# Abnormal Events and Behavior Detection in Crowd Scenes Based on Deep Learning and Neighborhood Component Analysis Feature Selection

Alaa Atallah Almazroey[1(✉)] and Salma Kammoun Jarraya[1,2]

[1] Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
aalmazroey0001@stu.kau.edu.sa, smohamadl@kau.edu.sa
[2] MIR@CL Laboratory, Sfax University, Sfax, Tunisia

**Abstract.** In the last few years, surveillance cameras have been massively distributed both indoors and outdoors in public, due to security concerns creating a need to monitor unexpected actions or activities in a scene. An intelligent automated approach is highly required to detect anomalies from the scene as well, to save the time and cost required by laborers to detect the anomalies manually from monitor screens. In this research, we propose a deep learning-based approach to detect abnormal events and behavior from surveillance videos in crowd scenes. Thus, by using the keyframe extractor to localize the keyframes that contain important information from video frames. The selected keyframes are used to compute the optical flow values of magnitudes, orientations, and velocities of each keyframe to generate several 2D templates. Then, the obtained 2D templates are supplied to a pre-trained model 'AlexNet' to extract high-level features. The Neighborhood Component Analysis (NCA) feature selection method is applied to select the appropriate features, then use these features to generate a classification model via Support Vector Machine (SVM) classifier. Results are evaluated on several public datasets, along with a new dataset that we built it that contains different videos covering abnormal events and behaviors. The obtained results proved that the proposed method outperforms other methods.

**Keywords:** Abnormal events detection · Keyframes selection · Neighborhood Component Analysis · Video surveillance

## 1 Introduction

Anomaly detection from a video is one of the fundamental tasks in a video surveillance system. Surveillance cameras are increasingly being distributed and used for different purposes, such as security reasons, avoiding traffic congestion, and crowd control. An abnormal event is an event that affected by external causes; for example, an escape that may be caused by a natural disaster i.e. earthquake, or by other abnormal behavior, i.e. explosion. While abnormal behavior is an attitude related to conducting individual or group activities, such as walking or driving in the wrong direction. However, it is still

© Springer Nature Switzerland AG 2020
A.-E. Hassanien et al. (Eds.): AICV 2020, AISC 1153, pp. 258–267, 2020.
https://doi.org/10.1007/978-3-030-44289-7_25

quite challenging to design a general framework for detecting all abnormal events and behaviors from a scene. This is because there is no exact definition or dictionary term that is currently used to explain what the abnormal events or abnormal behaviors are, since the boundary between normal and abnormal is often unclear and might vary depending on the situation, i.e. day/night time, peak hours/off-peak hours, etc. [1]. Therefore, activities related to each other by space and time form the context. The early technique to detect anomalies from a surveillance camera was a non-intelligent practice that monitored multiple screens continuously and were checked primarily by a human, which was considered as a critical task that needs high levels of attention, since the anomalies rarely occur when compared to normal activities. Therefore, to alleviate this issue, developing an intelligent system is much desired to detect abnormal events and behaviors in crowd scenes automatically. Although, detecting abnormal events is still a challenging task due to different scenarios, video quality, crowded and complex scenes, occlusions, and other factors. Most of the existing models are proposed based on a hand-crafted approach that trained and designed for a particular dataset to detect a specific kind of abnormal events or behavior from a particular scene under a specific condition. These approaches are facing a problem when using them with new datasets since each dataset has its own bias. In reality, any scene may contain complicated and different types of abnormal events and behaviors under different conditions that all need to be detected. In view of this, we propose a deep learning-based approach to extract deep features and detect abnormal events and behavior in crowd scenes. Firstly, by applying a keyframe extractor to extract keyframes, which are video frames that summaries the video and eliminate adjacent redundant frames. Thus, by using the Cosine Similarity algorithm (CS) [2] to only select the frames that contain new information. Then the selected keyframes are processed to extract the optical flow components magnitude, orientation, and spatial components via the Lucas–Kanade method [3] before being combined and fed to a pre-trained convolutional neural network AlexNet [4] to obtain deep features. Eventually, the Neighborhood Component Analysis (NCA) method [5] is performed to obtain the most effective features before submitting the selected features to a binary SVM classifier. The proposed method is tested with UCSD Ped1and UCSD Ped2 [6], Avenue [7], Hockey dataset [8], including our dataset that consists of collected videos that cover various abnormal events, such as panic, smoke and fire, and abnormal behaviors such as fighting, throwing an object, etc.

The rest of this paper is structured as follows: In Sect. 2, we present the literature review for some of the existing methods. In Sect. 3, we illustrate the proposed methodology; the experiment's performance, quantitative results are discussed in Sect. 4; finally, we conclude the paper in Sect. 5.

## 2 Literature Review

Abnormal events and behavior detection refer to detecting and responding to the abnormal changes in videos. Recently, many researchers have been exploring how to design an effective model in order to accurately detect an anomaly. One of the early attempts at anomaly detection approaches is a hand-crafted representation. In these approaches, features are extracted from the input videos, which then requires an expert to

design a model based on the specific extracted features. Such an approach can be classified into two categories: object trajectory-based approaches, as in [9, 10], and non-object trajectory-based approaches [11, 12]. In [9, 10], the authors proposed a trajectory-based approach that tracks objects and defines the global anomaly behavior. Such methods can provide an acceptable result when a scene contains few objects, but in crowded scenes, a trajectory-based approach becomes unreliable, as it is limited by occlusion. As alternatives, non- trajectory based approaches do not attempt to track objects, and propose to extract low-level features from pixel-level, such as optical flow (OF) [13], and Histogram of Oriented Gradients (HOG) [14], to detect both local and global anomaly in crowd scenes. However, the hand-crafted approach is suboptimal, as it is limited by the characteristics obtained. In recent years, deep learning has become gradually popular and commonly used in intelligent video surveillance systems to overcome the limitations of hand-crafted approaches, as no specific extraction procedures are needed. Deep learning approaches demonstrated significant performance on different domains like image classification [15], object detection [16], and activity recognition [17]. The generative approach is one of the deep learning approaches that used with unsupervised learning model that can understand and describe how the data is generated, in terms of a probabilistic model. One of the generative approach is Auto-Encoders model (AE), which has the ability to extract features efficiently without needing prior knowledge [18]. Sabokrou et al. [19] learned and serves two independent AE models, one to extract local features for local anomalies, and the second model to extract global features. Then, combine results from both models and trained Gaussian models to detect anomalies. Sparse Auto-Encoder (SAE) [20], and stack denoising auto-encoder (SDAE) [21, 22] are AE techniques that applied to speed up the model generating and reduce the dimensionality required for training. In [20], (SAE) is used to extract features from video frames. Then, the recurrent neural network (RNN) is used to generate a model. Feng et al. [21] applied two stacked denoising auto-encoder (SDAE), one for extracting appearance, texture, and short-term motion features, and the other for fusing the extracted features. Likewise, Xu et al. [22] applied stacked denoising autoencoders (SDAs) to extract both motion and appearance information. However, the networks used in their work are shallow and based on small image patches. Recently, fusion LSTM and Convolutional AE were proposed in [23] authors applied RNN model to learn appearance and motion features of long-term video. While in [24], the authors applied the Epic-Flow method to extract optical flow from the input data, then feeding the weighted convolutional Auto-Encoder long short-term memory (WCAE-LSTM) networks with these optical flows and raw input. Another deep learning approach is a discriminative approach that is a supervised learning method that uses labelled data for classification. Convolution neural networks (CNN) is an example of the discriminative model, which has a powerful tool to extract deeper and more discriminate hidden features. However, CNN requires a large number of training videos to avoid overfitting. To alleviate the limitations of CNN, fully convolutional neural networks (FCN) have been used for the first time by Sabokrou et al. [25]. Thus, by employing a pre-trained CNN model which reduces the computational cost and achieved three times faster detection of abnormal events than only using a traditional CNN. As in [26], proposed by computing the OF values and used these values to generate several templates that used as input into a pre-trained CNN to extract deep features. Wei et al. [27] proposed two-stream FCN network. The first stream of FCN for original frame input is to extract

appearance features, and the second stream uses OF to extract motion features between the video frames. Then, the combination of these features provides convolutional features. Lately, a hybrid approach was applied by combining the discriminative and generative architectures together. As in [28], the authors used (3D-GANs) to detect a temporal anomaly. They trained the model with only regular videos and detected irregularities according to the deviation estimated. In [29] proposed a combination of pre-trained CNN and LSTM to extract spatial-temporal features. However, the work in [30] enhanced the method used in [29] by proposing a Bidirectional Convolutional LSTM (BiConvLSTM) network. Thus, by applying a pre-trained network to extract appearance features, and then passing these features to the BiConvLSTM to encode temporal information in both directions.

In summary, there are many successful approaches in the related field of anomaly detection, as represented above. However, all these approaches are designed for detecting either abnormal events or abnormal behavior from video, but not both. In our work, we adopt a discriminative approach, as we provided labelled data. In addition, we tend to use a pre-trained CNN, since it does not require a large dataset for training.

## 3 Proposed Method Overview

The proposed method is based on a supervised deep learning-based approach for detecting abnormal events and behaviors from a crowd scene. The process consists through two stages: The first stage is the pre-processing step as illustrated in Fig. 1. Initially, we converted the input videos to a set of frames. After that, a keyframe selection method is applied to these frames using the Cosine Similarity (CS) algorithm, to calculate the difference between two frames (i.e. the current frame and the previous keyframe). The procedure starts by checking if the current frame is a first frame in the video frames, if it is then we save that frame in a buffer as the first keyframe; otherwise, (is not the first frame) then the CS algorithm is used to calculate the differences

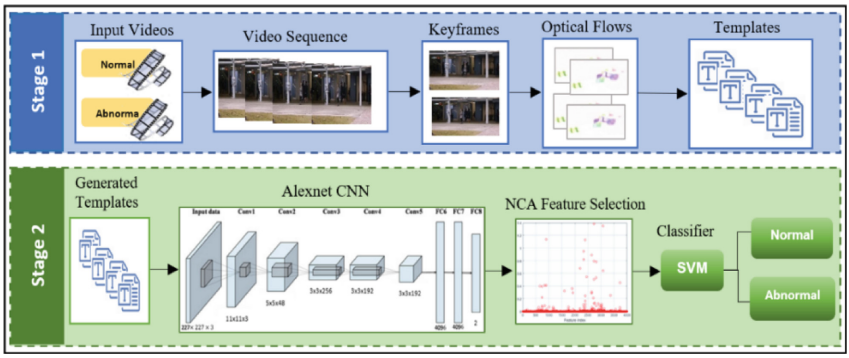


**Fig. 1.** The two stages of the proposed method to generate abnormal events and behavior detection model.

between the current frame and the previously extracted keyframe as in (1) [2]. As $A$ refers to the previous selected keyframe and $B$ refers to the current frame and $n$ states number of frames.

$$\text{Cosine Similarity (CS)} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

When the obtained CS result does not exceed the similarity threshold value, then it means that the two frames are different, and the current frame will be considered as a new keyframe. Then we saved it in a buffer and used it to extract the next keyframe.

The similarity threshold value was selected by calculating the CS between video frames for all input videos in order to determine the appropriate threshold value. The closer CS value to 1, the smaller changes between the two frames [2]. Figure 2 represents the average CS values for the abnormal and normal videos from Hockey, UCSD ped1, UCSD Ped2, and Avenue datasets. Moreover, we noticed that most of the CS values are ranged between 0.90 and 1. After several empirical experiments, we set a similarity threshold by 0.995. Secondly, in this stage, we used the nominated keyframes to estimate the optical flow of vertical and horizontal velocity, magnitude and orientation for each adjacent keyframes via the Lucas–Kanade method [3]. So, each video generates four different 2D templates [26]. The first template represents the orientation of motion by accumulating orientation values of each pixel. The second template calculates the dot product of magnitude and orientation values to obtain the speed of motion for an object. Moreover, the last two templates contain the horizontal and vertical velocity components along with magnitude information. Thus, the use of these templates makes the feature extraction step focus on motion regions instead of the all spatiotemporal features. In the second stage (Fig. 1), the generated templates from the previous stage are used as an input to a pre-trained CNN, Alexnet [4] to extract high-level features from each template. The network consists of five convolutional layers and three fully connected layers (FC). A dropout regularization [31] (50%) is placed between FC layers, to prevent an overfitting problem. Then we merged each template's extracted features to get the final big features collection. After that, we applied the Neighborhood Component Analysis (NCA) features selection method. Thus, by tuning the regularization parameter lambda ($\lambda$) for the NCA using the five-fold cross-validation for training dataset that assigns four folds for training set, and one-fold for validation. Then, we trained the NCA model for each lambda ($\lambda$) value in each fold based on used the training sets. In addition, we recorded the classification loss value for the corresponding validation set in the fold via the NCA model. After that, we selected the minimum loss value to get the best lambda, which used to generate the final NCA model used for feature selection. Furthermore, the number of extracted features in each generated model varies, and all the extracted features are important since the effective features are selected based on the weight of the feature, as the weight of features indicates how much each feature affects the classification issue. This procedure helps to reduce the loss values, maximize the accuracy and avoid overfitting. After that, we used the selected features to test the model through different classifiers, where the linear SVM is the one that produced better accuracy results.

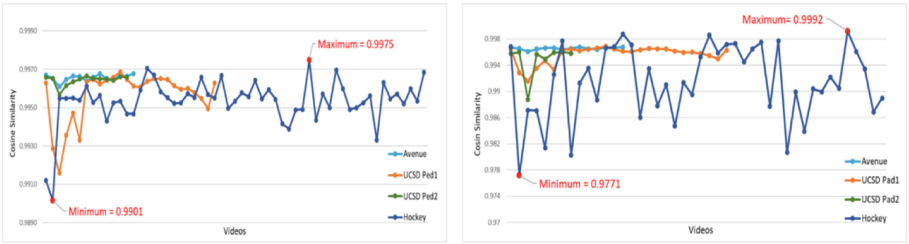


**Fig. 2.** The average of Cosine Similarity values (a) normal videos; (b) abnormal videos for all datasets.

## 4 Experiments and Results

### 4.1 Datasets Description

Our proposed model is trained and tested with four different public datasets, which are UCSD Ped1 [6], UCSD Ped2 [6], Avenue [7], and Hockey dataset [8]. Including our dataset that consists of other collected videos that contain abnormal events and behaviors. We divided the dataset by splitting 70% of the videos for the training set and 30% for the testing set. A combination of normal and abnormal videos is used for training.

***Public Datasets.*** **UCSD Ped1** has 70 videos for training and testing. Normal behavior is people moving towards and away from the camera. **UCSD Ped2** contains 28 videos. The normal behavior here is people moving parallel to the camera. The anomalies for both datasets include people walking on grass, skaters, bikers, and people in wheelchairs. **Avenue dataset** consists of 37 videos for training and testing. The anomalies are persons walking toward the camera, running, and throwing an object. **Hockey dataset** is used for fighting detection, it consists of 500 clips containing fighting and another 500 clips for non-fighting. Overall, all the previous public datasets contain only abnormal behaviors. However, in real-life, both abnormal events and behavior may occur simultaneously in the scene and need to be detected.

***Collected Videos.*** Due to the limitation in the previous public datasets (as all these datasets contain only abnormal behaviors), we recorded videos containing numerous abnormal events, such as panic from a natural disaster (earthquake), fire smoke, and flames from cars and motorcycles in a petrol station. These videos were collected from YouTube and we only selected videos captured by CCTV cameras. Then, we combined these videos with the previous public dataset videos to obtain a large dataset that contains both abnormal events and behaviors.

We applied the keyframes selection method to only work with the important frames from each dataset. Table 1 represents the average number of the selected keyframes from each video in these public datasets. As represented in Table 1, there is a significant reduction in the number of frames to be processed. Therefore, keyframe selection improves the efficiency of computation.

**Table 1.** Dataset description: public anomaly datasets and our training and testing dataset

| Dataset name | No. of videos | Average frames | Average keyframes | Anomaly type |
|---|---|---|---|---|
| UCSD Ped1 | 70 | 200 | 93 | Behavior |
| UCSD Ped2 | 28 | 163 | 57 | Behavior |
| Avenue | 37 | 180 | 76 | Behavior |
| Hockey | 1000 | 41 | 33 | Behavior |
| Collected videos | 73 | 395 | 156 | Events |
| **Our training and testing dataset** | **1,208** | **195** | **83** | **Events and behavior** |

### 4.2 Experimental Results Discussion

We used the Area Under ROC Curve (AUC) as the measurement to evaluate our proposed model. All experiments are conducted with processor Intel® Core™ i7–8750H and a graphics processor NVIDIA GeForce GTX 1050 Ti.

The keyframe selection method improved the required processing time for classification. Figure 3 shows the required time for classification of one video from the Hockey dataset with and without keyframe selection are 0.72 s and 0.89 s, respectively. The measurement includes the generation of templates, features extraction, and prediction.

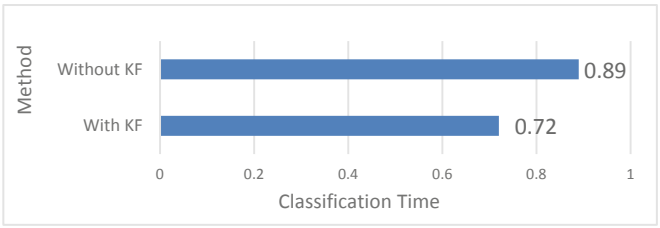


**Fig. 3.** Required time for classification with and without keyframe selection method for one video of Hockey dataset of 2 s.

We compared the results of our model with other existing methods on all the four datasets for anomaly detection as presented in Table 2. The obtained results showed that our model performs better than state-of-the-art methods on UCSD ped1, Avenue, and Hockey dataset. We record 93.75%, 87.5%, and 98.14%, respectively. Thus, the obtained model is sufficiently robust to detect anomalies from the crowd scene.

**Table 2.** AUC comparison of our model against other state-of-the-art methods.

| Methods | AUC (%) | | | |
|---|---|---|---|---|
| | UCSD Ped1 | UCSD Ped2 | Avenue | Hockey |
| Stack Denoising AE (SDAs) [22] | 92.1 | 90.8 | N/A | N/A |
| Convolutional AE + LSTM [23] | 89.9 | 87.4 | 80.3 | N/A |
| Convolutional AE [24] | 89.1 | **94.8** | 87.2 | N/A |
| 3D_GAN [28] | N/A | N/A | 79.68 | N/A |
| Optical flow + CNN [26] | N/A | N/A | N/A | 94.40 |
| CNN + BiConvLSTM [30] | N/A | N/A | N/A | 96.54 |
| CNN + LSTM [29] | N/A | N/A | N/A | 97.1 |
| **Proposed model on testing Datasets** | **93.75** | 83.3 | **87.50** | **98.14** |

In addition, we compared the results of our model with Keçeli et al. [26]. The authors in [26], applied the Relieff features selection method [31] to select the best 10% of features from all video frames. However, in our method we used the NCA feature selection method with k-fold cross-validation from only keyframes. The set of features selected by the NCA method gives better results in terms of accuracy and generalization. Table 3 shows the obtained accuracy for the Hockey dataset using keyframes is 98.14%, while in [26] it records an accuracy of 94.4% using all frames. As for our dataset, we tested it with both feature selection methods, with and without the keyframes method. The features selected by the NCA method that applied on keyframes provide higher accuracy than those selected by Relieff method for all frames. This improvement in accuracy is due to the NCA method tunes the regularization parameter lambda (λ) through k-fold cross-validation by updating the (λ) value for each fold to find the best lambda that produces the lowest classification loss.

**Table 3.** AUC percentage obtained from the Hockey and Our dataset, by using the Relieff method [26] with all frames in comparison with the NCA feature selection method with keyframes only.

| Features selection methods | AUC (%) | |
|---|---|---|
| | Hockey dataset | Our testing dataset |
| Relieff feature selection from all video frames [26] | 94.4 | 97.0 |
| **Proposed model based on the NCA feature selection from Keyframes** | **98.14** | **98.3** |

## 5 Conclusions

In this paper, we proposed a deep-learning approach to detect real-world anomalies from surveillance videos. This approach achieves fast anomaly detection from the videos since it is based on a keyframe selection method to process only with important

frames, then we calculated the optical flow values to generate multiple templates, which represent the continuity of the motion between frames in the video. After that, we fed the pre-trained CNN with these templates to extract deep features. We used the NCA feature selection method to select the appropriate features and trained the model using linear SVM. To validate the obtained model for abnormal events and behavior detection, we tested the model on different well-known public datasets such as UCSD Ped1, UCSD Ped2, Avenue, and Hockey dataset. Also, we have collected several videos from YouTube containing other anomalies like panic, fire smoke, and fire on cars and motorbike. The experiment results proved that our proposed abnormal events and behavior detection model performs significantly better than state-of-the-art methods.

## References

1. Yuan, Y., Feng, Y., Lu, X.: Statistical hypothesis detector for abnormal event detection in crowded scenes. IEEE Trans. Cybern. **47**(11), 3597–3608 (2016)
2. Li, Y., et al.: Key frames extraction of human motion capture data based on cosine similarity. Vectors **11**(12), 1 (2017)
3. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. Int. J. Comput. Vis. **12**(1), 43–77 (1994)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
5. Yang, W., Wang, K., Zuo, W.: Neighborhood component feature selection for high-dimensional data. JCP **7**(1), 161–168 (2012)
6. Mahadevan, V., et al.: Anomaly detection in crowded scenes. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE (2010)
7. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in MATLAB. In: Proceedings of the IEEE International Conference on Computer Vision (2013)
8. Nievas, E.B., et al.: Violence detection in video using computer vision techniques. In: International Conference on Computer Analysis of Images and Patterns. Springer, Heidelberg (2011)
9. Ghrab, N.B., Fendri, E., Hammami, M.: Abnormal events detection based on trajectory clustering. In: 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV). IEEE (2016)
10. Coşar, S., et al.: Toward abnormal trajectory and event detection in video surveillance. IEEE Trans. Circ. Syst. Video Technol. **27**(3), 683–695 (2016)
11. Li, N., et al.: Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing **155**, 309–319 (2015)
12. Cho, S.-H., Kang, H.-B.: Abnormal behavior detection using hybrid agents in crowded scenes. Pattern Recogn. Lett. **44**, 64–70 (2014)
13. Liu, W., et al.: Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
14. Li, F., Yang, W., Liao, Q.: An efficient anomaly detection approach in surveillance video based on oriented GMM. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2016)

15. Deecke, L., et al.: Image anomaly detection with generative adversarial networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Heidelberg (2018)
16. Kumaran, S.K., Dogra, D.P., Roy, P.P.: Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey. arXiv preprint arXiv:1901.08292 (2019)
17. Zhu, Y., Nayak, N.M., Roy-Chowdhury, A.K.: Context-aware activity recognition and anomaly detection in video. IEEE J. Sel. Top. Sig. Process. **7**(1), 91–101 (2012)
18. Liou, C.-Y., et al.: Autoencoder for words. Neurocomputing **139**, 84–96 (2014)
19. Sabokrou, M., et al.: Real-time anomaly detection and localization in crowded scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2015)
20. Zhou, X.-G., Zhang, L.-Q.: Abnormal event detection using recurrent neural network. In: 2015 International Conference on Computer Science and Applications (CSA). IEEE (2015)
21. Feng, Y., Yuan, Y., Lu, X.: Deep representation for abnormal event detection in crowded scenes. In: Proceedings of the 24th ACM International Conference on Multimedia. ACM (2016)
22. Xu, D., et al.: Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput. Vis. Image Underst. **156**, 117–127 (2017)
23. Chong, Y.S., Tay, Y.H.: Abnormal event detection in videos using spatiotemporal autoencoder. In: International Symposium on Neural Networks. Springer, Heidelberg (2017)
24. Yang, B., et al.: Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network. IEEE Trans. Cogn. Dev. Syst. **11**, 473–482 (2018)
25. Sabokrou, M., et al.: Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. Comput. Vis. Image Underst. **172**, 88–97 (2018)
26. Keçeli, A., Kaya, A.: Violent activity detection with transfer learning method. Electron. Lett. **53**(15), 1047–1048 (2017)
27. Wei, H., et al.: Crowd abnormal detection using two-stream Fully Convolutional Neural Networks. In: 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE (2018)
28. Yan, M., Jiang, X., Yuan, J.: 3D convolutional generative adversarial networks for detecting temporal irregularities in videos. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE (2018)
29. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE (2017)
30. Hanson, A., et al.: Bidirectional convolutional LSTM for the detection of violence in videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
31. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. **53**(1–2), 23–69 (2003)