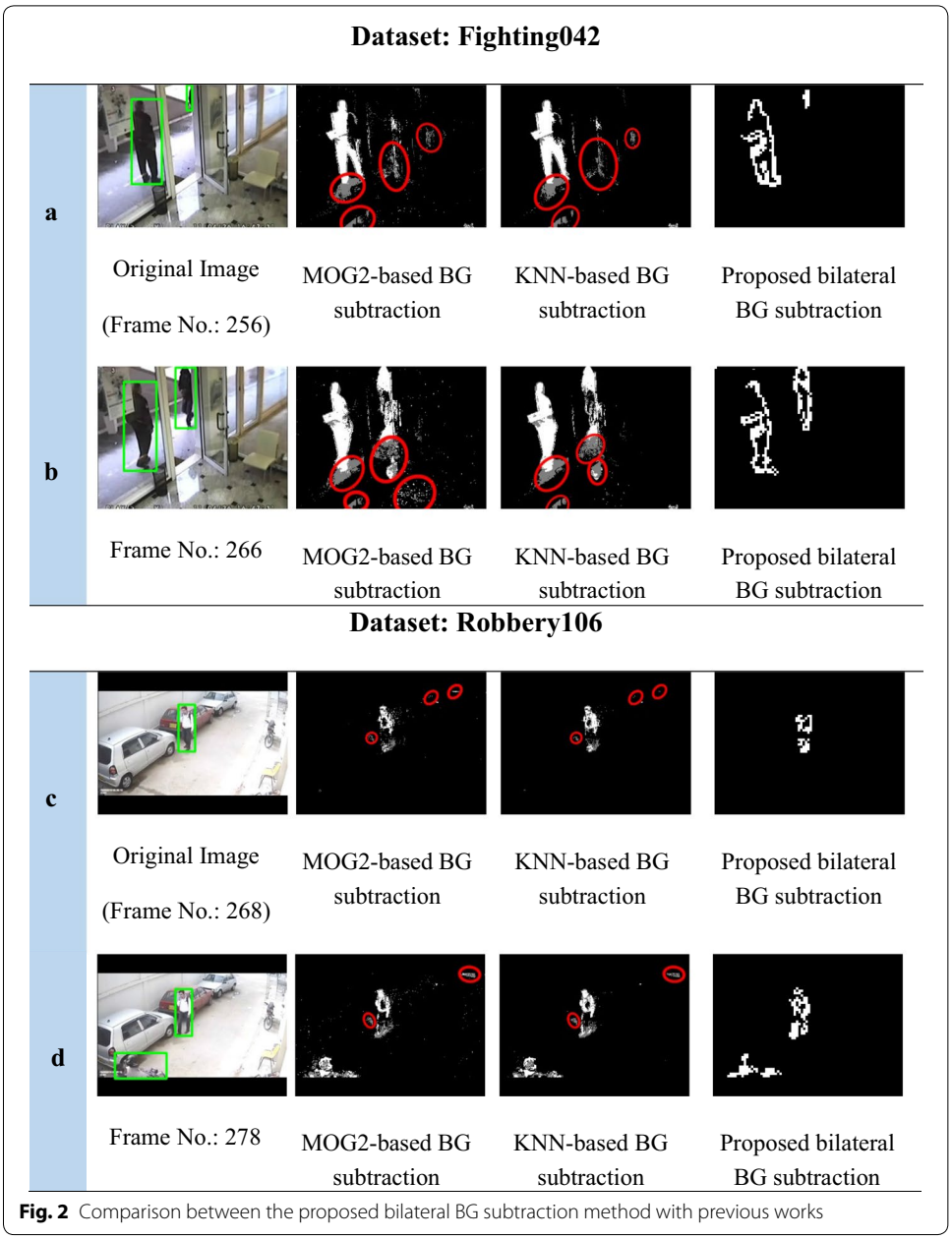


Fig. 2 Comparison between the proposed bilateral BG subtraction method with previous works

$$Dist\left(BM_{bil\_obs},\ BM_{mod}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(b_{ij}^{bil\_obs} \oplus b_{ij}^{mod}\right) \tag{1}$$

The bilateral filter is very slow compared to most filters while keeping the edges of the active area relatively sharp. Therefore, instead of using global memory, we use the texture memory of a CUDA to process an input frame and perform a bilateral GPU filter [41]. For clarity purposes, Fig. 3 describes the step-by-step generation of texture information.

In addition, Fig. 4 illustrates a simple example in order to generate the texture information on a single 4×4 pixel block. Figure 5 shows the image generated using

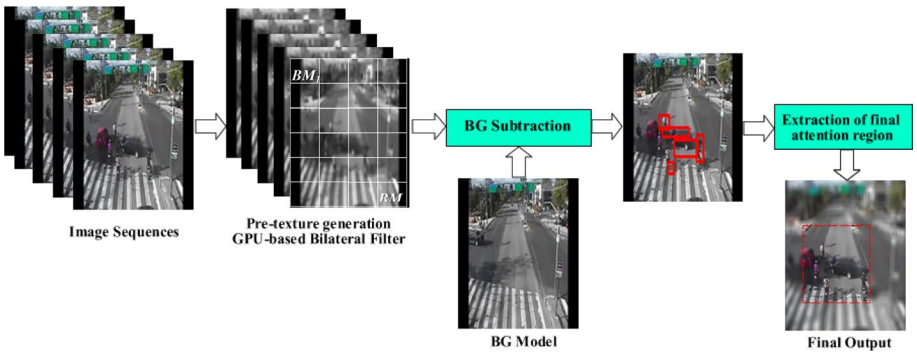


Fig. 3 Step-by-step of finding visual attention region

156	157	159	156
157	156	158	156
154	154	156	154
151	155	156	157

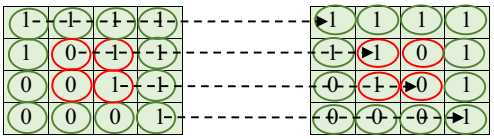
Step 1: Proceed every 4×4 block

156	156	156	157
156	155	156	157
154	154	156	157
153	154	155	157

Step 2: Calculate Bilateral filter for each 4×4 block and its block mean. Here the ***mean*** : **156**

1	1	1	1
1	0	1	1
0	0	1	1
0	0	0	1

Step 3: Determine Bilateral Bitmap $\boldsymbol{BM_{bil}}$. If $\boldsymbol{b_{i,j}} < \boldsymbol{mean}$, set to 0 Else, set to 1



Bilateral Bitmap $\boldsymbol{BM_{bil\_obs}}$ **Example of reference** BM_{mod}

Fig. 4 An example of calculating distance between incoming block $BM_{\text{bil\_obs}}$ and reference block model BM_{mod}

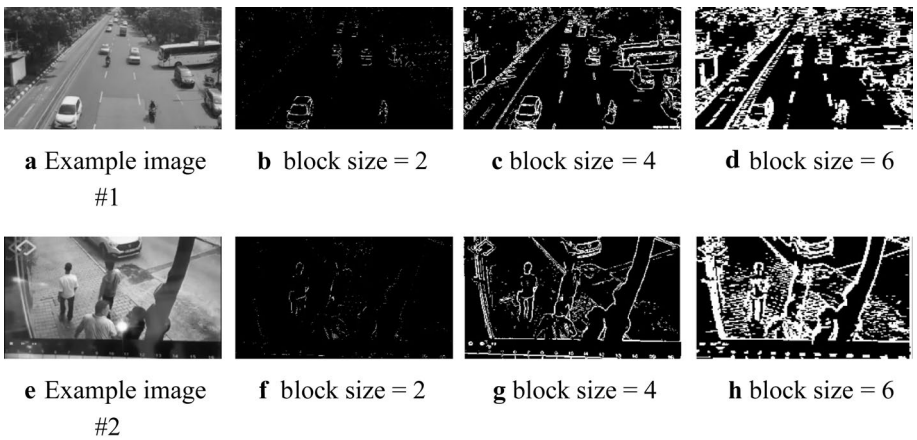


Fig. 5 The illustration of different block size used in the proposed texture descriptor

the proposed texture descriptor with different block sizes. Figure 5b, f remove too many details, whereas Fig. 5d, h shows excessive unimportant details. Figure 5c, g prove the validity of this texture descriptor and shows that block size = 4 is an excellent choice.

Event action detection and implementation details

Feature extraction through the pre-trained C3D model

The 3D CNN is commonly used for various computer vision applications, especially for classification, detection, and recognition task. Typically, the 3D CNN model consists of several layers, namely pooling, convolutional, and fully-connected (FC) layers. The preceding layer by means of kernels with a pre-defined, fixed-size receptive field is connected to every layer. The 3D CNN model learns the setup of hyper-parameters from a big data collection to represent the video clip's global or local characteristics. That model architecture has different layer types and activation functions to display better representational features than human-engineered software.

As illustrated in Fig. 6, for its good performance and efficiency, the popular Convolutional 3D Networks (C3D) [43] is selected as our pre-trained feature extractor. Recent studies [44–47] have shown that fine tuning of a more complex dataset results in excellent classification and detection performance using a pre-trained Sports-1 M dataset model [48]. The reason for this training procedure is that the 3D CNN receives general representation of video clips from pre-training. The model adjusts the parameter after the fine-tuning to show the specific features of the video segment, while retaining the ability to display the general video segment. This training strategy is implicitly implemented, coupled with a sampling of shuffles and cross-validation.

Implementation details of anomaly detection

Figure 7 shows the flow diagram of proposed visual attention-based anomaly detection. Specifically, we derive visual characteristics from the C3D network's fully connected (FC) layer FC6. We re-size each video frame to 240×320 pixels before computing features and set the frame rate to 30 fps. We compute C3D features for every 16-frame video clip followed by l_2 normalization. We take the average of all 16-frame clip features within that segment to get features for a video segment. These features (4096D) are input into a neural network of 3-layer FC. For detection

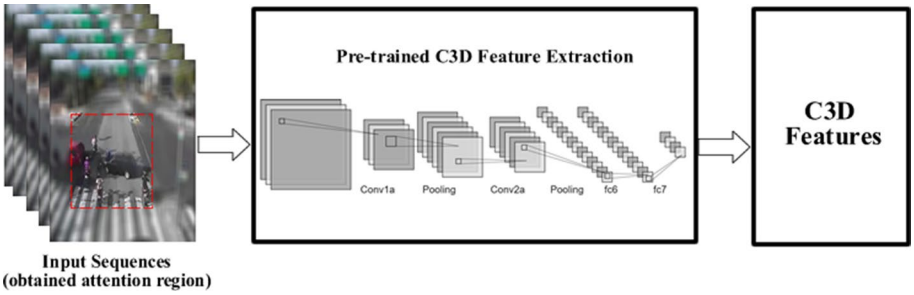


Fig. 6 Pre-trained C3D feature extraction of obtained attention regions

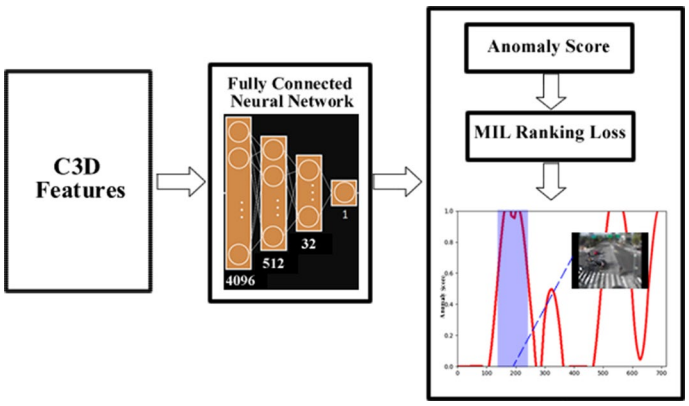


Fig. 7 Flow diagram of proposed visual attention-based anomaly detection

Table 1 Parameter used in the C3D feature extraction

Parameter	Value
Batch size	30
Dropout	0.5
Image resize (width × height)	128 × 171
# feature extractor layers	6

purposes, we inference every 160 frames (every 10 C3D extracted files) gradually in order to convince whether the anomaly scenes exist or not.

The regression network outputs the video anomaly score. Since the score ranges from 0 to 1, we can interpret it as the likelihood of an unusual event occurring in the localized segment being investigated. Inspired by [6], we utilize the MIL ranking loss as sparsity and smoothness constraints [8] and consider each video segment as an instance of the bag. Given an input video M, its anomaly score Sc(M) must comply with the following:

$$\begin{cases} 0 \leq Sc(M) < T, \ if \ M \ is \ unanomalous; \\ T \leq Sc(M) \leq 1, \quad if \ M \ is \ anomalous. \end{cases} \tag{2}$$

where the threshold T drives the binary classification into normal and anomalous videos. Ideally, anomalous segments score close to 1 while regular videos map values close to 0. In typical setting, the T is set to 0.5. We use the activation function of rectified linear activation unit (ReLU) and adopt a 50% dropout regularization [49] between the layers of FC.

It is important to note that our training sample consists of a 16-frame video segment M, where each frame already output a localized area from previous steps. We train the localized model using the public and a comprehensive dataset called UCF-Crime [6]. The number of training data are 800 un-anomalous videos and 810 anomalous videos, respectively. In general, there are 14 classes of event that provided by the authors [6] Table 1.

Table 2 Statistics about the localized UCF-crime dataset. Number in brackets represent the number of videos in the training set

Type	Value
Number of videos	Anomalous: 950 (810), Normal: 950 (800)
Average total number of frames	13 million
Dataset length	128 h
Anomalous classes tested	Robbery, fighting, road accident

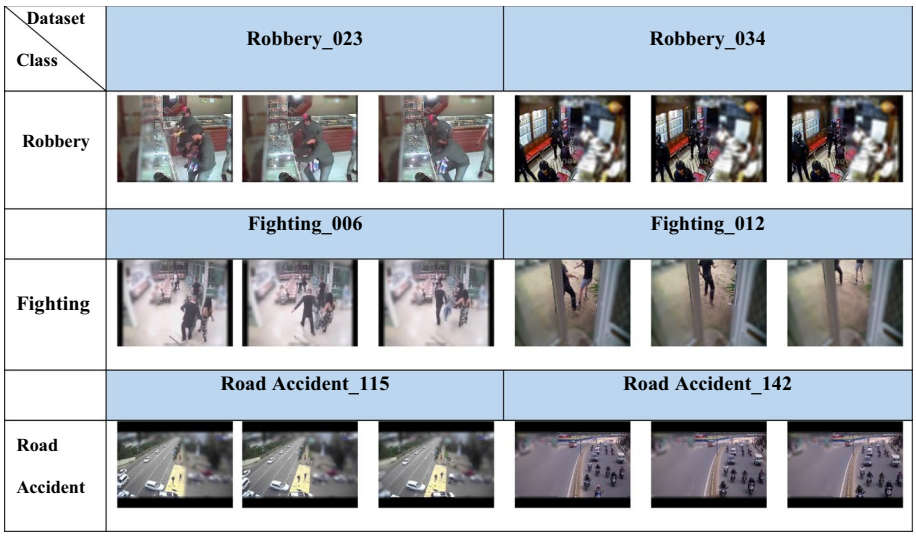


Fig. 8 Example of different anomalies in trained videos (ROI localized)

Results and discussion

During experiment, the PC is equipped with Intel i7-7700HQ processor, 16 GB of memory, and NVIDIA GeForce GTX 1050 Ti 4 GB. The PyTorch 1.2 is employed as a framework and pre-trained model of C3D is used for spatial–temporal feature extractions. In Table 2, we describe the statistics about the localized UCF-Crime dataset which is used throughout the training and testing stage.

In addition, Fig. 8 shows some examples of localized anomalous regions in various classes. The UCF Crime dataset [6] consists of surveillance videos which are data obtained from LiveLeak and YouTube. In this experiment, we evaluated three classes from the UCF Crime dataset to act as a baseline test set for evaluating the accuracy. The dataset provides the ground-truth label in binary classification for each tested video. Therefore, it is straightforward to perform the evaluation of detection thoroughly. In Fig. 9, we demonstrate an anomalous event example from the UCF-Crime dataset through qualitative comparison, namely robbery, fighting and road accident, respectiv ely.

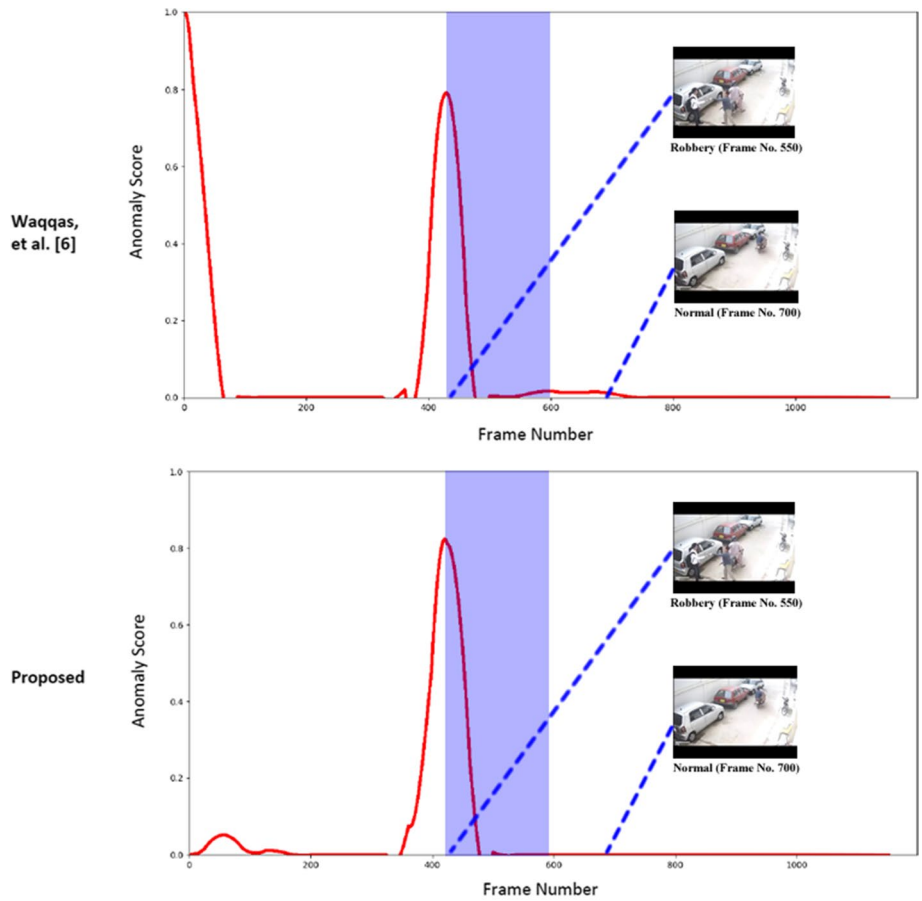
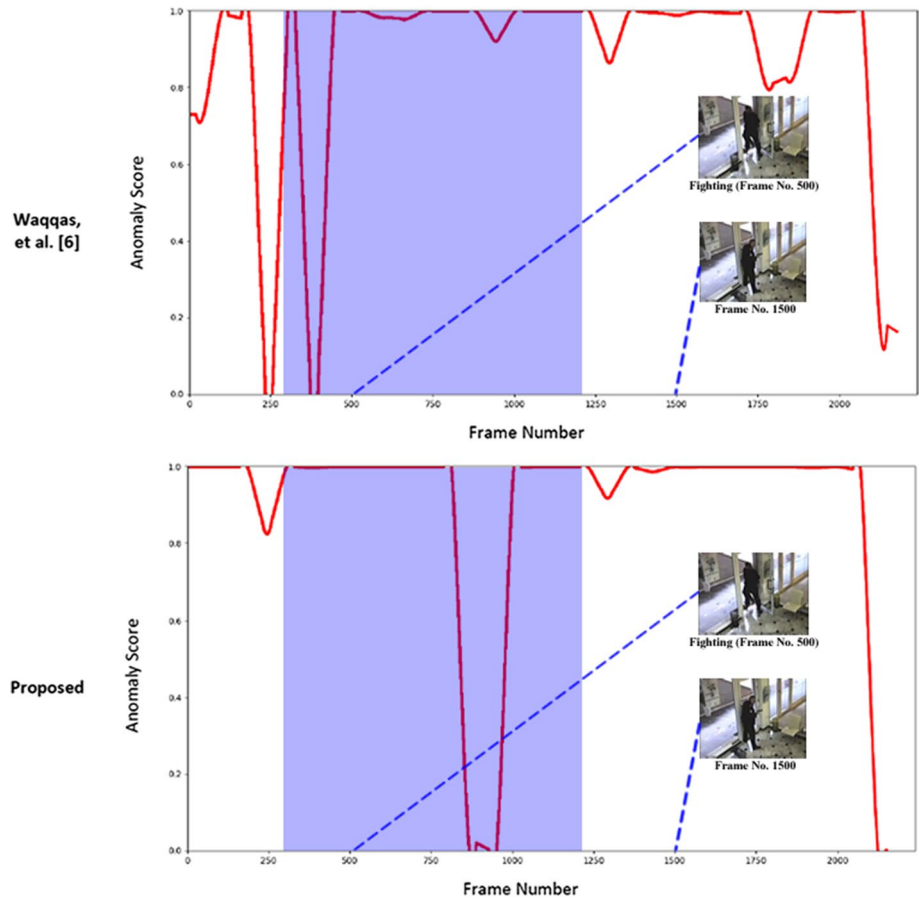


Fig. 9 Qualitative results of our proposed method and previous work by Waqqas et al. [6] on testing videos. Colored window (blue) shows ground truth anomalous region. **a**–**c** show videos containing robbery, fighting, and road accident, respectively

As clearly shown in Fig. 9 that we only feed the attention regions to the deep-learning pipeline. In other words, the uninterested region will be blurred and will not produce any important visual features during extraction, training and inference process. Note that, the x and y-axis indicate frame numbers and probability of anomaly scores, respectively. We compare our proposed method with previous work by Waqqas [6]. Although there are several types of research which introduce the localization in anomaly detection, their approach focused on one anomaly event, such as fighting action (as proposed by Xu [10]). Therefore, in order to evaluate several anomalous events, we compare the proposed work with the full-frame approach. In Fig. 9a shows two-person approaching a man and rob his mobile phone, then leaving the area by motorbike. Clearly, both works are able to detect the robbery event with high probability (see the highlighted frame No. 550 in Fig. 9a). In addition, in subsequent highlighted frame No. 700 in Fig. 9a, we visualize the normal event when the two robbers leaving the scene by motorbike. Our approach is more accurate by yielding a significantly lower score anomalous probability (almost zero). Note that, the higher the probability score, the more likely anomaly events will be. In an anomaly fighting scene, a security officer is trying to protect the area from



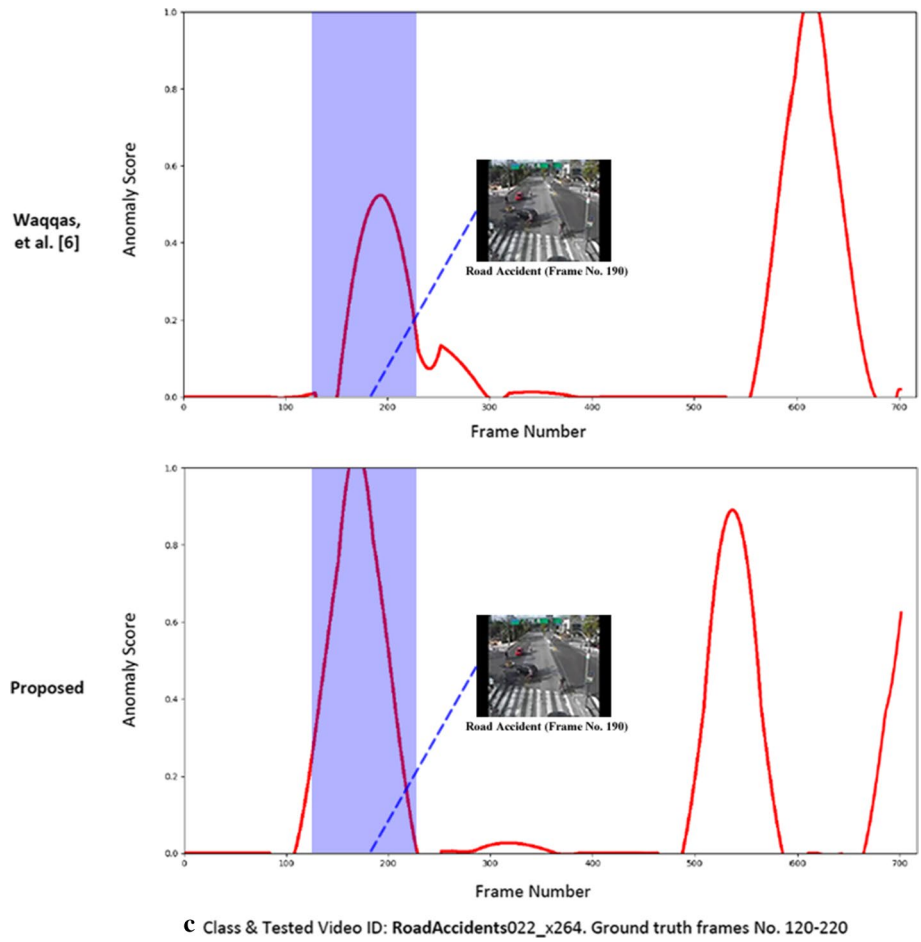
b Class & Tested Video ID: Fighting042_x264. Ground truth frames No. 290-1200

Fig. 9 continued

the intruder, while in normal scene illustrates the intruder leaves the area after failing to fight against the officer. As visualized in Fig. 9b, the anomaly score of Waqqas [6] and the proposed work is very competitive. Similarly, Fig. 9c visualizes the road accident scene that the anomaly score of our proposed work outperforms the previous work [6].

In Fig. 10a–c, the respective highlighted frames from Fig. 9 in higher resolution are provided. It is clearly shown that our method is able to localize the anomalous regions successfully through the BG subtraction idea. From the industrial application point of view, the extraction of this attention region can assist the security officer on focusing on the corresponding anomaly region, instead of a wider, full-framed inspection.

As shown in Table 3 above, it is obvious by applying the proposed localization approach achieves higher accuracy on several tested videos. The accuracy of each tested video was calculated from each segment that contains anomalous events. For example, the robber started their action from video segment 14 to 15, the fighting was initially begun from segment 4 to 18, and road accidents occurred very quickly start from segment 5 to 7. Therefore, the accuracy needs to be evaluated on several segments and calculate the average score. In average, our proposed approach is able to obtain stable accuracy on every tested video (as concluded in Fig. 11).



c Class & Tested Video ID: **RoadAccidents022_x264. Ground truth frames No. 120-220**

Fig. 9 continued

Furthermore, in order to compare the accuracy of trained model, we also collect 135 test videos from UCF-Crime dataset and extract corresponding C3D features. The accuracy is simply calculated by accumulating the correct predictions over the number of tested videos. In [6], 133 of 135 videos are labeled correctly, while our proposed visual attention learning can classify 134 of 135 videos correctly. The details can be found in Table 4 below:

In the real-world scene, multiple events are possibly occurring in one CCTV footage. For example, a robber tries to rob something but at the same time, the victim is

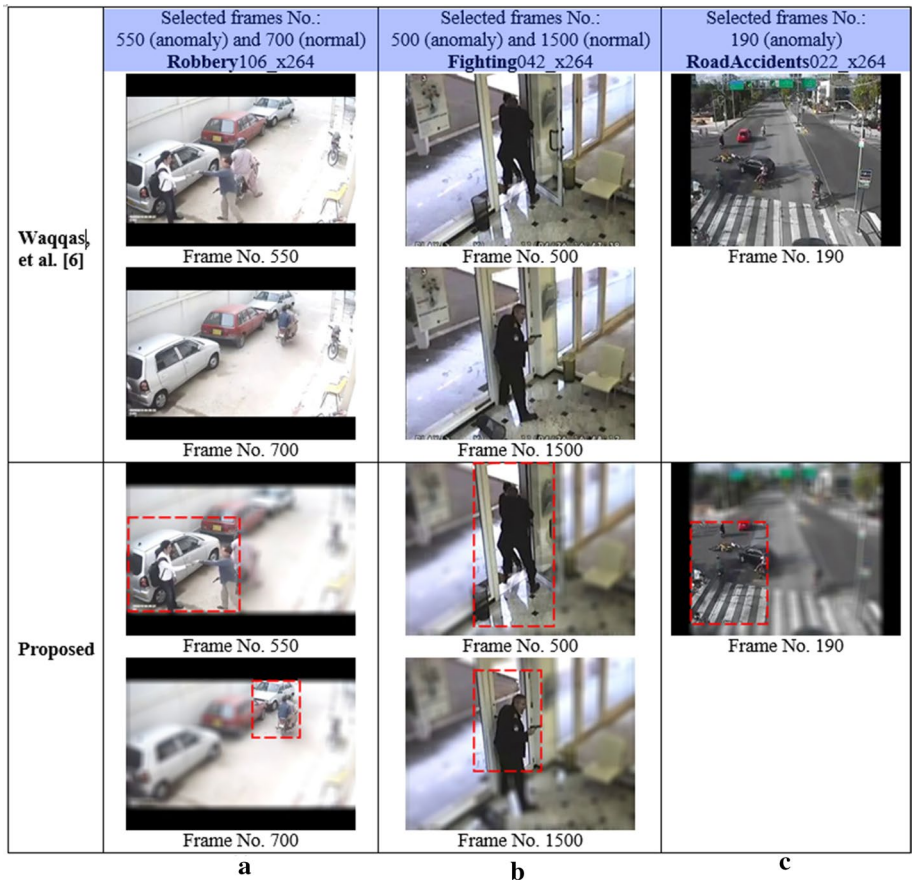


Fig. 10 Corresponding highlighted frames from Fig. 9 in higher resolution and clear illustration of the localized region from the proposed method. **a**–**c** show videos containing robbery, fighting, and road accident, respectively

Table 3 The accuracy (%) evaluation of tested videos between full-frame and our proposed locality learning

Method	Dataset		
	Robbery_106	Fighting_042	Road accident_022
Sultani et al. [6]	*99.9*	77.7	35.63
Proposed method	96.5	*98.2*	*92.53*

fighting back. Therefore, we also visualize one tested video that we obtained arbitrarily through YouTube. This scene has shown a robber was approaching a group of people in the station, but he failed to accomplish the action and those people were successfully self-defense their goods. Similar to the previous UCF- Crime evaluation, we conduct a thorough analysis by examining each video segment and measure its accuracy. As shown in qualitative measurements below, the proposed localized approach is able to detect two separated anomalous segments, while the previous work is failed to detect the event when the group of people was fighting back and force the robber escaping from the area.

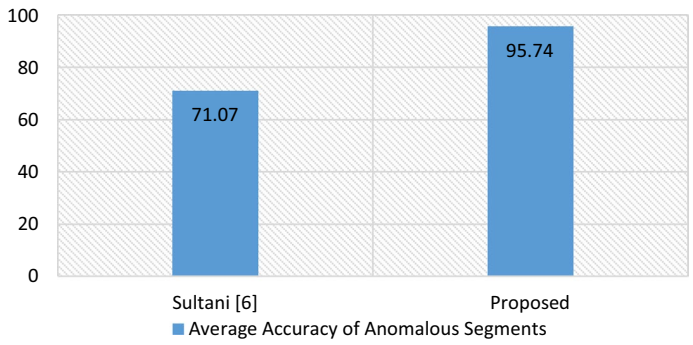


Fig. 11 Average comparisons of accuracies in UCF-Crime dataset

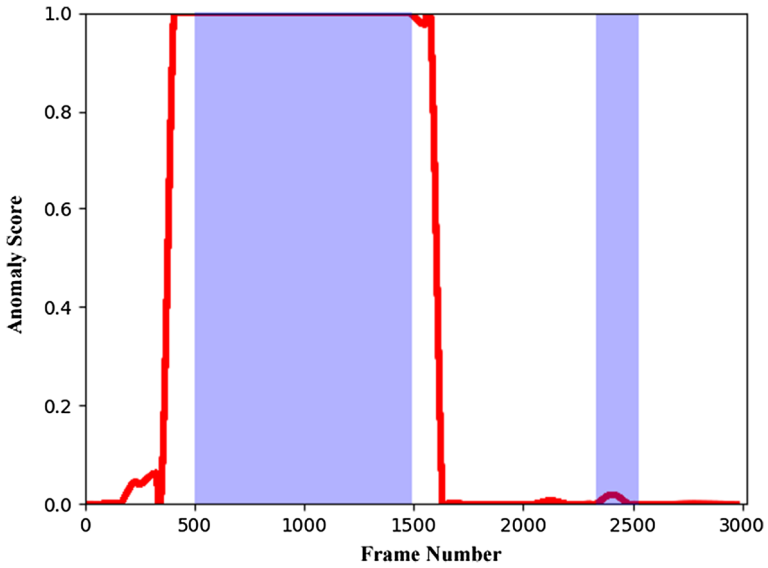
Table 4 The comparison of accuracy (%) between full-frame and our proposed locality learning

Method	Accuracy
Sultani et al. [6]	98.51
Proposed method	*99.25*

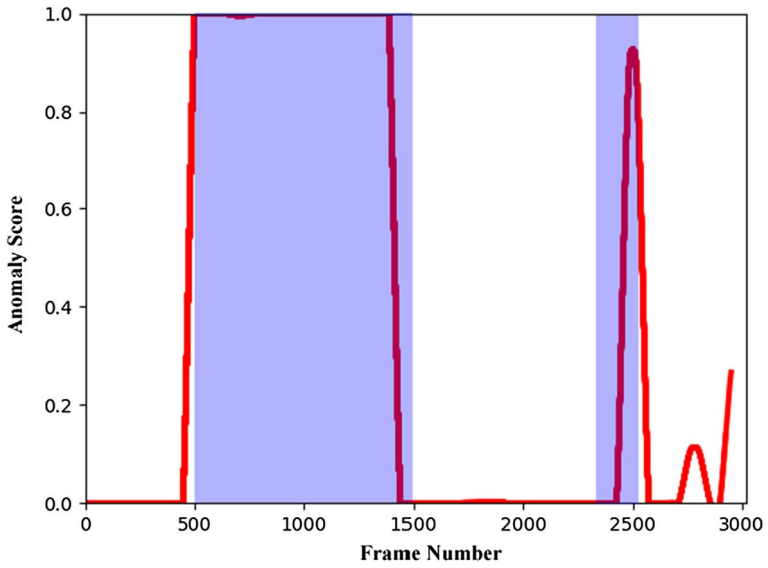
Table 5 The accuracy (%) evaluation of a Multi-events tested video from youtube between full-frame and our proposed locality learning

Method	Dataset
	Robbery and fighting
Sultani et al. [6]	50.84
Proposed method	*88.13*

The ground-truth are manually labeled through manual inspection frame-by-frame. The events are occurring from frame no. 500 to 1500, then continuing from frame no. 2300 to 2500. The corresponding accuracies of anomalous segments are provided in Table 5.


1.0
0.8
0.6
0.4
0.2
0.0
Anomaly Score
0
500
1000
1500
2000
2500
3000
Frame Number


Frame No. 250
1000
1500
2000
2500
Normal Scene
Localized Anomalous Event
Localized Anomalous Event


1.0
0.8
0.6
0.4
0.2
0.0
Anomaly Score
0
500
1000
1500
2000
2500
3000
Frame Number

Conclusions

In this paper, an automatic localization through robust computer vision techniques in anomaly detection is proposed. Experimental results show that: (1) finding a localized attention region from each segment helps anomaly detection; (2) our method is able to obtain accurate results in different kinds of event, e.g. road accident, robbery, and fighting and (3) Incorporating a robust BG subtraction can help to find the region of interest (ROI) as correct as possible. In terms of accuracy, the proposed visual attention model can obtain 99.25 of accuracy. It is noteworthy, we utilize a weakly-supervised network for training. More generally, we believe that our approach of extracting the visual attention region could benefit many other online tasks, such as video object localization and classification, and plan to pursue this in future work. Our work is limited to the anomaly events which contain the moving object.

Acknowledgement
This work was supported by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

Authors' contributions
Conceptualization, NN; methodology, KM and APJD; software, KM; validation, AA; project administration, NN. All authors read and approved the final manuscript.

Funding
This research was funded by The Ministry of research, technology and higher education of the Republic of Indonesia, Grant number: 90/UN11.2.1/PT.01.03/DPRM/2020.

Availability of data and materials
https://tinyurl.com/y85ff62d

Competing interests
The authors declare no conflict of interest.

Author details
[1] Department of Electrical and Computer Engineering, Syiah Kuala University, Aceh, CO 23111, Indonesia. [2] Telematics Research Center (TRC) Universitas Syiah Kuala, Aceh, CO 23111, Indonesia. [3] Nodeflux, Jakarta, CO 12730, Indonesia.

Received: 16 June 2020 Accepted: 8 October 2020
Published online: 16 October 2020

References

1. Mohammadi S, Perina A, Kiani H, Murino V. Angry crowds: detecting violent events in videos. In: ECCV, 2016.
2. Esen E, Arabaci MA, Soysal M. Fight detection in surveillance videos. In: 11th Int. Workshop on Content-Based Multimedia Indexing, 2011.
3. Nievas EB, Suarez OD, Garc´ıa GB, Sukthankar R. Violence detection in video using computer vision techniques. In: CAIP, 2011.
4. Kamijo S, Matsushita Y, Ikeuchi K, Sakauchi M. Traffic monitoring and accident detection at intersections. IEEE Trans Intell Transp Syst. 2000;1:108–18.
5. Sultani W, Choi JY. Abnormal traffic detection using intelligent driver model. In: ICPR, 2010.
6. Sultani W, Chen C, Shah M, Real-world anomaly detection in surveillance videos. In: CVPR, 2018.
7. Andrews S, Tsochantaridis I, Hofmann T, Support Vector Machines for Multiple-Instance Learning. In: Advances in neural information processing systems, 2003.
8. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89:31–71.
9. Landi F, Snoek CGM, Cucchiara R. Anomaly Locality in Video Surveillance, arXiv preprint arXiv:1901.10364, 2019.
10. Xu Q, See J, Lin W. Localization guided fight action detection in surveillance videos. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019.
11. Jain M, Gemert JV, e. J´egou H, Bouthemy P, Snoek CG. Action localization with tubelets from motion. In: CVPR, 2014.
12. Lessard FBA, Bilodeau G-A, Saunier N. The countingapp, or how to count vehicles in 500 hours of video. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2016.
13. Xu D, Ricci E, Yan Y, Song J, Sebe N. Learning deep representations of appearance and motion for anomalous event detection. In: BMVC, 2015.
14. Wu S, Moore BE, Shah M. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: CVPR, 2010.
15. Basharat A, Gritai A,Shah M. Learning object motion patterns for anomaly detection and improved object detection. In: CVPR, 2008.

16. Cui X, Liu Q, Gao M, Metaxas DN Abnormal detection using interaction energy potentials. In CVPR, 2011.
17. Antic B, Ommer B. Video parsing for abnormality detection. In ICCV, 2011.
18. Hospedales T, Gong S, Xiang T. A markov clustering topic model for mining behaviour in video. In: ICCV, 2009.
19. Zhu Y, Nayak IM, Roy-Chowdhury AK. Context-aware activity recognition and anomaly detection in video. IEEE J Select Topics Signal Process. 2012;7(1):91–101.
20. Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell. 2013;36(1):18–32.
21. Kratz L, Nishino K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: CVPR, 2009.
22. Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab. In: ICCV, 2013.
23. Zhao B, Fei-Fei L, Xing EP. Online detection of unusual events in videos via dynamic sparse coding. In: CVPR, 2011.
24. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS. Learning temporal regularity in video sequences. In: CVPR, 2016.
25. Cheng K-W, Chen Y-T, Fang W-H. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In: CVPR, 2015.
26. Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection. In: CVPR, 2011.
27. Dutta JK, Banerjee B Online detection of abnormal events using incremental coding length. In AAAI, 2015.
28. Ionescu RT, Smeureanu S, Popescu M, Alexe B. Detecting abnormal events in video using narrowed normality clusters. In: WACV, 2019.
29. Kim J, Grauman K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In: CVPR, 2009.
30. Mehran R, Oyama A, Shah M, Abnormal crowd behavior detection using social force model. In: CVPR, 2009.
31. Ren H, Liu W, Olsen SI, Escalera S, Moeslund TB. Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection. In: BMVC, 2015.
32. Xu D, Yan Y, Ricci E, Sebe N. Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst. 2017;156:117–27.
33. Zhang Y, Lu H, Zhang L, Ruan X, Sakai S. Video anomaly detection based on locality sensitive hashing filters. Pattern Recogn. 2016;59:302–11.
34. Kooij J, Liem M, Krijnders J, Andringa T, Gavrila D. Multi-modal human aggression detection. Comput Vis Image Underst. 2016;144:106–20.
35. Saleemi I, Shafique K, Shah M. Probabilistic modeling of scene dynamics for applications in visual surveillance. IEEE Trans Pattern Anal Mach Intell. 2009;31(8):1472–85.
36. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Proc Image Commun. 2016;47:358–68.
37. Jian M, Lam K-M, Dong J. Illumination-insensitive texture discrimination based on illumination compensation and enhancement. Inf Sci. 2014;269:60–72.
38. Lin C-Y, Muchtar K, Lin W-Y, Jian Z-Y. Moving object detection through image bit-planes representation without thresholding. IEEE Transact Intell Transport Syst. 2019;21:1–11.
39. Zivkovic Z. Improved adaptive Gaussian mixture model for background subtraction. In Cambridge: ICPR, 2004.
40. Zivkovic Z, d Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognit Lett. 2006;27(7):773–80.
41. Tomasi C, Manduchi R. Bilateral Filtering for Gray and Color Images. In: IEEE International Conference on Computer Vision, 1998.
42. Yeh C-H, Lin C-Y, Muchtar K, Kang L-W. Real-time background modeling based on a multi-level texture description. Inf Sci. 2014;269:106–27.
43. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: ICCV, 2015.
44. Fa L, Song Y, Shu X, Global and Local C3D Ensemble System for First Person Interactive Action Recognition. In: International Conference on Multimedia Modeling, 2018.
45. Bendali-Braham M, Weber J, Forestier G, Idoumghar L, Muller P-A, Transfer learning for the classification of video-recorded crowd movements. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), 2019.
46. Liu K, Liu W, Ma H, Tan M, Gan C. A Real-time Action Representation with Temporal Encoding and Deep Compression, IEEE Transactions on Circuits and Systems for Video Technology (Early Access), 2020; p. 1.
47. Fan Y, Lu X, Li D, liu Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016.
48. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L, Large-scale Video Classification with Convolutional Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
49. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Machine Learn Res. 2014;15:1929–58.

Publisher's Note