International Journal of Theoretical and Applied Research, 2024, Vol. 3, No. 1, 328-343

International Journal of Theoretical and Applied Research (IJTAR)
ISSN: 2812-5878

Homepage: https://ijtar.journals.ekb.eg

Original article

ANOMALY DETECTION IN SURVEILLANCE VIDEOS: SURVEY

Esraa A. Mahareek [1]*, Eman K. ElSayed [2], Nahed M. ElDesouky [3], Kamal A. ElDahshan [4]

[123]Mathematics department Faculty of Science Al-Azhar University (Girls Branch), Cairo, Egypt
[2] School of Computer Science, Canadian International College CIC, Cairo, Egypt
[4]Mathematics department Faculty of science Al-Azhar University, Cairo, Egypt

ARTICLE INFO

Received 16/09/2023
Revised 09/01/2024
Accepted 15/01/2024

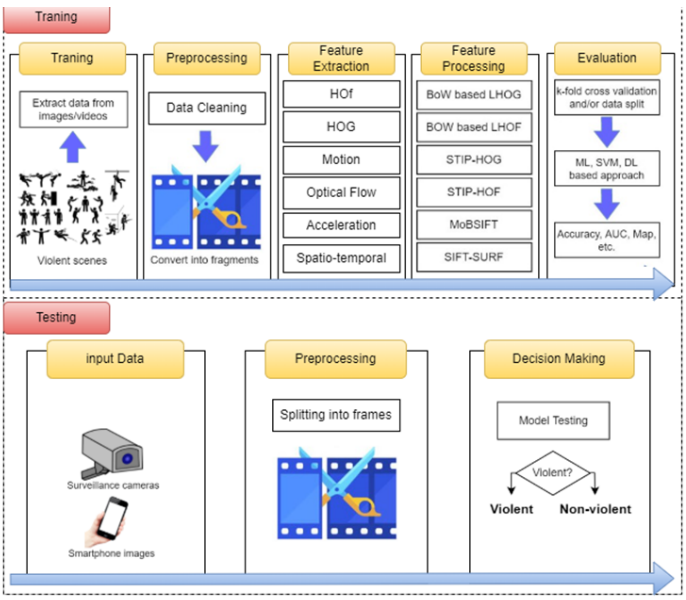
Keywords

Anomaly detection
Computer vision
Deep learning
Fight detection
Violence detection

ABSTRACT

The scientific community is paying more attention to the highly developed field of anomaly detection in video surveillance. Intelligent systems that can automatically spot unusual events in streaming videos are in high demand. This survey article gives a thorough summary of the several methods for spotting irregularities in surveillance videos. Both conventional methods—such as statistical modeling and motion analysis—and more current strategies—such as deep learning and artificial intelligence—are included in these methodologies. The study also identifies each technique's advantages and disadvantages as well as prospective uses in real-world situations. It also covers the difficulties in developing efficient anomaly detection algorithms for surveillance movies and points out potential future research topics. Overall, it is a useful tool for academics and professionals involved in the study of violent behavior detection (VioBD). It proposes a road map for future research on anomaly identification in surveillance films and provides insights into the state of the field now. To ensure the best possible performance of the anomaly detection system, it is crucial to keep in mind that the success of anomaly identification in surveillance videos significantly depends on the availability and quality of training data. As a result, future studies should concentrate on creating reliable feature extraction methods and enhancing the readability of anomaly detection models. The survey also says that in order for large-scale video data to be used in real-world applications that use anomaly detection systems, future studies should look into new ways to make these systems more scalable and effective.

Graphical abstract



⁎ *Corresponding author*
E-mail address: esraa.mahareek@azhar.edu.eg

DOI: 10.21608/IJTAR.2024.229847.1073

Special issue "selected papers from the 2nd International Conference on Basic and Applied Science (2nd IACBAS-2023)"

1. Introduction

Due to improvements in camera technology, storage capacity, and processing power, the field of video surveillance has grown significantly over the last few decades. These developments have made it possible to set up massive surveillance systems that can record, store, and process enormous volumes of video data. The vast amount of video data produced by these systems, however, presents a huge issue for human operators, who are unable to continuously examine all the material. The development of automated anomaly detection systems that can review surveillance footage and alert operators of potential security breaches or peculiar occurrences that require additional investigation has drawn increasing attention as a solution to this problem.

Finding anomalies in surveillance films requires developing algorithms that can identify behaviors or events that are out of the ordinary in a particular environment. This task can be particularly challenging given the complexity and variety of real-world scenarios as well as the requirement that algorithms operate in real-time without any lag or delay. Furthermore, the algorithms' capacity to distinguish between typical and abnormal occurrences must be extremely accurate if false positives or negatives are to be avoided.

Machine learning algorithms that can automatically learn from training data and recognize patterns of normal behavior are only one of the many methods that have been proposed to accomplish this. Utilizing deep learning strategies like convolutional neural networks (CNNs), which have shown promising outcomes in the identification of anomalies in surveillance films, is an additional strategy. But how effectively these algorithms' function depends a lot on the type and quantity of training data offered as well as the particular features used to characterize both normal and abnormal behavior. The following are this study's main contributions:

- Extensive Survey: The techniques utilized for anomaly identification in security footage are covered in detail in this research. It covers more modern tactics like deep learning and artificial intelligence in addition to more traditional ones like statistical modelling and motion analysis. This survey provides scholars and industry experts with a helpful resource by compiling and summarizing different approaches.
- Identification of Benefits and Drawbacks: The research lists the benefits and drawbacks of several anomaly detection methods for surveillance footage. Through a well-rounded examination of the advantages and disadvantages of every methodology, scholars and professionals can make knowledgeable choices when choosing and applying these techniques in practical settings.
- Prospective Uses in Real-World Circumstances: This paper explores the possible real-world uses for anomaly detection methods. Through elucidating the pragmatic applications of these techniques, the study offers discernments into their effective implementation to augment security and detect anomalies in surveillance recordings.
- Future Research Roadmap: Based on the work, a future research roadmap for anomaly identification in surveillance videos is proposed. It highlights issues and possible research directions that should be pursued in order to raise anomaly detection systems' efficacy and efficiency. This road map can direct upcoming studies and stimulate fresh approaches to the field's exploration.

Overall, the thorough review, benefits, and drawbacks analysis, potential applications in practical settings and research roadmap for anomaly detection in surveillance footage are the main contributions of this work.

This paper is set up as follows: Section 2 represents usual datasets in violence detection field. The approaches for segmentation and feature extraction and selection are briefly explained in Section 3. Section 4 explains the violence detection techniques. Finally discuss some of research challenges briefly.

1. Overview of surveillance videos datasets

To learn and understand activity recognition, anomaly detection has been thoroughly explored in a variety of domains, including computer vision. Due to the intricacy of real-world events, this work might become exceedingly difficult. It is hard to compile all abnormal incidents because there are many of them. Thankfully, a variety of datasets have been created to aid scientists and researchers in this effort. These datasets are sensitive to occlusion and variations in lighting because they were primarily collected in the visible spectrum. In this part, we provide a quick overview of various well-known datasets that are currently being used by academics to identify behavioral anomalies. The datasets are arranged chronologically, going from the oldest to the newest. We list the dataset's release year, kind (single-scene or multi-scene), sensor information (RGB or thermal, resolution, FPS), description of aberrant activity, and sample photographs for each dataset.

- **Subway dataset**

Adam et al. first mentioned the tube dataset in their 2008 publication [1]. Another kind of single-scene dataset is one like this. Figure 1 shows two extensive video recordings from this dataset of people being observed at a tube entry and departure. No spatial ground truth is available. The video comprises a total of 125,475 frames and was captured in grayscale format at 15 frames per second with a resolution of 512 x 384. The main anomalous events are a caretaker cleaning the walls, incorrect turns, loitering, no payment, persons jumping or squeezing through turnstiles, and loitering.

Figure 1: **S**hows a selection of photos from the Subway Dataset.

- **UMN Crowd Abnormality Dataset**

The UMN dataset was first introduced in the 2009 study by R. Mehran.et.al[2]. The video scenario represented a populous region where actors were seen walking about and acting strangely as they fled. Because it was captured in many locations, this dataset can be regarded as a multi-scene dataset.

A total of 11 small videos from the dataset are combined into one 4min 17s long video with 7739frames. The brief videos start out with normal behavior before switching to bizarre. There is only one interior setting and two outside scenarios. The videos were all shot with a static camera at a resolution of 640×480 and a frame rate of 30FPS. A temporal annotation represents the true situation. Several examples of the video's pictures are shown in Figure 2.

- **Anomalous Behavior Dataset**

In 2010, York University published The Anomalous Behavior Dataset [3].

The dataset includes eight multi-scene recordings that were captured under a variety of difficult circumstances, including lighting effects, scene clutter, changeable target appearance, quick motion, and camera jitter.

Figure 2: Shows a selection of photos from the UMN dataset[2].

Along with algorithms for identifying anomalous events in specific regions of each video, a spatiotemporal ground truth was also made available to the dataset. Figure 3 shows the image sequences in this collection primarily highlight the activities of people and vehicles in certain common places, such as a railway, river, sea, and airport.

Figure 3: **S**hows a selection of photos from the Anomalous Behavior Dataset

• Traffic-rail: This video shows routine rail operations. Given the significant changes in illumination and camera jitter, this film presents many difficulties. The film consists of 19,218 RGB image frames with a frame rate of 25 FPS and a resolution of 288 x 386. The movement of a passenger is an unusual event.

• Belleview: This clip shows automobiles navigating a junction. There are 2918 total frames in the grayscale video, which is captured at a resolution of 320 x 240 at a frame rate of 10 FPS. Cars entering the highway from the left or right constitute an abnormal incident.

• Boat-Sea: In this video, a passing boat is referred to as an unusual occurrence. The video comprises 450 total frames and was captured in RGB format at 720x576 resolution and 19 frames per second.

• Boat-River: In this film, a boat crossing a river is depicted as an unusual occurrence. The movie contains 250 total frames and was captured in RGB format at a resolution of 720x576 with a frame rate of 5FPS.

• Canoe: In this video, a canoe crossing a river is described as an unusual occurrence. The movie contains 1050 total frames and was captured in RGB format at 320x240 resolution and 30 frames per second.

• Camouflage: In this video, a person wearing camouflage is shown moving. The proper motion is taught as the norm, while the incorrect motion is the aberrant behavior. The movie has a total of 1629 frames and was captured in RGB format at 320×240 resolution and 30 frames per second.

• Airport-WrongDir: In this video, a queue of travelers is seen moving through an airport. The movie has a total of 2200 frames and was captured in RGB format at a resolution of 300x300 at a frame rate of 25FPS. People traveling in the wrong way is a strange occurrence.

- **Avenue Dataset**

The 37 movies in this dataset [4], which was published in 2013, are split into 16 typical videos for training and 21 atypical videos for testing. RGB single-scene is the dataset's type. 47 anomalous occurrences in all, divided into three primary categories: odd behavior, odd object, and improper direction. These recordings, which comprise 30,652 frames (15,328 for training and 15,324 for testing), were shot on the CUHK campus avenue. Each image series has a 640 x 360 resolution and a 25 FPS frame rate. Both temporal and spatial annotations were given by the author. Below are three examples of typical anomalous events in Figure 4:

1.Strange behaviors: running, tossing things, and lingering are examples.
2. Wrong direction: individuals travelling against the flow.
3. Abnormal items: individuals walking about with odd objects like bicycles.

Figure 4: Shows a selection of photos from the Avenue Dataset.

- **UCSD Anomaly Detection Dataset**

Two sub datasets, Pedestrian 1 and Pedestrian 2, are part of this dataset [5], which was published in 2013. Both have a 10 FPS, 238 x 158 grayscale image sequence for pedestrian number one and 360 x 240 for pedestrian number two. Each dataset consists of only one scenario. Both have testing recordings with anomalous events and training videos with only normal behaviors. The dataset was captured using a static camera that was mounted on a lift and looked down at the paths for pedestrians as Figure 5. The unusual occurrences consist of the following:

- Unusual pedestrian motion patterns, such as people strolling across a sidewalk or in the grass surrounding it.
- The movement of non-pedestrian entities such as bikers, skaters, and small carts on the walkways.

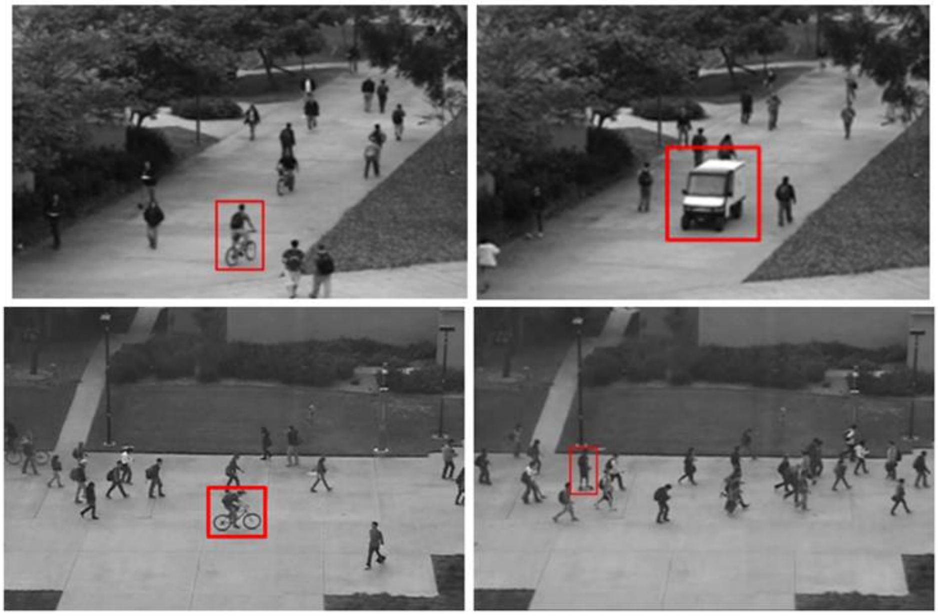


Figure 5: Shows a selection of photos from the UCSD Anomaly Detection dataset.

- **ShanghaiTech Campus Dataset**

In 2016, this dataset [6] was made available. It has 107 test movies with 130 anomalous occurrences and 330 training videos with only normal events. There are 317,398 total frames, and 17,090 of them are irregular. An RGB camera with a resolution of 856×480 at 24FPS was used to gather the dataset while looking over pedestrian paths. It is made up of 13 scenes (multi-scenes), each having its own complicated lighting, camera angles, and anomaly types, most of which are connected to odd things, peculiar directions, and odd activities as Figure 6.

- Weird behaviors include fleeing, robbing, pushing, jumping over fences, dropping things, hurling things, and fighting. Here are some examples of pictures.
- Wrong way: There are times when someone goes in the opposite direction of what they generally do.
- Abnormal objects: In this scenario, a person is seen carrying an odd object, like a bicycle or pram for a child.

- Weird behaviors include fleeing, robbing, pushing, jumping over fences, dropping things, hurling things, and fighting. Here are some examples of pictures.
- Wrong way: There are times when someone goes in the opposite direction of what they generally do.
- Abnormal objects: In this scenario, a person is seen carrying an odd object, like a bicycle or pram for a child.

Figure 6: Shows a selection of photos from the ShanghaiTech dataset.

- **UCF-Crime Dataset**

In 2018 [7], the UCF-Crime dataset was made public. This dataset is a collection of 128 hours' worth of 1900 online films that were shot with numerous RGB cameras in various locations (multi-scene). Abuse, arrest, arson, assault, car accident, burglary, explosion, fight, robbery, shooting, stealing, shoplifting, and vandalism are a few examples of unusual occurrences. The event recognition of 13 group activities and anomaly detection in each individual group are the two main objectives that these movies, which represent 13 real-world events, can be utilized for. Only temporal annotations were provided by the authors. Some sample photographs from the dataset are shown in Figure 7.

Figure 7: **S**hows a selection of photos from the UCF-Crime Dataset.

2. Preprocessing

The benchmark datasets that have been discussed are merely compilations of raw video and image data. Prior to being fed into the ML algorithms, these data must be preprocessed. The performance of a video surveillance system is significantly impacted by the procedures of feature engineering and data preparation. These procedures involve a number of steps for the vision-based domain, such as feature representation extraction, foreground extraction, and backdrop creation. Reduced noise, important representation feature selection, high-dimension feature transformation into the sub-space domain without losing vital information, and reduced overfitting problems are the primary goals of the feature engineering process. There are many obstacles, though, like shifting lighting, complex backgrounds, occlusion, or fictitious interactions between the subjects.

- **Segmentation**

Segmentation is the first step in the preparation of data. The target subjects are extracted from pictures or videos using segmentation. Techniques for foreground extraction and background construction are included in segmentation. backdrop creation algorithms strive to identify a scene's overall representative qualities and the subjects they identify are subsequently analyzed depending on how the current frame differs from the produced backdrop [8]. When employed with fixed cameras, background-building techniques are extremely effective for tracking moving items in an image.

In the multi-modal domain with adaptive parameters, several statistical approaches of background generation, such as [9][10], can perform well; nevertheless, their effectiveness is significantly hampered by ambient noise or bad lighting circumstances. The neural network-based methods described in [8][11][12] can get beyond these restrictions, but they tend to overfit the data. The aim of the application heavily influences the technique selected. A foreground extraction-based segmentation technique is required for dynamic backdrop recording with moving cameras. To separate target subjects from the backdrop of a video sequence, both spatial and temporal data are analyzed. The authors in [13][14] effectively recovered the targets from the video recording by moving the camera while dealing with occlusion and distortion using an optical flow technique; nevertheless, their methods are intricate and time-consuming. Foreground segmentation, which was susceptible to noise but needed little computational resources, was also carried out using temporal information in [15]. Markov random fields were employed in references [16] to maintain the borders and manage complex backgrounds, but these techniques are not computer-efficient.

Both background and foreground information have been attempted to be used by several of the proposed approaches for video anomaly identification. Two decoders were introduced by Lai et al. [17] in their network that produced a future frame and RGB difference. By identifying objects in a movie, Doshi et al. [18] employed a pretrained object detection model (YOLO) to capture location and appearance data. Contrarily, Cai et al. [19] employed an optical flow clip and an image clip as input to record structure and motion data. The fused feature was then taken from two encoders and used by two decoders to create a future frame and optical flow image.

- **Feature Extraction and Selection**

The right properties for comprehending human behaviors were extracted using a handcrafted feature-based extraction method [20]. These techniques can only be used under specific circumstances and are not adaptable to different situations. They take a lot of time and are ineffective on computers. New representative characteristics, which may be divided into local, global, and semantic features, were used by researchers. These attributes have amply demonstrated their benefits and resistance to noise and dynamic settings.

• Utilizing local descriptor techniques, local representative features regulate the local quantification of an image's input region. They describe each section in an image separately, considering their specific location. A HOG (histogram of gradients) is a fundamental method for extracting a local description of the gradient magnitude and direction of picture, claim [20]. Although a HOG is only useful for human detection at fixed sizes, it is unaffected by changes in photometry. However, it contains high-dimensional features, which are wasteful for calculating and unsuitable for real-time applications. The Scale-Invariant Feature Transform (SIFT) was used in [21] and demonstrated invariance to geometric and photometric translation, even with 3D projection. The speed-up robust feature (SURF) method [22] is faster than SIFT while maintaining the excellent quality of the detection points. Finally, by preserving the edge structures of the target subjects, the [23] shape-based local feature descriptor demonstrated that it was noise-resistant. But silhouette segmentation is the major technique used in the preprocessing stage.

• In order to regulate the global quantization of an input image and create a feature vector that reflects the contents of the image abstractly, global representative features need an image descriptor. In [24], The crucial details of the corners, edges, ridges, and optical flow were meticulously captured by the global descriptor. The camera depth can be used to readily retrieve these characteristics; however, they are scene-specific and lack general information. In [25], some researchers exploited 3D-space-time volume to extract background-independent 3D global feature vectors. These 3D characteristics were, however, extremely susceptible to occlusion and noise. The discrete Fourier transform (DFT) was also used in [26] to convert spatial features into frequency features, despite the possibility that the inverse procedure may lose the spatial and temporal information necessary to identify the anomalous target subjects.

The process of extracting, identifying, and matching a feature or object from a single video frame is known as Feature extraction (FE). This explains the Feature extraction and encoding methods used in Anomaly detection in surveillance videos. Chandrakar et al. [27] recommended a video anomaly detection (VAD) centered on Gaussian Process Regression (GPR) in addition to localization and presentation of hierarchical features. An interaction codebook was created and modelled by GPR. A method for estimating the likelihood of an observed interaction was devised using inference. The global steady anomaly masks and these local likelihood scores were combined. As a result, anomalies might then be quickly discovered. GPR was used to model the link between the nearby STIP for Anomaly detection. This method, which required less computing, outperformed the best ones. However, this strategy did not address complex causality.

For the FE of VAD, Roberto Leyva et al. [28] constructed a Gaussian Mixture Model. An inference process that approximated the compact feature set using the Gaussian Mixture Model, Markov Chains, and Bag-of-Words was used to look for unusual events. The joint reactions of the models to the local spatiotemporal neighborhood were also taken into consideration to improve detection accuracy. Popular current datasets were collected, which included a wide range of real-world videos that were recorded using security cameras. Other online methods outperformed them. The framework acquired a competitive detection performance when measured against the best non-online approaches. However, lengthy processing durations were required.

A particle filter-centered approach for feature series retrieved from videos was devised by Xinwen Gao et al. [29]. The entire procedure included particle filter tracking in addition to feature series development. An L2-norm extractor was modelled with an optical flow as its focal point to represent the features of the video. The particle filter then maintained these feature series on course. The occurrence of anomalous occurrences may be the source of the shift in the features series as well as a larger mistake on the PF tracking. This made it possible for the computers to understand and describe the instances of abnormalities. considered as the UMN dataset. The algorithm's frame-level detection accuracy was 90%. However, when compared to deep neural networks (NN), the particle filter's generalization capacity was substantially lower in terms of data fitting.

A fully interconnected Convolution 3D(C3D) was created by Muhammad Zaigham Zaheer et al. [30] with the goal of FE of anomalous films. The self-reasoning-centered training was carried out using pseudo labels created by binary clustering of spatiotemporal video characteristics. It helped to reduce the noise that was displayed on the labels of anomalous videos. To achieve the goal of more accurate AD, the primary network and clustering were urged to function in tandem. We used freely available real-world AD datasets from UCF-crime, ShanghaiTech, UCSD, Ped2, and others. The framework's advantage over the most effective methods already in use was demonstrated. For larger fragments, the processing time was increased.

To reduce the FE in surveillance footage, JunY. Lim et al. [31] constructed a deep multiple-level feature network. For depiction learning, a dataset was built in a setting with no rules and handguns. For creating a dataset, a collection of 250 recorded videos and more than 2500 unique labeled frames was considered. According to the comparative receptive field, the backbone's shallow, medium, and deep features were combined to improve the base features. Focal Loss was added as its categorization loss to help with the detection of smaller weapons. The accuracy of the suggested model was 87.42%.

3. Techniques for Detecting Anomaly in Surveillance Videos

Violence is often understood to be unexplained events or behaviors. Utilizing computer vision to recognize these actions in security cameras is a common issue in the realm of action detection [32]. Scientists have discussed numerous strategies and techniques for spotting violent or unexpected events, stressing the necessity for more effective identification as seen by the sharp increase in crime rates.

Over the past few years, numerous techniques for detecting violence have been created. Machine learning and deep learning are two of the three categories of violence detection methods, and they are separated according to the classifier that is employed. The main objective of a violence detection system is to identify events as they take place so that potentially harmful situations can be avoided. Nevertheless, it is crucial to understand a few fundamental ideas. The basic steps of video-based violence detection systems are shown in Figure 8.

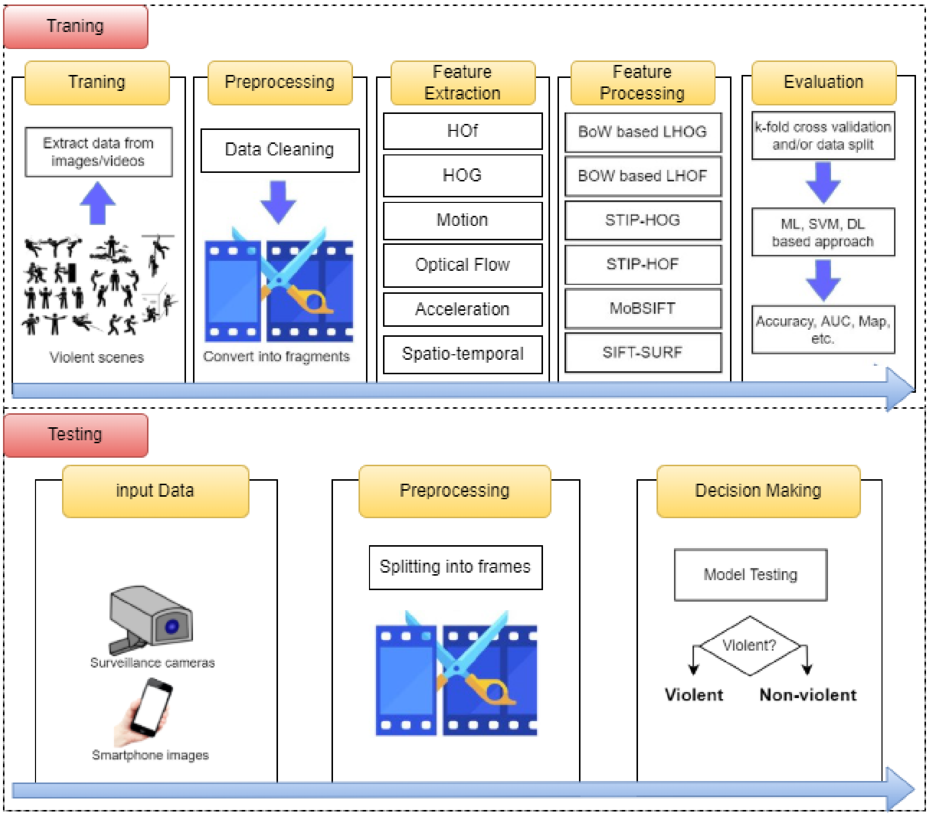


Figure 8: Basic phases of violence detection in surveillance videos [32].

A K-means algorithm was developed by Sahu et al. [33] with the goal of effective crowd AD. Many DL-centered as well as manually crafted feature-centered approaches were outperformed in terms of accuracy. The presentation of a low-power FPGA implementation. Non-overlapping pixels were used in the feature extraction process. This made it possible to gate inputs to several modules, resulting in high power efficacy. The maximum amount of energy per pixel required was 2.43nJ, and 126.65 M-pixels could be processed in a second. Outliers and noisy data weren't dealt with.

A single-class Support Vectors Machine (SVM) was created by Claudio Piciarelli et al.[34] with the intention of detecting anomalous occurrences that are trajectory-centered. Trajectory analysis, a method with several applications—most notably VS along with traffic monitoring—was used to specifically solve AD. The approach was centered around SVM clustering. The detection SVM capabilities were used to identify anomalous trajectories.

Trajectory categorization required particular attention because there was no prior knowledge of how the outliers were distributed. The efficacy of the method was established. However, it did not demonstrate that the system had strong generalization capabilities when classifying previously unknown patterns.

Xinfeng Zhang et al. did a K-Nearest Neighbor (KNN) analysis with a focus on localization and VAD. A method for detecting anomalies spanning time and space was presented by examining whether the similarities between the testing samples and the recovered K-NN samples followed the pattern distributions of homogeneous intra-class similarities. It was '1'-class learning that was unsupervised and didn't require prior knowledge or clustering. Because probability was employed to make the assessment and the computation of probability was unaffected by motion distortions brought on by perspective distortion, this system had an advantage over current ones in that it could adapt to the entire picture. However, because this system required tailored training for various settings, the exceedingly sluggish training tempo limited its full implementation.

Sondos Fadl et al. [35] suggested using a Gaussian RBF multiple class SVM (RBF-MSVM) to identify fake surveillance videos. A technique for detecting interframe forgeries was created using a 2DCNN of spatiotemporal information and feature extraction fusion. Duplication, insertion, and frame deletion were all present. It was demonstrated that the technique was effective in finding every interframe fake. However, the lack of motion modeling made it unsuitable for usage in videos.

Shaoci Xie et al.[36] suggested using machine learning to identify aberrant behavior models and detect crowds in video footage. The objective was to learn divergent data mining while also attempting to improve detection accuracy. It was decided to render an IDSAD model in the absence of user behavior AD. It was made easier to visualize user activity patterns and behavior profiles. The job of similarity was used. The results of an experiment using data from UNIX user shell commands showed that the detection architecture offered better detection performances. Estimating the ideal sequence length for certain users was challenging.

We examine the DL for actual anomaly detection on surveillance films in this section. Numerous techniques for classifying video activities have been developed as a result of the successful demonstration of DL for image classification [37]. Table 1 discusses the deep learning methods for anemology detection.

Table 1: List of deep-learning classifiers for anomaly detection in CCTV.

Ref. &year	Techniques	Datasets	Results	disadvantages
[38] 2020	IBaged-FCNet	UCF-Crime-dataset	AUC:92.06%	Significantly slower than other systems.
[39] 2019	Deep 3-dimensional-Convolutional-network(C3D)	UCSD and Avenue,Subway.	AUC:82.1%	Poor knowledge about temporal order information in the training videos.
[40] 2019	Convolution-Long-shorts-Term-memory(CLSTM)	Ped2	AUC:96.5%	Not appropriate for detecting non-obvious anomalies.
[41] 2020	CNN	PETS2009, and UMN	Acc:98.39%	High computational complexity.
[42] 2020	Deep-Spatiotemporal-Translations-Network(DSTN)	UMN,CUHK ,Avenue,UCSD, pedestrians	AUC:83.1%	Degrade the performance for complex scene detection.
[43] 2018	Fuzzy-theory	Publicly available dataset	Acc:93.4%	Increase the false rate for complex situations.
[44] 2018	Generative-NNs	UCSD,Avenue, UMN,additionto-PETS	Acc:98.8%	The model parameters oscillate, destabilize in addition to never converge.
[45] 2016	Adaptive Intra-frame Classifications Network	UCSDPed1 datasets	AUC:95.1%	Manual segmentation of sub regions.

Ref. &year	Techniques	Datasets	Results	disadvantages
[46] 2020	CNNs	publically accessible data set	Acc:95%	Reduce the system accuracy for camera redirecting detection.
[47] 2021	CNN	COCO	Acc:78.9%	Increase the processing time for complex features.
[48] 2021	Reinforcements Learning Model	UCSD	Acc:85.30%	Degrade the performance for complex features.
[49] 2021	MRCNN	UCSDped2	Acc:94%	Trained with particular real-time scenarios.
[50] 2020	Multi-layer perception (MLP) recurrent NN(RNN)	Real-world-dataset	Acc: 98.30%	Less efficiency.
[51] 2021	3DCNN	UCSD,Ped2, Shang-haiTechs, and Avenue	AUC:96.9%	Trained with small datasets.
[52] 2020	Fully Convolutional Net-work(FCN)	UCSD and Avenue	Acc:94.9%	Intricate conditions like the changing scenes have difficulties to study the proper appraisal of the distribution parameters due to lack of enormous similar nor-mal events.
[53] 2021	Cognitive-memory-augmented-network (CMAN)	Ped 2	96.2%	Trained with small datasets.
[54] 2021	Single and multi-frame-anomaly-detection	Ped 2,Avenue	97.5%-87%	Not appropriate for detecting non-obvious anomalies.
[55] 2021	Multi-Level-Memory-modules in an Autoencod-er with-Skip-Connections(ML-MemAE-SC)	Ped 2,Avenue,ShanghaiTech	99.3%,91.1%, 76.2%	The model parameters oscillate, destabilize in addition to never con-verge
[56] 2021	Autoencoder with a MemoryModule (AMM)	Ped 2,Avenue,ShanghaiTech	97.2%,87.9% ,70.2%	Significantly slower than other sys-tems.
[57] 2021	Explanation for Anom-alyDetection	Pad1,Ped 2	73%,80%	Reduce the system accuracy for camera redirecting detection.
[58] 2022	Attention-based adversari-al-autoencoder (A3N)	Pad1,Ped 2,Avenue,ShanghaiTech	90.7%,97.7% 89.4%,86.9%	Poor knowledge about temporal order information in the training videos.
[59] 2022	GroupActivities for AD	Pad1,Ped 2,Avenue	84.4%,95%, 82.3%	High computational complexity.
[60] 2022	Variationa-lAnomaly-Detection-Network (VADNet)	Ped 2,Avenue,ShanghaiTech	96.8%,87%, 75%	Less efficiency.
[61] 2022	Context-related video anomaly-detection	Ped 2,Avenue,ShanghaiTech	96.3%,87.1% ,73.6%	Manual segmentation of sub re-gions.
[62] 2022	Localization-based-Reconstruction(LBR)	Pad1, Ped 2,Avenue,ShanghaiTech	81%,97.2% 92.8%,72%	The model parameters oscillate, destabilize in addition to never con-verge
[63] 2022	Foreground–Background SeparationMutual GenerativeAdversarial Network(FSM-GAN)	Ped 2,Avenue,ShanghaiTech	98.1%,80.1% ,73.5%	Trained with particular real-time scenarios.
[64] 2023	Dual-stream-memory-network	Ped 2,Avenue,ShanghaiTech	98.3%,88.6% ,75.7%	Increase the processing time for complex features
[65] 2023	Attention-based residual-autoencoder	Ped 2,Avenue,ShanghaiTech	97.4%,86.7% ,73.6%	Less efficiency.

Ref. &year	Techniques	Datasets	Results	disadvantages
[66] 2023	Bi-directional-Frame Interpolation	Ped 2,Avenue,ShanghaiTech	98.9%,89.7% ,75%	Not appropriate for detecting non-obvious anomalies.
[67] 2023	Zero-shot-Cross-domain-VideoAnomaly Detection(zxVAD)	Pad1, Ped 2,Avenue	78.6%, 95.8%,83.2%	Increase the processing time for complex features

4. Research Challenges

The following list of key challenges is based on the review of the studies mentioned above:

1. Lack of better datasets: Due to the uneven distribution of normal and abnormal data behavior, there are very few publicly available benchmarked datasets for anomaly identification.
2. Environmental difficulties: Difficulties in the environment, such as background noise, occlusions, and illuminations, might make it difficult to detect anomalies accurately.
3. High computational space and temporal complexity characterize the majority of the known algorithms. Therefore, creating a straightforward, effective, and precise method remains a difficult task.
4. Dynamic behavior in anomalies: Since anomalous events are uncommon and their behavior is typically sparse, no single technique can be used to identify every kind of anomaly.
5. Atmospheric Turbulences: The visuals in the film are blurred by fluctuations in the refraction and reflection of light, smoke, and fog, which are common atmospheric problems.

5. Research Gaps

The following are a few possible research gaps that this study could look into further: Limited Assessment of Real-World Performance: Although the study covers the benefits and drawbacks of several anomaly detection methods, a thorough assessment of their effectiveness in actual use may be lacking. Subsequent investigations may concentrate on carrying out tests and analyses utilizing authentic surveillance video collections in order to appraise the efficiency and resilience of these techniques in real-world situations.

Absence of Standardized Benchmarks: Evaluation metrics and benchmarks are not standardized in the field of anomaly identification in surveillance films. This makes it difficult to fairly compare and benchmark various methods. Subsequent investigations may focus on developing uniform standards and assessment procedures to enable impartial comparisons and foster progress in the domain.

Privacy and Ethical Issues: The paper makes a passing reference to the necessity of protecting the privacy and addressing ethical issues when implementing anomaly detection systems. However, more research can examine alternative privacy-preserving strategies, go deeper into the ethical implications of surveillance technologies, and look into ways to strike a balance between security and privacy when analyzing surveillance video.

Interpretability and Explainability of Algorithms: The study recognizes that there is a need to enhance anomaly detection algorithms' interpretability and explainability. Subsequent investigations may concentrate on creating comprehensible models and methodologies that offer perspectives into the logic underlying anomaly identification. This can promote human comprehension and decision-making while also strengthening systemic trust.

Scalability and Effectiveness: Improving the surveillance video anomaly detection systems' scalability and efficacy is touched upon in passing in the paper. With the volume and complexity of surveillance video data growing, there is a need for more studies to determine how to make these algorithms more scalable. Furthermore, studies can concentrate on enhancing computational effectiveness to facilitate real-time anomaly identification in massive video streams.

Future research can improve anomaly identification in surveillance footage and offer workable solutions to improve security and public safety by filling in these research gaps.

6. Conclusion

The necessity for systems that automatically recognize violent incidents grows in direct proportion to the expansion of surveillance cameras to monitor human activity in all spheres of life. The identification of violent activity has emerged as a hot topic in computer vision, drawing new researchers. For identifying these activities in movies, numerous scholars have proposed numerous techniques. Examining the most recent research in violence detection is the main goal of this systematic review. In this study, various feature extraction, machine learning, and deep learning-based video violence detection methods were investigated. We started by examining the most popular methods for extracting and characterizing features. Additionally, all datasets and video attributes were used across all methodologies, and those that are essential to identification are thoroughly recorded in tables. The precision of the feature extraction, object recognition, and classification methods are all crucial factors. Then, we provide a thorough analysis of the terms used to describe violence. Finally, we talked about challenges for violence detection in video. Our study could provide insight on techniques and methods for spotting violent conduct in surveillance videos.

References

1. 1. A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE . 30(3) (2008) 555–560 . doi: 10.1109/TPAMI.2007.70825.
2. R. Mehran, A. Oyama, and M. Shah, Abnormal crowd behavior detection using social force model. IEEE Conf. Comput. Vis. Pattern Recognition. 2 (2009) 935–942. doi: 10.1109/CVPRW.2009.5206641.
3. A. Zaharescu and R. Wildes, Anomalous behaviour detection using spatiotemporal oriented energies. subset inclusion histogram comparison and event-driven processing. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6311 (PART 1) (2010) 563–576. doi: 10.1007/978-3-642-15549-9_41.
4. C. Lu, J. Shi, and J. Jia, Abnormal event detection at 150 FPS in MATLAB. Proc. IEEE Int. Conf. Comput. Vis.(2013) 2720–2727. doi: 10.1109/ICCV.2013.338.
5. V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, Anomaly detection in crowded scenes. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2010) 1975–1981. doi: 10.1109/CVPR.2010.5539872.
6. W. Liu, W. Luo, D. Lian, and S. Gao, Future Frame Prediction for Anomaly Detection - A New Baseline. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.(2018) 6536–6545. doi: 10.1109/CVPR.2018.00684.
7. W. Sultani, C. Chen, and M. Shah, Real-world Anomaly Detection in Surveillance Videos. (2018). Available: http://arxiv.org/abs/1801.04264
8. M. Babaee, D. T. Dinh, and G. Rigoll, A deep convolutional neural network for video sequence background subtraction. Pattern Recognit. 76 (2018) 635–649. doi: 10.1016/j.patcog.2017.09.040.
9. B. N. Subudhi, S. Ghosh, S. C. K. Shiu, and A. Ghosh, Statistical feature bag based background subtraction for local change detection. Information Sciences. 366 (2016) 31–47. doi: 10.1016/j.ins.2016.04.049.
10. S. Jiang and X. Lu, WeSamBE: A Weight-Sample-Based Method for Background Subtraction. IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2105–2115. doi: 10.1109/TCSVT.2017.2711659.
11. D. Sakkos, H. Liu, J. Han, and L. Shao, End-to-end video background subtraction with 3d convolutional neural networks, Multimedia Tools and Applications .77(17) (2018) 23023–23041. doi: 10.1007/s11042-017-5460-9.
12. T. Minematsu, A. Shimada, H. Uchiyama, and R. ichiro Taniguchi, Analytics of deep neural network-based background subtraction. J. Imaging. 4(6) (2018). doi: 10.3390/jimaging4060078.
13. A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. De Albuquerque, Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM. IEEE Trans. Ind. Electron . 66(12) (2019) 9692–9702. doi: 10.1109/TIE.2018.2881943.
14. A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z. Mahfouf, Human activity recognition via optical flow: decomposing activities into basic actions. Neural Comput. Appl. 32(21) (2020) 16387–16400. doi: 10.1007/s00521-018-3951-x.
15. H. Mliki, F. Bouhlel, and M. Hammami, Human activity recognition from UAV-captured video sequences. Pattern Recognit. 100 (2020) 107-140. doi: 10.1016/j.patcog.2019.107140.
16. L. Liu, S. Wang, Y. Peng, Z. Huang, M. Liu, and B. Hu, Mining intricate temporal rules for recognizing complex activities of daily living under uncertainty. Pattern Recognit. 60 (2016) 1015–1028. doi: 10.1016/j.patcog.2016.07.024.
17. Y. Lai, Y., & Han, Anomaly Detection with Prototype-Guided Discriminative Latent Embeddings. IEEE Int. Conf. Data Min. (2020) 300–309.
18. K. Doshi and Y. Yilmaz, Any-shot sequential anomaly detection in surveillance videos. IEEE Comput. Soc. Conf. Computer Vision and Pattern

Recognition. Work. June (2020) 4037–4042. doi: 10.1109/CVPRW50498.2020.00475.
19. R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, Appearance-Motion Memory Consistency Network for Video Anomaly Detection. 35th AAAI Conf. Artif. Intell. 2A (2021) 938–946. doi: 10.1609/aaai.v35i2.16177.
20. K. P. Sanal Kumar and R. Bhavani, Human activity recognition in egocentric video using HOG, GiST and color features. Multimed. Tools Appl. 79(5–6) (2020) 3543–3559. doi: 10.1007/s11042-018-6034-1.
21. J. B and C. M Patil, Video Based Human Activity Detection, Recognition and Classification of actions using SVM. Trans. Mach. Learn. Artif. Intell. 6(6) (2018), doi: 10.14738/tmlai.66.5287.
22. D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, Vision-based human activity recognition: a survey. Multimed. Tools Appl. 79(41–42) (2020) 30509–30555. doi: 10.1007/s11042-020-09004-3.
23. F. Patrona, A. Chatzitofis, D. Zarpalas, and P. Daras, Motion analysis: Action detection, recognition and evaluation based on motion capture data. Pattern Recognit. 76 (2018) 612–622. doi: 10.1016/j.patcog.2017.12.007.
24. K. Kim, A. Jalal, and M. Mahmood, Vision-Based Human Activity Recognition System Using Depth Silhouettes: A Smart Home System for Monitoring the Residents. J. Electr. Eng. Technol. 14(6) (2019) 2567–2573. doi: 10.1007/s42835-019-00278-8.
25. F. Baumann, A. Ehlers, B. Rosenhahn, and J. Liao, Recognizing human actions using novel space-time volume binary patterns. Neurocomputing. 173 (2016) 54–63. doi: 10.1016/j.neucom.2015.03.097.
26. M. S. Seyfioğlu, A. M. Özbayoğlu, and S. Z. Gürbüz, Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. IEEE Trans. Aerosp. Electron. Syst. 54 (4)(2018) 1709–1723. doi: 10.1109/TAES.2018.2799758.
27. U. S. R Chandrakar, R Raja, R Miri, RK Patra, Computer succoured vaticination of multi-object detection and histogram enhancement in low vision. Int. J. Biom. 15(3–4) (2023) 255–271.
28. L. C.-T. Leyva R, Sanchez V, Video anomaly detection with compact feature sets for online performance. J. Latex Cl. Files, 14(8) (2015) 1–16.
29. X. Gao, G. Xu, S. Li, Y. Wu, E. Dancigs, and J. Du, Particle Filter-Based Prediction for Anomaly Detection in Automatic Surveillance. IEEE Access, 7 (2019) 107550–107559. doi: 10.1109/ACCESS.2019.2931820.
30. M. Z. Zaheer, A. Mahmood, H. Shin, and S. Lee, A Self-Reasoning Framework for Anomaly Detection Using Video-Level Labels. IEEE Signal Process Lett. 27(July) (2020) 1705–1709.
31. J. Y. Lim, M. I. Al Jobayer, V. M. Baskaran, J. M. Y. Lim, J. See, and K. S. Wong, Deep multi-level feature pyramids: Application for non-canonical firearm detection in video surveillance. Eng. Appl. Artif. Intell. 97(October) (2020) 104-119. doi: 10.1016/j.engappai.2020.104094.
32. B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, State-of-the-art violence detection techniques in video surveillance security systems: A systematic review. PeerJ Comput. Sci. 8 (2022) 1–41. doi: 10.7717/PEERJ-CS.920.
33. A. K. Sahu, S. Sharma, M. Tanveer, and R. Raja, Internet of Things attack detection using hybrid Deep Learning Model. Comput. Commun. 176(April) (2021) 146–154. doi: 10.1016/j.comcom.2021.05.024.
34. C. Piciarelli, C. Micheloni, and G. L. Foresti, Trajectory-based anomalous event detection. IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1544–1554. doi: 10.1109/TCSVT.2008.2005599.
35. S. Fadl, Q. Han, and Q. Li, CNN spatiotemporal features and fusion for surveillance video forgery detection. Signal Process. Image Commun. 90 (2021) 116066. doi: 10.1016/j.image.2020.116066.
36. S. Xie, X. Zhang, and J. Cai, Video crowd detection and abnormal behavior model detection based on machine learning method. Neural Comput. Appl. 31(2019) 175–184. doi: 10.1007/s00521-018-3692-x.
37. R. Raja, P. C. Sharma, M. R. Mahmood, and D. K. Saini, Analysis of anomaly detection in surveillance video: recent trends and future vision. Multimed. Tools Appl. 82(8) (2023) 12635–12651. doi: 10.1007/s11042-022-13954-1.
38. Y. Zahid, M. A. Tahir, N. M. Durrani, and A. Bouridane, IBaggedFCNet: An Ensemble Framework for Anomaly Detection in Surveillance Videos. IEEE Access, 8 (2020) 220620–220630. doi: 10.1109/ACCESS.2020.3042222.
39. W. Chu, H. Xue, C. Yao, and D. Cai, Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos, IEEE Trans. Multimed. 21(1)(2019) 246–255. doi: 10.1109/TMM.2018.2846411.
40. Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, Spatio-Temporal Unity Networking for Video Anomaly Detection. IEEE Access. 7(2019)172425–172432. doi: 10.1109/ACCESS.2019.2954540.
41. C. Direkoglu, Abnormal Crowd Behavior Detection Using Motion Information Images and Convolutional Neural Networks. IEEE Access. 8(2020) 80408–80416. doi: 10.1109/ACCESS.2020.2990355.
42. T. Ganokratanaa, S. Aramvith, and N. Sebe, Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network. IEEE Access. 8 (2020) 50312–50329. doi: 10.1109/ACCESS.2020.2979869.
43. Y. Li, T. Guo, R. Xia, and W. Xie, Road Traffic Anomaly Detection Based on Fuzzy Theory. IEEE Access. 6 (c) (2018) 40281–40288. doi: 10.1109/ACCESS.2018.2851747.
44. S. Wang, E. Zhu, J. Yin, and F. Porikli, Video anomaly detection and localization by local motion based joint video representation and OCELM. Neurocomputing. 277(2018) 161–175. doi:

10.1016/j.neucom.2016.08.156.
45. K. Xu, T. Sun, and X. Jiang, Video Anomaly Detection and Localization Based on an Adaptive Intra-Frame Classification Network. IEEE Trans. Multimed. 22(2) (2020) 394–406. doi: 10.1109/TMM.2019.2929931.
46. L. Dong, Y. Zhang, C. Wen, and H. Wu, Camera anomaly detection based on morphological analysis and deep learning. Int. Conf. Digit. Signal Process. DSP, 0(2016) 266–270. doi: 10.1109/ICDSP.2016.7868559.
47. Ragedhaksha, Darshini, Shahil, and J. Arunnehru, Deep learning-based real-world object detection and improved anomaly detection for surveillance videos. Mater. Today Proc. 80(8)(2021). doi: 10.1016/j.matpr.2021.07.064.
48. R. F. Mansour, J. Escorcia-Gutierrez, M. Gamarra, J. A. Villanueva, and N. Leal, Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model. Image Vis. Comput. 112(2021) 104229. doi: 10.1016/j.imavis.2021.104229.
49. L. Tiwari et al., Detection of lung nodule and cancer using novel Mask-3 FCM and TWEDLNN algorithms. Meas. J. Int. Meas. Confed. 172(October) (2020) 108882. doi: 10.1016/j.measurement.2020.108882.
50. M. Murugesan and S. Thilagamani, Efficient anomaly detection in surveillance videos based on multi layer perception recurrent neural network. Microprocess. Microsyst. 79(2020) 103303. doi: 10.1016/j.micpro.2020.103303.
51. Y. Hao, J. Li, N. Wang, X. Wang, and X. Gao, Spatiotemporal consistency-enhanced network for video anomaly detection. Pattern Recognit. 121(2022) 108232 doi: 10.1016/j.patcog.2021.108232.
52. Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder. Comput. Vis. Image Underst. 195(2020). doi: 10.1016/j.cviu.2020.102920.
53. T. Wang, X. Xu, F. Shen, and Y. Yang, A Cognitive Memory-Augmented Network for Visual Anomaly Detection. IEEE/CAA J. Autom. Sin. 8(7) (2021) 1296–1307. doi: 10.1109/JAS.2021.1004045.
54. M. Bahrami, M. Pourahmadi, A. Vafaei, and M. R. Shayesteh, A comparative study between single and multi-frame anomaly detection and localization in recorded video streams. J. Vis. Commun. Image Represent. 79(February) (2021) 103232. doi: 10.1016/j.jvcir.2021.103232.
55. Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. Proc. IEEE Int. Conf. Comput. Vis. (2021)13568–13577. doi: 10.1109/ICCV48922.2021.01333.
56. J. Tang, Wei and Feng, Yunjian and Li, An Autoencoder with a Memory Module for Video Anomaly Detectiontle. 36th Youth Acad. Annu. Conf. Chinese Assoc. Autom.(2021) 473–478. doi: 10.1109/YAC53711.2021.9486538.
57. C. Wu, S. Shao, C. Tunc, P. Satam, and S. Hariri, An explainable and efficient deep learning framework for video anomaly detection. Cluster Comput. 25(4) (2022) 2715–2737. doi: 10.1007/s10586-021-03439-5.
58. M. H. Aslam, N.; Rai, P.K.; Kolekar, A3N: Attention-based adversarial autoencoder network for detecting anomalies in video sequence. J. Vis. Commun. Image Represent. 87(C) (2022). doi: https://doi.org/10.1016/j.jvcir.2022.103598.
59. L. F. Borja-Borja, J. Azorín-López, M. Saval-Calvo, A. Fuster-Guilló, and M. Sebban, Architecture for Automatic Recognition of Group Activities Using Local Motions and Context. IEEE Access. 10(August) (2022) 79874–79889. doi: 10.1109/ACCESS.2022.3195035.
60. B. Z. and X. C. X. Huang, Y. Hu, X. Luo, J. Han, Boosting Variational Inference With Margin Learning for Few-Shot Scene-Adaptive Anomaly Detection. IEEE Trans. Circuits Syst. Video Technol. 33(6)(2022) 2813–2825. doi: 10.1109/TCSVT.2022.3227716.
61. Y. Li, Daoheng ; Nie, Xiushan ; Li, Xiaofeng ; Zhang, Yu ; Yin, Context-related video anomaly detection via generative adversarial network. Pattern Recognit. Lett. 156(2022) 83-189.doi: 10.1016/j.patrec.2022.03.004.
62. G. Yu, S. Wang, Z. Cai, X. Liu, C. Xu, and C. Wu, Deep Anomaly Discovery from Unlabeled Videos via Normality Advantage and Self-Paced Refinement. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. June (2022) 13967–13978. doi: 10.1109/CVPR52688.2022.01360.
63. Y. L. Zhi Zhang, Sheng-hua Zhong , Ahmed Fares, Detecting abnormality with separated foreground and background: Mutual Generative Adversarial Networks for video abnormal event detection. Comput. Vis. Image Underst. 219(june) (2022) 103416. doi: https://doi.org/10.1016/j.cviu.2022.103416.
64. Y. C. Zhongyue Wang, Anomaly detection with dual-stream memory network. J. Vis. Commun. Image Represent. 90 (February) (2023). doi: https://doi.org/10.1016/j.jvcir.2022.103739.
65. V. T. Le and Y. G. Kim, Attention-based residual autoencoder for video anomaly detection. Appl. Intell. 53(3)(2023) 3240–3254. doi: 10.1007/s10489-022-03613-1.
66. H. Deng, Z. Zhang, S. Zou, and X. Li, Bi-directional Frame Interpolation for Unsupervised Video Anomaly Detection. IEEE Winter Conf. Appl. Comput. Vision. (2023) 2633–2642. doi: 10.1109/WACV56688.2023.00266.
67. A. Aich, K. C. Peng, and A. K. Roy-Chowdhury, Cross-Domain Video Anomaly Detection without Target Domain Adaptation. IEEE Winter Conf. Appl. Comput. Vision. (2023) 2578–2590. doi: 10.1109/WACV56688.2023.00261.