Proceedings of CONF-MLA Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/80/2024CH0061

A Review on Computer Vision-Based Methods for Abnormal Human Action Recognition

Mengdi Yuan

School of Computer Science, Southwest Petroleum University, Chengdu, China

yuanmengdi@ldy.edu.rs

Abstract. Human behavior recognition constitutes a crucial research domain within both computer vision and behavior recognition. As a branch of human behavior recognition, abnormal behavior recognition has witnessed rapid advancement in recent years, which is capable of enhancing the governance level of public safety. To investigate the theoretical and technical progress of abnormal behavior recognition in public places within the realm of computer vision, this paper initially delineates the definition of abnormal behavior in public places. Secondly, as computer vision and pattern recognition technologies have progressed, algorithms are now divided into two distinct categories: traditional methods and those leveraging deep learning techniques. The identification of atypical human behavior can be divided into two approaches depending on the detection of body key points: one approach uses skeletal key points, while the other focuses on analyzing temporal and spatial features. Finally, this paper conducts a review of the mainstream datasets of abnormal human behavior both at home and abroad, analyzes the performance of related algorithms on the datasets, and gives future research directions as well as optimization suggestions.

Keywords: abnormal behavior recognition, computer vision, deep learning.

1. Introduction

In crowded public places such as high-speed rail stations and airports, failure to detect and handle emergencies such as stampedes and terrorist attacks in a timely manner will result in severe consequences. Research indicates that pedestrian abnormal behavior in crowd activities is a significant factor contributing to sudden public safety incidents [1,2]. Simultaneously, with the widespread application of surveillance and other monitoring devices, techniques for recognizing abnormal behavior, rooted in the field of computer vision, are continuously evolving and being refined, establishing themselves as the predominant approach for identifying unusual conduct in public spaces. To reduce accidents and prevent various offenses, the recognition of abnormal human behavior has become extensively utilized in public safety, traffic control, intelligent monitoring and so on [3]. The ultimate goal of the research related to human abnormal behavior recognition is to liberate the human eye, to solve the low recognition rate and high leakage rate existing in the traditional monitoring system. The study is therefore of great practical importance.

Traditional methods require manual feature extraction from video images, which mainly include basic visual features at a low level such as motion flow, optical flow, and Histogram of Oriented Gradient(HOG). The combination of features into fixed-size video-level descriptions and the final

DOI: 10.54254/2755-2721/80/2024CH0061

prediction using classifiers such as Support Vector Machine(SVM) are mainly used to detect anomalous behaviors through traditional machine learning methods. With the ongoing advancement of deep learning technology, researchers are exploiting deep learning to extract advanced features like appearance and motion from videos, thereby resulting in an enhancement in accuracy and robustness in tasks such as recognition and classification. The methods relying on deep learning for the recognition of anomalous behaviors have emerged as the dominant approach [4].

Meanwhile, depending on whether it detects key points of the human body or not, the deep learning-based method can be further classified into the methods for recognizing abnormal human behaviors based on skeletal key points and temporal-spatial feature analysis.

Recent research by scholars has focused extensively on recognizing atypical human behavior in public areas. POPOOLA et al. [5] have studied the methods for anomaly detection in video surveillance, including the definition of the anomaly detection task, the problems faced, and the anomaly judgment strategies for the point model, the sequence model, and the graph model. MABROUK et al. [6] undertook in-depth research regarding behavior representation and behavior modeling, and proffered a framework and method for the classification and modeling of behaviors. Ji Xiaofei et al. [7] undertake an analysis of abnormal behavior from the viewpoints of recognition and detection, and devise a joint algorithm for human detection and abnormal behavior recognition through feature enhancement to overcome the deficiencies of previous methods in handling complex environments, high similarity of actions, and human occlusion. In conclusion, the present research fails to offer a clear definition of abnormal behavior in public places, and the technological approaches employed lack data analysis, while the comprehensive review of existing algorithms is inadequate. Hence, this paper will categorize the current principal human abnormal behavior recognition methods in accordance with the types of data involved and the technical basis; conduct a statistical analysis of the performance of related algorithms on abnormal behavior datasets; explore the problems existing in the current abnormal behavior recognition research and anticipate the future development.

2. Abnormal human behavior in public places

2.1. Definition of Abnormal Behavior

Regarding "abnormality", it typically pertains to infrequent occurrences that are conspicuously distinguishable from other behaviors [8]. Researchers denominate deviations from normal behavior as abnormal behavior. In public spaces, abnormal behavior typically denotes actions that are incompatible with the surrounding circumstances and may even pose a threat to public safety. Under this definition, prevalent human abnormal behaviors encompass: falling, transgressing boundaries, carrying hazardous items, engaging in fights, committing robbery, perpetrating theft, trampling, and damaging public facilities, etc. Table 1 presents instances of human abnormal behaviors in diverse scenarios.

Table 1. Common abnormal human behavior in various contexts

Context	abnormal human behavior
subway station	detention、collision、begging
gas station	detention、smoke、call
park	trample the lawn、cultivate
freeway	walk、run、cycling
escalator	climb、vigorous exercise
bank	run、wandering、gather
exam room	turning one's head、turn around、look around
scenic area	crowd surge、fare evasion

DOI: 10.54254/2755-2721/80/2024CH0061

2.2. Research methods and current status of abnormal behavior

During the course of literature retrieval, the bibliometric analysis approach was employed. Within the China National Knowledge Infrastructure (CNKI) and Web-based Citation Database, relevant literature spanning from 2000 to 2024 was retrieved through keyword searching and subjected to analysis. Firstly, conduct a search for the Chinese keywords on CNKI: "abnormal behavior recognition", "public places", "detection", and "crowds". In Web of Science (WOS), undertake a search for the English keyword phrases "Anomaly", "Abnormal", "Anomalous", "Pedestrian behavior" and "Crowd activity". Following this, the pertinent literature related to the research topic is selected for further analysis.

In recent years, anomalous behavior recognition has evolved into a research focus. Since 2010, a remarkable escalation has been witnessed in the number of research papers concerning anomaly detection.

3. Methods for Identifying Abnormal Human Behavior

At present, human abnormal behavior recognition methods can be categorized into traditional approaches and those utilizing deep learning, based on their technical foundations. Deep learning methods are capable of extracting richer image information. Sufficient training data enables the model to be more consistent, and it can acquire abundant semantic features in the video, thus making such methods more applicable for human behavior recognition.

Built upon deep learning, methods can be classified into two types depending on human keypoint detection: the methods for recognizing abnormal human behaviors based on skeletal key points and temporal and spatial feature analysis.

3.1. Traditional methods of recognizing abnormal human behavior

Conventional methods extract local high-dimensional visual features from specific regions of the video, integrate them into fixed-size descriptions at the video level, and subsequently employ classifiers to make the ultimate predictions. The dense trajectory algorithm put forward by Wang et al. [9] constitutes an efficient video recognition approach that extracts video tracking trajectories on the basis of dense trajectories and motion boundary histograms. The algorithm can sample feature points densely on multiple spatial scales for each frame, utilize the optical flow field to obtain the trajectories in the video sequence, extract and encode features along these trajectories, and finally, train an SVM classifier using the encoded results. In the same year, the team [10] enhanced the algorithm by optimizing optical flow images, improving feature regularization and refining feature encoding, all based on the Dense Trajectories algorithm, leading to a considerable improvement in the final performance.

Nevertheless, traditional approaches necessitate abundant experience and specialized background knowledge for the manual extraction of features from images, which is not only time-consuming but also labor-intensive.

3.2. Abnormal behavior recognition method based on deep learning

At the core of deep neural networks lies the development of a multi-layered framework that captures targets at multiple levels of abstraction. Through the utilization of this architecture, the networks strengthen feature robustness, facilitating the efficient extraction and representation of abstract semantic information via high-level features distributed throughout the various layers. Deep learning approaches are capable of acquiring more abundant image information. Sufficient training data enable the model to fit better and obtain rich semantic features in the video. Such methods are more appropriate for human behavior recognition.

The approach for recognizing human abnormal behavior by analyzing temporal and spatial features mainly encompasses Two-Stream networks, 3D Convolution, and Long Short-Term Memory (LSTM) models.

3.2.1. Recognition method of abnormal human behavior based on spatio-temporal feature analysis.

Spatio-temporal features are capable of capturing human motion information concurrently in both the

DOI: 10.54254/2755-2721/80/2024CH0061

temporal and spatial dimensions, and have a considerable advantage in characterizing human motion with a rich three-dimensional information quantity. The approach for recognizing human abnormal behavior by analyzing temporal and spatial features mainly encompasses Two-Stream networks, 3D Convolution, and Long Short-Term Memory (LSTM) models.

The Two-Stream model is a deep learning architecture prevalently employed for recognizing behaviors in videos. The feature extraction process for action recognition is segregated into two discrete branches: one branch extracts spatial features by utilizing RGB data, while the other captures temporal optical flow features. Eventually, it combines the two types of features for action recognition. Karen et al. [11] propose the Two-Stream Convolutional Neural Networks(CNN) algorithm. CNN are adept at handling static appearance information rather than motion information. Hence, the features of motion information are extracted using the optical flow network. CNN merely needs to learn the mapping between the optical flow input and the classification of the temporal flow network. The two networks do not interfere with each other, and the model performs very well. Wang et al. [12] aim to tackle the issue that two-stream convolutional networks(TSN) are incapable of modeling long-range temporal structures and propose the TSN algorithm. Contrary to Two-stream which employs single frames or stacked frames, TSN utilizes sampling and aggregation from the entire video, thereby facilitating effective learning of action models via the entire motion video. By utilizing average pooling and integrating multi-scale temporal windows, the trained model can be easily adjusted for action recognition in both trimmed and untrimmed videos. Zhu et al. [13] put forward the Hidden Two-Stream Network, employing an unsupervised pre-training approach, to tackle the challenge of capturing the temporal connections between video frames. Zhou et al. [14] put forward the TRN network structure. Based on the original framework of TSN, they improved the original fusion function to characterize the relationships of different temporal segments, forming an MLP structure. Meanwhile, through multi-scale feature fusion in the temporal dimension, the video-level robustness and the anti-interference ability between actions of different speeds are enhanced.

Convolution 3D(C3D) expands upon 2D convolution by adding a temporal dimension, enabling the direct extraction of features encompassing both temporal and spatial aspects. The C3D algorithm put forward by Tran et al. [15] employs 3D convolutional kernels to capture the spatio-temporal features, featuring high efficiency and high accuracy. Considering the issues of high computational cost and large model storage of 3D convolutions, Qiu et al. [16] propose Pseudo-3D Residual Networks(P3D), they transformed the 3D convolutions employed in the video field. Instead of using 3*3*3 convolutions, they utilized 1*3*3 convolutions and 3*1*1 convolutions to simulate 3D convolution operations on 2D image data for capturing richer spatial information, thereby significantly reducing the computational load. Tran et al. [17] refine the N×t×d×d in the convolution module by decomposing it into N×1×d×d 2D spatial convolution and M×t×1×1 1D temporal convolution. Compared with full 3D convolution, the R(2+1)D convolution decomposition enhances the complexity of the representable functions without altering the quantity of parameters. Meanwhile, it forces 3D convolution to be transformed into separate spatial and temporal components, leading to a lower training error. Zolfaghari et al. [18] propose the Efficient Convolutional Network(ECO) algorithm. This algorithm adopts an end-to-end architecture, integrates the feature information in various time periods, and classifies actions based on temporal information. It can not only handle temporal information more effectively but also process information over longer time spans.

LSTM networks typically employ CNN to extract spatial features. Donahue et al. [19] put forward the Long-term Recurrent Convolutional Networks(LRCN), which integrates the conventional CNN networks and LSTM. It holds the capacity to handle both temporal video inputs and single-frame images, and also has the capability to output single-value predictions or sequence predictions, thereby making the LRCN a consummate network for processing sequential input and output information.

3.2.2. Abnormal Behavior Recognition Based on Key Points of Human Skeleton. Compared with other previous input methods, the input information of the skeleton conspicuously lacks the visual appearance details of the human body and mainly conducts action recognition by using keypoint coordinates. The

DOI: 10.54254/2755-2721/80/2024CH0061

skeleton sequence possesses three remarkable characteristics: Firstly, there exists a strong correlation between each node and its adjacent nodes, thus the skeleton framework encompasses abundant information regarding the body structure. Secondly, temporal continuity exists not only within the same joints but also within the body structure. Ultimately, a symbiotic relationship exists between the spatial and temporal domains.

Yan et al. [20] pioneer the Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition(ST-GCN) algorithm. This algorithm employs the temporal sequence representations of human joint positions to create dynamic skeletons and enhance graph convolutions for developing spatio-temporal convolutional networks, enabling the capture of spatio-temporal variation relationships. Its hierarchical property eliminates the requirement of manually crafting part allocation or traversal rules, which not only endows the model with stronger expressiveness and higher performance, but also makes it more prone to generalization to different scenarios.

Tang et al. [21] put forward a deep progressive reinforcement learning (DPRL) approach for behavior recognition based on skeleton videos. By employing the deep progressive reinforcement learning method, it gradually extracts frames with substantial information and discards those with little information in the sequence. Finally, based on the diagrams of joints and skeletal structure, action recognition is carried out using a GCNN.

Fan et al. [22] present an algorithm of gated recurrent unit - feedforward network (GRU-FFN) based on the human skeleton. The GRU units are introduced into the feedforward network to establish a feedback loop for memory and weight sharing. Multiple GRU units are interconnected, and eventually, a message passing mechanism is utilized to predict the differences from the real frames.

Pang et al. [23] incorporate Transformer networks into the traditional skeleton framework, thereby establishing the Skeleton-Transformer algorithm. The original skeleton is decomposed into local and global pose components, and temporal information is added through the utilization of position embeddings. Abnormality is detected based on the discrepancy between the predicted and real skeleton sequences.

Sun et al. [24] propose the Hierarchical Semantic Contrast(HSC) algorithm. Founded on the human skeleton and in alliance with appearance features and background information, a binary classifier is exploited to detect abnormal behaviors. This model is conducive to dealing with the diversity of normal behavior patterns and can effectively differentiate positive and negative samples associated with the scene.

4. Human Abnormal Behavior Dataset and Test Performance

4.1. Mainstream dataset of abnormal crowd behavior

During the process of recognition and detection of abnormal behavior, numerous public datasets are accessible for utilization. The following text will present an introduction to the commonly used public datasets for abnormal behavior detection from aspects such as video content, scenes, and characteristics.

Hollywood2: The Hollywood2 dataset was published by IRISA in 2009. It encompasses 3,669 samples from 12 action categories and 10 scenes, all of which were sampled from 69 Hollywood films. The expressions, postures, and attires of the actors in the video samples, along with the variations in camera movements, lighting changes, occlusions and backgrounds are highly diverse, closely resembling real-world scenarios. Consequently, it is extremely challenging to analyze and recognize the behaviors.

UCF50: UCF50 is an action recognition dataset issued by the University of Central Florida, composed of real videos from YouTube and encompassing 50 action categories, such as baseball pitching, basketball shooting, cycling, billiards, breaststroke swimming, diving, and drumming.

HMDB51: HMDB51 was issued by Brown University in 2011. The videos within the database primarily originate from movies, public databases, YouTube, and other online video libraries. The database holds 6,849 samples classified into 51 categories, each featuring at least 101 samples. It chiefly comprises facial expressions and body movements.

DOI: 10.54254/2755-2721/80/2024CH0061

UCF-101: UCF-101 is a series of databases that were issued by the University of Central Florida (UCF) of the United States since 2012. The database samples are derived from a diverse array of sports footage gathered from BBC/ESPN television broadcasts, as well as from video content hosted on YouTube. Comprising a total of 13,320 video clips, the sample collection encompasses categories such as makeup artistry, musical instrument, and various sports disciplines.

ShanghaiTech: This dataset encompasses video data compiled from 13 varied scenarios, captured under complex lighting conditions and diverse camera perspectives, totaling 330 training videos and 107 test videos, encapsulating a comprehensive array of 130 anomalous events. The analysis of abnormal behaviors within is multifaceted and intricate, designating this dataset as a substantial repository for abnormal behavior detection.

CUHK Avenue: This dataset is utilized for scholarly inquiry into activities or behaviors within densely populated environments. It comprises two distinct sub-datasets: the traffic dataset, which encompasses the MIT traffic video collection, and the pedestrian dataset. The dataset comprises a 90-minute video sequence capturing traffic conditions, supplemented by manually annotated pedestrian ground truth for select frames. Additionally, it encompasses a 30-minute video of New York's Grand Central Station, which lacks any annotations or descriptive facts.

HR-Avenue: The HR-Avenue dataset consists of a single scene with 16 training videos and 21 test videos. The abnormal dataset comprises 29 synthesized scenes generated via Cinema4D and natural background synthesis, encompassing 186 normal training videos and 211 test videos.

4.2. Evaluation Indicators for Identification of Abnormal behavior

In the domain of abnormal behavior identification, the prevalent metrics for assessment encompass accuracy as well as the area under the receiver operating characteristic curve (AUC).

Accuracy is defined as the ratio of correctly predicted samples to the total number of samples in a classification task, and it constitutes one of the most frequently utilized metrics for evaluating the performance of a classification model.

In various contexts, the sample distribution of abnormal behavioral data may exhibit bias, and diverse algorithms may render discrepant judgments regarding the same behavior within the same scenario. In the domain of anomaly detection, the ROC curve is impervious to the influence of both positive and negative samples, thereby obviating sample distribution bias. It is commonly employed to evaluate the efficacy of algorithms. Consequently, AUC is quantified as a value ranging between 0 and 1, where a higher magnitude of this value signifies enhanced algorithmic efficacy.

4.3. Performance Comparison of Human Abnormal behavior Recognition Test

In accordance with the review's analytical framework, the precision and performance metrics of representative algorithms against the dataset are delineated in Table 2.

Table 2. Abnormal Behavior Recognition Test Performance

Category	Algorithm	Year	Dataset	Performance
Traditional methods of recognizing abnormal human behavior	DT	2013	Hollywood2,UCF50,HMDB51	accuracy —— 58.2%,84.5%,46.6%
	iDT	2013	Hollywood2,UCF50,HMDB51, UCF-101	accuracy —— 64.2%,91.2%,57.2%, 86.4%
Recognition method of abnormal human behavior based on	TwoStream CNN	2014	UCF-101, HMDB51	accuracy —88.0%,59.4%
	TSN	2016	UCF-101, HMDB51	accuracy —94.2%,69.4%

DOI: 10.54254/2755-2721/80/2024CH0061

Table 2. (continued).

spatio-temporal feature analysis	Hidden Two-Stream Convolutional Networks	2018	UCF-101	accuracy —89.82%
	TRN	2018	UCF-101	accuracy —83.83%
	C3D(3 nets)+linear SVM	2014	UCF-101	accuracy —85.2%
	C3D (3 nets) + iDT + linear SVM		UCF-101	accuracy —90.4%
	P3D ResNet	2017	UCF-101	accuracy —88.6%
	R(2+1)CD-RW	2018	UCF-101, HMDB51	accuracy —93.6%,66.6%
	R(2+1)D-Flow			accuracy —93.3%,70.1%
	R(2+1)D-TwoStream			accuracy —95.0%,72.7%
	ECO	2018	UCF-101	accuracy —90.3%,61.7%
	LRCN	2014	UCF-101	accuracy —82.9%
	LSTM composite model	2015	UCF-101	accuracy —84.3%
Abnormal Behavior Recognition Based on Key Points of Human Skeleton	ST-GCN	2018	Kinetics	accuracy —72.4%
	DPRL	2018	SYSU,UT	accuracy —76.7%,98.0%
	DPRL+GCNN			accuracy —76.9%,98.5%
	GRU-FFN	2021	ShanghaiTech, CUHK Avenue	AUC—82.6%,91.7%
	Skeleton-Transformer	2022	HR-Avenue	AUC—86.7%
	HSC	2023	ShanghaiTech, CUHK Avenue	AUC—82.4%,93.7%

From Table 2, it is evident that deep learning-based human abnormal behavior recognition technologies exhibit superior recognition accuracy and enhanced algorithmic performance when compared to traditional methods.Concurrently, traditional methodologies are not antithetical to deep learning; the integration of iDT and C3D on the UCF-101 dataset yielded an accuracy rate of 90.4%, [15] outperforming the accuracy of the TwoStream CNN [11] and P3D [16] on the same dataset.

In recent years, owing to the strong correlation between skeletal keypoints and human behaviors, the approach of human abnormal behavior recognition based on skeletal keypoints has evolved rapidly and emerged as a current research focus. With the continuous refinement of the algorithm, the overall performance manifested is relatively favorable.

DOI: 10.54254/2755-2721/80/2024CH0061

5. Conclusion

The recognition technology of abnormal human behaviors in public spaces possesses considerable research value in production and daily life. On this basis, this article has conducted a systematic analysis of the research on human abnormal behavior recognition technology under computer vision. Firstly, the definition and characteristics of abnormal human behaviors in public spaces are presented. The common application scenarios and typical abnormal behaviors in the current abnormal behavior technology are enumerated. Secondly, based on the aforementioned definitions, a classification and summary of the human abnormal behavior recognition technology was carried out. Ultimately, the public datasets frequently employed for anomaly behavior recognition and detection are enumerated. Starting from the two indicators of accuracy and AUC, the performance of the algorithms listed previously on these datasets is consolidated.

Furthermore, taking into account the current research status and the challenges witnessed in human abnormal behavior recognition technology employing computer vision, this article puts forward several future research directions in this domain.

Enhance the recognition technology. The existing recognition methods have failed to achieve multi-type and precise identification of abnormal behaviors. It is proposed that the existing methods be improved to realize an abnormal behavior identification method based on multi-data fusion.

Realize cross-scenario models. Most existing methods train models for a single and fixed scenario, and thus are unable to effectively conduct abnormal behavior recognition across scenarios. In practical applications, training models separately for each monitoring scene will bring about a considerable workload. Incorporating scene information into abnormal behavior detection models is one of the research directions available in the future.

Forecast abnormal behavior. At present, most of the recognition technologies for abnormal behaviors are based on the existing data and are used for detecting and discriminating the events that have already taken place. If abnormal behavior can be detected and alerted prior to the occurrence of an incident, the applications and scope of this technology will be considerably broadened. At present, there is still ample opportunity for further research in this domain.

References

[1] Kok, V. J., Lim, M. K., Chan, C. S. (2016). Crowd Behavior Analysis: A Review where Physics meets Biology. Neurocomputing, 177:342-362.

[2] Zhang, M., Han, Y. X., and Liu, Z. C. (2022). A Method for Detecting High-altitude Safety Protective Equipment for Construction Workers Using Deep Learning. Chinese Journal of Safety Science, 32(5):140-146.

[3] Zhang, Y. X. (2022). Research on Human Pose Recognition and Abnormal Behavior Understanding. Xi'an Technological University.

[4] Xu, Tao., Tian, C. Y., Liu, T. (2021). A Survey on Abnormal Behavior Detection of Crowds Based on Deep Learning. Computer Science, 48(09):125-134.

[5] Popoola, O. P., & Wang, K. (2012). Video-based abnormal human behavior recognition—a review. IEEE Transactions on Systems Man & Cybernetics Part C, 42(6):865-878.

[6] Ben Mabrouk, A., & Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: a review. Expert Systems with Applications, 91(jan.):480-491.

[7] Ji, X. F., Zhao, D. Y. (2023). Combined algorithm for human detection and abnormal behavior recognition. Science and Engineering, 23(08):3370-3378.

[8] Liu, Y. P. (2023). Method of Abnormal Behavior Recognition in Video Surveillance Images Based on Neural Networks. Digital Communication World, (7):71-73.

[9] Wang, H., Klser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103: 60-79.

[10] Wang, H., & Schmid, C. (2014). Action Recognition with Improved Trajectories. 2013 IEEE International Conference on Computer Vision. IEEE.

DOI: 10.54254/2755-2721/80/2024CH0061

[11] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 1.
[12] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., & Tang, X., et al. (2016). Temporal segment networks: towards good practices for deep action recognition. Springer, Cham.
[13] Hauptmann, A., Newsam, S., Lan, Z. Z. & Zhu, Y. (2018). Hidden two-stream convolutional networks for action recognition. Springer, Cham.
[14] Zhou, B., Andonian, A., & Torralba, A. (2017). Temporal relational reasoning in videos.
[15] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. IEEE.
[16] Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. IEEE.
[17] Tran, D., Wang, H., Torresani, L., Ray, J., & Lecun, Y. (2017). A closer look at spatiotemporal convolutions for action recognition.
[18] Zolfaghari, M., Singh, K., & Brox, T. (2018). Eco: efficient convolutional network for online video understanding. Springer, Cham.
[19] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., & Saenko, K. (2015). Long-term recurrent convolutional networks for visual recognition and description. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[20] Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition.
[21] Tang, Y., Tian, Y., Lu, J., Li, P., & Zhou, J. (2018). Deep progressive reinforcement learning for skeleton-based action recognition. IEEE.
[22] Fan, B., Li, P., Jin, S., & Wang, Z. (2021). Anomaly Detection based on Pose Estimation and GRU-FFN. 2021 IEEE Sustainable Power and Energy Conference (iSPEC), 3821-3825.
[23] Pang, W., He, Q., & Li, Y. (2022). Predicting skeleton trajectories using a skeleton-transformer for video anomaly detection. Multimedia systems,1481–1494
[24] Sun, S., & Gong, X. (2023). Hierarchical Semantic Contrast for Scene-aware Video Anomaly Detection. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 22846-22856.