DOI 10.15622/ia.23.2.5

N. SUJATA GUPTA, K. RAMYA, R. KARNATI

A REVIEW WORK: HUMAN ACTION RECOGNITION IN VIDEO SURVEILLANCE USING DEEP LEARNING TECHNIQUES

Sujata Gupta N., Ramya K., Karnati R. **A Review Work: Human Action Recognition in Video Surveillance Using Deep Learning Techniques.**

Abstract. Despite being extensively used in numerous uses, precise and effective human activity identification continues to be an interesting research issue in the area of vision for computers. Currently, a lot of investigation is being done on themes like pedestrian activity recognition and ways to recognize people's movements employing depth data, 3D skeletal data, still picture data, or strategies that utilize spatiotemporal interest points. This study aims to investigate and evaluate DL approaches for detecting human activity in video. The focus has been on multiple structures for detecting human activities that use DL as their primary strategy. Based on the application, including identifying faces, emotion identification, action identification, and anomaly identification, the human occurrence forecasts are divided into four different subcategories. The literature has been carried several research based on these recognitions for predicting human behavior and activity for video surveillance applications. The state of the art of four different applications' DL techniques is contrasted. This paper also presents the application areas, scientific issues, and potential goals in the field of DL-based human behavior and activity recognition/detection.

Keywords: face recognition, emotion recognition, action recognition, anomaly recognition, DL, human behavior and activity recognition/detection.

1. Introduction. Numerous actual environments have applications for human behavior identification, such as intelligent video surveillance and purchasing behavior evaluation [1]. There are many uses for surveillance footage, particularly in indoor, outdoor, and public spaces. Safety includes surveillance as a crucial component. For the sake of security and protection, surveillance cameras are now a must [2]. Among the key goals of the Indian government's growth initiative, Digital India is e-surveillance. It still includes surveillance footage in some form. Efficient surveillance, a need for less labor, cost-effective auditing capabilities, adopting of recent safety trends, etc. are all benefits of surveillance footage [3]. Until now, monitoring was done manually. We have to manage enormous amounts of video footage that can easily wear individuals out. Furthermore, omissions brought on by manual intervention will significantly reduce the structure's efficacy [4]. Video surveillance automation has provided a solution for this. Nowadays, it is difficult to manually watch every incident captured on a CCTV (Closed Circuit Television) camera. Even if the incident occurred previously, manually looking for it in the video footage is a laborious procedure [5].

Among the oldest and most active areas of computer vision and pattern identification study is monitoring footage. In earlier times, operator humans who watched dozens of displays at once were the mainstay of video-based

monitoring systems [6]. Individuals are shown to be extremely inconsistent in identifying the so-called "unusual events" while evaluating either online video clips or archival data because of the amount of data and relatively lengthy monitoring recordings [7]. The key difficulty is creating an intelligent, autonomous, video-based monitoring system that doesn't need human involvement. An emerging area in the field of automated video surveillance structures is the analysis of anomalous occurrences from video [8].

In recent times, DL-based video surveillance systems (VSSs) have produced a range of impressive outcomes when used for diverse purposes, including crowd counting (CC) [9], abnormal event detection (AED) [10], object detection (OD) [11], human action recognition (HAR) [12], etc. Deep networks mimic human vision by modeling high-level abstractions via several levels of non-linear transformations. DL algorithms must be trained on a large quantity of data to do this [13]. These techniques, meanwhile, have drawbacks to numerous other uses and only perform effectively for specific applications when getting the data is simple [14].

The most important details are that there are insufficient funds and data scales to train DL algorithms from scratch, it is costly and takes time to gather massive databases, the majority of DL algorithms rely on supervised learning, and enlisting the help of human experts to label training datasets is a significant expense and effort [15].

Many other strategies are being put forth; however, the initial research heavily depends on trajectory-based methods. These methods use visual tracking to make an effort to predict the target's trajectories, while a model is acquired to explain typical activities [16]. The activity associated with trajectory deviations from the learned model is therefore considered an anomaly. However, because of their great temporal complexity and the occlusion problem brought on by objects moving, these approaches are unsuitable for difficult and dense situations [17]. As a result, non-object-centered unsupervised techniques have become increasingly popular recently. By learning typical patterns of activity from the behavior-related traits of individuals and things in geographical and temporal settings, these techniques address the issue of recognizing anomalies [18]. Target size, gradient, speed, and direction are all typically considered to be behavioral characteristics along can be represented with low-level illustrations like dense spatial-temporal interest points (dense STIPs), histograms of optical flow (HOF), and histograms of oriented gradients (HOG). These techniques are superior to trajectory-based techniques because they operate at the pixel level, which renders them more reliable in challenging settings [19]. Although there are numerous distinct kinds of anomalous behavior, all of these approaches rely on hand-crafted characteristics that are challenging to explain a priori. They are also incapable of adapting to defects that were never seen before [20].

In the last decade, DL approaches have been primarily employed to address a variety of computer vision problems, beating the state-of-the-art in a variety of challenging scenarios, based on the depth of hidden layers we can differentiate the neural network into four various categories as illustrated in Figure 1.

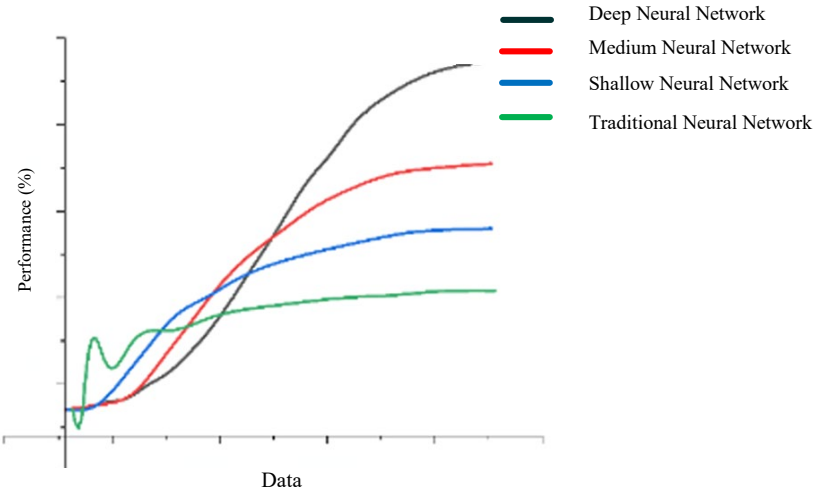

Fig. 1. Graphical representation of various neural networks' performance in human activity recognition

Traditional neural networks, also known as single-layer perceptrons, are simple linear classification models that cannot handle complicated problems. Shallow neural networks feature a limited number of hidden layers, which makes them successful for basic tasks but ineffective for more complicated ones. With a reasonable number of hidden layers, medium neural networks may learn more complicated features and patterns. With over six hidden layers, deep neural networks excel at learning highly abstract characteristics and can perform complicated tasks such as picture and speech recognition. Deep network training, on the other hand, maybe computationally costly and require a huge quantity of data to avoid overfitting. Integrating recognizing objects, object classification, and action identification. Since DL, a subtype of ML, trains to interpret the input as a hierarchy of nested concepts within various stages of the neural network, this study focuses on DL advancement to achieve outstanding results. As data volumes increase for recognizing behavior and activities, DL outperforms classical machine learning.

2. Literature Survey. The review has covered papers under the years 2019-2023 are provided in Figure 2. Totally 40 research articles from different sources are collected and the works are elaborated clearly in the following section with their respective pros and cons.

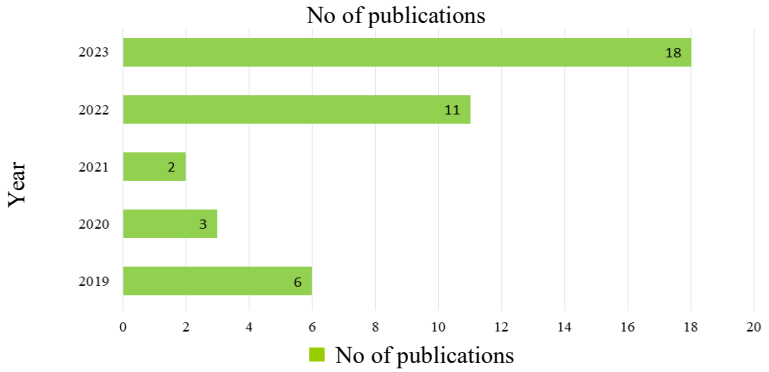

Fig. 2. Year-wise weightage of papers considered in the review

From Figure 2 it is clear that recent 2023 papers are considered in more numbers compared to the other papers. The variation in the study from 2019 to 2023 can be evaluated from this review. The collection of journals that have shown a lot of curiosity in recognizing human conduct and activities for video surveillance is depicted in Figure 3.

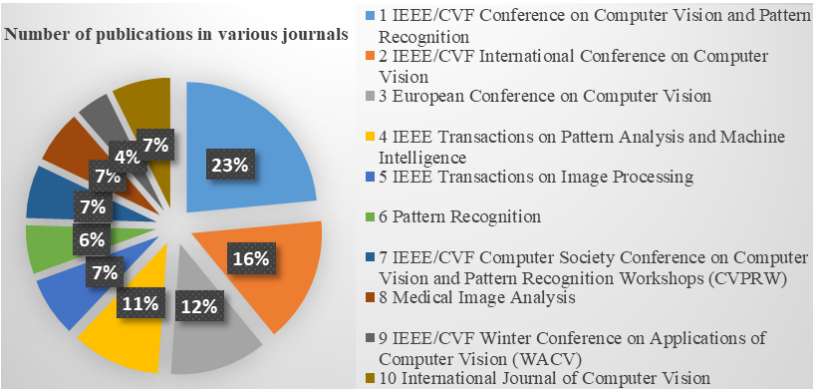

Fig. 3. Top 10 journals which have shown interest in video surveillance

Face recognition. In study [21] the authors suggested a technique for updating model parameters that will allow the dispersed EC

environment to synchronize the global DL model. A dynamic data movement strategy is further suggested to resolve the disparity between the workload and processing capabilities of edge nodes. The suggested DIVS system can effectively handle video surveillance and analysis duties, according to experimental results, which also demonstrated the EC architecture's ability to deliver elastic and scalable processing capacity.

In paper [22] the authors suggested an attribute-driven strategy for feature disentangling and frame re-weighting. In the study disentangling the characteristics of a single frame into sub-feature groups that individually relate to distinct semantic properties is described. The final representation is created by aggregating the sub-features at the temporal dimension after the sub-features have been reweighted by the attribute recognition confidence. This method enhances the most informative areas of each frame and helps the representation of the sequence be more discriminative. Numerous ablation experiments support the value of both feature disentangling and temporal re-weighting. The proposed strategy outperforms current state-of-the-art methods, as shown by the experimental findings on the iLIDS-VID, PRID-2011, and MARS datasets.

Paper [23] offered a straightforward method for achieving this goal using several key machine learning technologies, including TensorFlow, Keras, OpenCV, and Scikit-Learn. The proposed method detects the face in the picture or video and then assesses whether or not it is covered by a mask. It can distinguish a face and a mask in motion and a video as a surveillance task performance. The method achieves superb accuracy. We look at the most effective parameter settings for the Convolutional Neural Network model (CNN) to precisely detect the presence of masks without producing over-fitting.

In study [24] the authors state that AdaFocus is an adaptive focus technique that, although requiring a challenging three-stage training pipeline, has achieved a favorable accuracy-to-inference-speed trade-off. Through the use of an enhanced training scheme and a differentiable interpolation-based patch selection operation, this work reformulates the AdaFocus training as a straightforward one-stage approach. Extensive tests on six benchmark datasets show that the model performs far better than the original AdaFocus and other industry benchmarks while being much easier and more effective to train.

Study [25] presented a face mask identification technique for static photos and real-time videos that separates images into "with mask" and "without mask" categories. The Kaggle data set is used to train and assess the model. The collected data set has a performance accuracy rating of 98% and contains around 4,000 images. Comparing the proposed model to DenseNet-121, MobileNet-V2, VGG-19, and Inception-V3 reveals that it is more accurate and computationally economical. Schools, hospitals, banks, airports, and many other public or commercial institutions can use this work as a digital scanning tool.

In paper [26] the hybrid driver and vehicle identification module that can identify both the driver and the vehicle is presented. It can identify the driver and the car using facial recognition, voice recognition, and license plate identification. FaceNet was utilized for face identification, multi-task cascaded convolutional networks were used to crop the faces for facial recognition, and a three-layer long short-term memory model was used for speech verification. A tesseract was employed to identify vehicle license plates. The trials' findings demonstrate that the suggested method can consistently recognize both drivers and autos with zero error, which is a key advancement for guaranteeing the security of institutions.

In study [27] the authors sought to apply a model to complicated data by focusing on tasks for face identification in the picture and real-time video footage of persons wearing and without masks. The suggested technique performed well, with a 99.64% accuracy rate for images and a respectable accuracy rate for real-time video images. According to experimental findings, the suggested technique worked effectively, achieving an accuracy of 99.64% for images and a respectable accuracy for real-time video images. With an accuracy of 100%, recall of 99.28%, f1-score of 99.64%, and an error rate of 0.36%, additional measures demonstrated that the approach beat earlier models. In current technology, face mask detection is frequently employed in fields like artificial intelligence and smartphones.

In paper [28] the authors use a mix of a neural network and a genetic algorithm to pick and categorize face traits. The efficiency of the suggested technique was recently evaluated using both individual and composite elements of the face region. In experimental experiments, composite features outperform face region features. This research also includes a thorough comparison with different face recognition methods found in the FERET database. The classification accuracy achieved by the suggested approach is 94%, which represents a considerable increase and the highest classification accuracy among the findings from earlier investigations.

Study [29] presented real-time face recognition along with a DL framework for person identification and authentification in live or recorded CCTV feeds. The recommended approach is based on the VGGFace DL neural architecture and utilizes transfer learning to retrain the algorithm using a lesser originally designed dataset of 7500 photos of 26 different people. The presented approach provides the maximum level of recognition accuracy, as evidenced by a mean average of 96 percent on real-time inputs and confidence levels that vary from 78.54 percent to 100%.

In paper [30] the simple and effective facial detection method QMagFace is built using a recognition algorithm that utilizes a magnitude-aware angular margin loss and a quality-aware comparison score. The

suggested method incorporates model-specific facial picture characteristics into the comparison process to improve identification performance in unrestricted environments. The tests performed on various face recognition benchmarks and databases show that introducing quality awareness consistently improves recognition performance. The QMagFace source code is accessible to everyone. The information is presented in Table 1.

Table 1. Systematic Survey on Face Recognition

Ref no.	Year	Technique	Significance	Limitation/Future scope
[21]	2019	Upgrading parameter values to ensure global DL network synchronization in a distributed EC scenario	– Effectively handle video surveillance and analysis duties, – Deliver elastic and scalable processing capacity	Low contrast images
[22]	2019	An attribute-driven approach for feature disentangling and then frame re-weighting	– Enhances the most informative areas of each frame – Represents the sequence being more discriminative	Low-resolution movies often feature background movement which is not considered
[23]	2021	Distinguish a face and a mask in motion of a video	Distinguish a face and a mask in motion and a video as a surveillance task performance	Biometric scans can be carried out in the future while wearing a face mask
[24]	2022	Differentiable interpolation-based patch selection procedure	Simple and efficient	The time-consuming phase of feature representation
[25]	2022	A face mask identification technique for static photos and real-time videos	Accuracy = 98%	Yet to increase the reliability
[26]	2023	The hybrid driver and vehicle identification module	Capable of recognizing both drivers and autos with 0% error	Scaling up the recommended technique for estimating the risk factors
[27]	2023	Face detection tasks with an effort to apply a model to complex data	Accuracy rate of 99.64%	Space complexity costs
[28]	2023	Neural network with genetic algorithm	Accuracy = 94%	Non-local change between frames increases the complexity
[29]	2023	DL system for person identification and recognition in live or recorded CCTV feeds	Confidence = 78.54 to 100 % Mean average = 96 %	To reduce the dataset's size while increasing the number of photographs it contains
[30]	2023	QMagFace	Accuracy 98.98 %	Should concentrate on quality-based fusion methods

Despite significant progress, there are still several limitations as given in Table 2 that make video-based physical detection more difficult and demanding.

Table 2. Systematic Survey on Emotion Recognition

Ref no.	Year	Technique	Significance	Limitation/Future scope
[31]	2019	Brand-new RAN	Enhanced FER's performance with occlusion and alternative poses	Makes content interpretation complicated
[32]	2019	The DL-based emotion identification system	The effectiveness of the system is suggested using CNNs and ELMs	The future study assesses system performance in an edge-cloud computing environment
[33]	2019	The research uses the DL technique to categorize emotions through an iterative process	DL methods effectively classify emotions using a large number of sensors	Integrating sensors and modalities
[34]	2020	Self-care Network effectively suppresses uncertainty in deep networks	SCN outperforms advanced techniques with high scores	-
[35]	2021	The article introduces the DL technique for focusing on facial characteristics	Sensitivity is very high	-
[36]	2022	The article lists facial expressions and outlines four steps for execution	System performance was evaluated for databases and compared to current approaches	-
[37]	2022	Transformer-based fusion module combines static vision with dynamic multimodal properties	The performance of the model is increased	-
[38]	2023	Paper reduces processing power using a hierarchical Swin Transformer for expression recognition	Optimal speed-precision balance through high computational effort	Researchers can improve convolutional modules for expression recognition
[39]	2023	Research develops CNN for facial expression recognition using SMM dataset	The model achieves 93.94% accuracy and 67.18% FER2013 score on CK+	Future SMM facial expression collection should include emotions like fear and contempt
[40]	2023	Low-light image enhancement, convolutional neural network for facial emotion recognition	Experimental assessment shows suggested technique outperforms others with 69.3% accuracy	Researchers developed smart glasses prototypes for vision-impaired identification

In actuality, choosing the traits that make a moving object is a difficult process since they have a significant influence on the description and analysis of the activity. For example, when the scene's backdrop changes often or when brand-new things appear out of nowhere, it could be challenging to depict the action. Furthermore, a variety of elements, like scene location (outdoor/indoor) and outfit (dress, suit, footwear, etc.), can influence how the moving object seems.

Survey on Emotion Recognition. In [31] the authors looked at several in-the-wild FER datasets with pose and occlusion characteristics to answer the real-life pose with occlusion robust FER challenge. A unique Region Attention Network (RAN) was suggested in addition to the pose variation FER where an area biassed loss was added to imaginatively capture the significance of facial areas for occlusion. High attention weights for very significant locations might be encouraged by this method. Numerous investigations show that RAN and area biassed loss produce novel findings on FERPlus, AffectNet, RAF-DB, and SFEW, significantly enhancing FER performance with occlusion and changing position.

Study [32] demonstrated a DL-based emotion recognition mechanism built on emotional Big Data. Both speech and video data are recovered and fed to a CNN after being processed in the frequency domain to create a Mel-spectrogram. The outputs of the two CNNs are combined for the last classification using two ELMs and an SVM. The utility of the suggested technology was demonstrated by experimental results.

In study [33] an ongoing procedure that includes incorporating and eliminating enormous quantities of sensor data from several modalities is offered, this research applied a DL approach for emotion categorization. CNN-LSTM is employed in the approach, which reduces the requirement for human feature discovery and engineering by applying a hybrid strategy to raw sensor data. The findings show that DL algorithms are effective at classifying human emotions when a lot of sensors are being utilized (average accuracy is 95% and F-Measure is %95). In addition, hybrid models perform better than previously created Ensemble approaches that train the model through feature engineering (average accuracy 83%, F-Measure = 82%).

In study [34] it is described how to avoid deep networks from over-fitting ambiguous face images, this research proposed a simple but efficient Self-care Network (SCN) that efficiently suppresses uncertainty. Two methods that SCN particularly suppresses the uncertainty are through a self-attention mechanism over a mini-batch to weight every sample used for training with a ranking regularization with a meticulous relabeling procedure to update the labels of those specimens in the lowest-ranked group. Both

simulated FER datasets with gathered Web Emotion datasets have been used to assess the effectiveness of the proposed technique. SCN overcomes state-of-the-art approaches according to outcomes from open benchmarks, scoring 89.35% on FERPlus, 60.23% on Affect-Net, and 88.14% on RAF-DB.

In article [35] the author proposed an attentional convolutional network-based DL technique that can target important face characteristics and surpasses previous approaches on a range of datasets, such as FER-2013, CK+, FERG, and JAFFE. In addition, based on the classifier's output, we use a visualization technique that may help us pinpoint key facial features that indicate different moods. This research study's findings indicate that various emotions are responsive to various facial characteristics.

In research [36] a technique to identify facial expressions was suggested. Its four primary components are face recognition, a CNN framework founded on DL, data augmentation techniques, and a trade-off among data augmentation with DL characteristics. For thorough results from experiments, three benchmark databases – KDEF, GENKI-4k, and CK+ – have been employed. The performance of the suggested approach is being compared to currently employed state-of-the-art techniques, demonstrating its advantages.

Paper [37] offered a transformer-based fusion module that fuses static vision with dynamic multimodal features. The fusion module's cross-attention module focuses the output integrated features on the critical sections, easing downstream detection tasks. To increase model performance even further, we employ various data balance, data augmentation, and post-processing strategies. In the EXPR and AU tracks of the official ABAW3 Competition test, the model wins first place. On the Aff-Wild2 dataset, extensive quantitative evaluations and ablation experiments show how effective the recommended method is.

In article [38] the author uses the hierarchical Swin Transformer for the expression recognition job, which significantly reduces its processing power. The Swin Transformer with CNN and utilize it in an expression recognition job. At the same time, it is fused with a CNN model to suggest a network design that integrates the Transformer and CNN. We next test the suggested strategy using certain expression datasets that are available to the public, and we can achieve competitive results.

Paper [39] proposes a CNN design to distinguish facial expressions and create a facial expression dataset for the SMM. The suggested technique was evaluated for facial expression identification on two distinct benchmark datasets, FER2013 and CK+. We tested the suggested model on CK+ and

obtained accuracy for FER2013 of 93.94% and 67.18%, respectively. To investigate as well as assess the recommended algorithm's accuracy, we used the SMM Facial Expression dataset and attained 96.60% accuracy.

Study [40] offered a method for recognizing facial emotions in masked facial photographs by utilizing low-light image enhancement and feature analysis utilizing a CNN. The suggested method makes use of the AffectNet picture collection, which contains eight different types of facial emotions and 420,299 photos. The head and upper features of the face are represented using boundary and regional representation approaches. A facial landmark identification method-based feature extraction methodology is used to extract features. In an experimental test using the AffectNet dataset, the recommended method achieved an accuracy of 69.3%.

From Table 3 we can find that tracking gets challenging when analyzing photos with fluctuating light, which is a common aspect of actual environments. Outside CCTV cameras are subjected to external illumination fluctuations while gathering footage at night, which may provide low-contrast images that are challenging to comprehend. The adaptive background subtraction method also offers a consistent means of handling recurrent and long-term situation changes, as well as fluctuations in light. Noise reduction is necessary because low-resolution videos frequently have background movement brought on by camera movement or changes in lighting. While the optical flow vector's amplitude is a very effective signal for determining how much movement there is, the flow direction also may provide extra motion data.

Action Recognition. Study [41] presents a deep neural network that collects and categorizes activity characteristics by fusing convolutional layers with LSTM by fusing convolutional layers with LSTM, collects and categorizes activity characteristics. The proposed architecture comprises a two-layer LSTM followed by convolutional layers, a GAP level to reduce the parameters of the model, and then a BN layer to speed up convergence. The efficacy of the model was assessed using three publicly accessible datasets. The accuracy of the model was 95.78%, 95.85%, and 92.63% overall. The results demonstrate that the suggested theory looks more robust and effective in spotting activity than many of the previous results.

In paper [42] the author suggested a deep human action detection framework that is view-invariant and incorporates two crucial action cues: motion and shape temporal dynamics (STD). The motion stream encodes the motion content of the action as RGB-DIs, whereas the STD stream learns long-term view-invariant shape dynamics of action by mining view-invariant features from structural similarity index matrix (SSIM) dependent key depth human pose images. Research that employed cross-subject and

cross-view validation methodologies to measure the performance utilized three publicly accessible benchmarks. In regards to accuracy, ROC curve, and AUC, the technique greatly beat the state-of-the-art at the time.

Study [43] presented a brand-new end-to-end method for identifying unsupervised human actions using skeletons. We provide an innovative design that employs a convolutional autoencoder and graph Laplacian regularization to describe the skeletal geometry across the temporal dynamics of activities. Due to this approach including a self-supervised gradient reverse layer that provides generalization between camera perspectives, it is resistant to viewpoint fluctuations. The proposed method outperforms all earlier unsupervised skeleton-based methods on the cross-subject, cross-view, and cross-setup protocols on the large datasets NTU-60 and NTU-120. Though unsupervised, the system even outperforms a few supervised skeleton-based action recognition techniques owing to its learnable representation.

Research [44] offered a complete method for recognizing human motion in real-time from unprocessed depth picture sequences. It is based on a 3D fully CNN called the 3DFCNN, which dynamically encodes spatiotemporal patterns from raw depth data. The suggested 3DFCNN has been adjusted to operate in real-time with a respectable accuracy performance. On three well-known public datasets, it was recently compared to different state-of-the-art systems, showing that 3DFCNN surpasses other non-DNN-based current methods with an optimal precision of 83.6% yet maintains a noticeably lower computational cost of 1.09 seconds.

In paper [45] the authors used LSTM and CNN to construct a hybrid approach for activity identification. 20 individuals used the Kinect V2 sensor to build a brand-new challenging dataset with 12 distinct groups of human physical activity. A detailed ablation investigation was conducted using several traditional ML and DL neural networks to discover the optimal HAR solution. The accuracy of 90.89% achieved with the CNN-LSTM method shows that the model suggested is suitable for HAR applications.

In paper [46] the authors suggested a unique deep ConvLSTM network for skeletal-based activity identification and then fall detection. In sequence, LSTM systems, fully linked layers, and CNNs are combined. The acquisition method uses human identification and posture assessment to pre-calculate skeletal coordinates from an image/video sequence. From the raw skeleton coordinates and their unique geometrical and kinematic properties, the ConvLSTM network generates fresh directed features. On

the KinectHAR dataset, the recommended approach surpassed CNNs and LSTMs, which recorded accuracy of 93.89% and 92.75%, respectively.

In study [47] a spatially adaptive residual graph convolutional network (SARGCN) based on skeleton feature extraction was suggested for action recognition. It employs a learnable parameter matrix to decrease the number of parameters and improve feature extraction and generalization. To achieve greater accuracy at reduced computing costs and learning challenges, a residual connection is added. The effectiveness of the offered strategy has been confirmed by extensive trials on two substantial datasets.

In [48] the authors examined how well cuboid-aware feature aggregation performed when huge amounts of activity were presented. The authors also suggested monitoring actors and conducting temporal feature aggregation along the corresponding tracks to improve actor's feature representation under big motion. The intersection-over-union (IoU) between the boxes of action tubes/tracks was used by the authors to describe the actor's motion at various fixed time scales. Large-motion activities would eventually have reduced IoU, but slower actions would keep IoU higher. Researchers discover that, as compared to the cuboid-aware baseline, track-aware feature aggregation regularly produces a significant boost in action identification performance, particularly for actions with significant motion. As a result, the authors also provide the most recent findings using the extensive multi-sports dataset.

Paper [49] suggested the Spatio-TemporAl cRoss (STAR)-transformer, that is capable of successfully representing two cross-modal information as a vector. The encoder consists of a full spatio-temporal attention (FAttn) module and a proposed zigzag spatio-temporal attention (ZAttn) module, whilst the continuous decoder comprises a FAttn component with a recommended binary spatio-temporal attention (BAttn) module. Investigations show that the recommended method enhances performance excitingly on the Penn-Action, NTU-RGB+D 60, and 120 datasets.

Study [50] focused on the body occlusions for Skeleton-based One-shot Action Recognition (SOAR) in their study. It primarily takes into account two types of occlusions: arbitrary occlusions and more realistic occlusions brought on by various commonplace items. The authors formalize the first benchmark for SOAR from partly occluded body postures by using the suggested process to blend out sections of the skeleton sequences of three widely used action identification datasets. A novel transformer-based model called Trans4SOAR uses mixed attention fusion and three data streams to lessen the negative impact of occlusions. On all datasets, it performs better than alternative designs, outperforming the best-reported method on NTU-120 by 2.85%.

Table 3. Systematic Survey on Action Recognition

Ref no.	Year	Technique	Significance	Limitation/Future scope
[41]	2020	Deep neural network combining LSTM and convolutional layers	The model achieves 95.67% accuracy	-
[42]	2020	Shape temporal dynamics in deep human behavior perception	The method outperforms the current state-of-the-art in accuracy, ROC curve, and AUC	The author seeks to improve action identification with skeletal and depth details
[43]	2022	New end-to-end approach for skeleton-based unsupervised human action identification	Improved methodology outperforms previous unsupervised skeleton-based algorithms on large datasets	The author focuses on real-time AE-L deployment and spatiotemporal connection enforcement
[44]	2022	Real-time human activity recognition method using unprocessed depth picture sequences	3 DFCNN outperforms non-DNN techniques with 83.6% accuracy	Enhancing recognition accuracy in behaviors remains an ongoing research concern
[45]	2022	The author's mixed approach combines LSTM and CNN for activity identification	CNN-LSTM algorithm achieves 90.89% accuracy in HAR applications	Develop a model for identifying multiple people's activities and expanding advanced physical activities
[46]	2022	The article presents the ConvLSTM network for skeletal-based activities and fall identification	ConvLSTM achieves 98.89% accuracy, surpassing CNNs and LSTMs	
[47]	2023	The study proposes SARGCN for action identification using skeleton feature extraction	The efficiency of the model is high	A writer explores feature extraction and spatiotemporal graph structure analysis
[48]	2023	Investigating cuboid-aware feature aggregation performance in high activity	Track-aware feature aggregation enhances action recognition performance, particularly for significant motion actions	
[49]	2023	STAR-transformer represents cross-modal characteristics as identifiable vectors	Study shows suggested strategy improves performance compared to older methodologies	Scientists develop algorithms without overfitting using limited data
[50]	2023	The study focuses on SOAR body occlusions	NTU-120 outperforms the best SOAR technique by 2.85%	Future investigation into one-shot video identification is excluded

Conquering the following problem will be challenging for optical flow-based motion assessments as given in Table 4. The mathematical feature points of the head, arms, legs, elbows, and shoulders form distinctive abstractions of various postures. The phase of correlation technique should be utilized to determine global motion among every pair of succeeding frames. If global motion is found, a Point Spread Function must be created using the projected slope and length of displacement, and the following frame can be deconvolved using the iterative deconvolution method.

Table 4. Systematic Survey on Anomaly Recognition

Ref no.	Year	Technique	Significance	Limitation/Future scope
[51]	2019	Neural network for anomaly recognition	Detecting abnormal occurrences	
[52]	2022	Real-world traffic surveillance records require ongoing monitoring to ensure proper response to fatal situations. Nevertheless, maintaining constant human supervision of them is time-consuming and prone to mistakes	Real-world video traffic surveillance datasets are used to run the model, and both qualitatively and quantitatively substantial results have been obtained	The code may be implemented on PYNQ hardware in the future to process video frames more quickly for anomaly identification. The application of active learning to identify anomalies might also be the focus of the study
[53]	2022	The researcher proposed the Deep Residual Spatiotemporal Translation Network (DR-STN), an innovative unsupervised Deep Residual conditional Generative Adversarial Network (DR-cGAN) architecture using an online hard negative mining (OHNM) algorithm	The frame-level evaluation for the three benchmarks has an average AUC score of 96.73%. Between DR-STN and cutting-edge techniques, there is a 7.6% improvement in AUC at the frame level	The author's future work will concentrate on ongoing learning of unknown events, helping to determine if they are truly aberrant or merely unusual typical happenings
[54]	2022	This paper proposed a cheating detection system to deal with plagiarism and other forms of academic dishonesty	The authority is informed of the unusual conduct by an automated alarm, reducing the possibility of error that may arise from manual monitoring	
[55]	2023	In this study, researchers proposed a weakly supervised deep temporal encoding-decoding approach based on multiple instances of learning for anomaly detection in surveillance videos	The results show that the recommended method works as well as or is superior to the state-of-the-art techniques for detecting anomalies in video surveillance uses, achieving a state-of-the-art false alarm rate on the UCF-crime dataset	

Continuation of Table 4

Ref no.	Year	Technique	Significance	Limitation/Future scope
[56]	2023	The suggested method efficiently uses both geographic with temporal information by adopting a geographical branch with a temporal branch in a single network	The outcomes show that the network surpasses state-of-the-art techniques, obtaining, in terms of Area Under Curve, 97.4% for UCSD Ped2, 86.7% for CUHK Avenue, and 73.6% for the Shanghai Tech dataset	Practical surveillance needs to improve in the future
[57]	2023	This study demonstrated the creation of an automated safety mechanism that can quickly assist victims and identify suspicious activity in real time	On the test set database, the suggested approach's AUC was equal to 94.21%, and the detection accuracy was equivalent to 88.46%	Future studies will examine new feature extraction theories, feature selection strategies, and decreasing dimensionality approaches to increase the precision of the indicator
[58]	2023	This paper describes Ancilia, an end-to-end scalable, intelligent video surveillance platform for the IoT	To create safer and more secure communities, Ancilia intends to change the surveillance environment by introducing more efficient, intelligent, and fair security to the sector without asking individuals to give up their right to privacy	Future studies will examine new feature extraction theories, feature selection strategies, and decreasing dimensionality approaches to increase the precision of the indicator
[59]	2023	The author of this work using isolation tree-based unsupervised clustering divides the deep feature space of the video segments	According to experimental findings, the suggested framework outperforms state-of-the-art video anomaly detection techniques in terms of accuracy	Data training and quality must be improved in the future
[60]	2023	In video surveillance, finding frames that differed noticeably from the norm was the aim of anomaly detection. To solve this problem, the author created a unique bi-directional frame interpolation-based video anomaly recognition framework	The recommended method's value was confirmed by the excellent frame-level video anomaly detection results on open benchmarks	The suggested method's key is to interpolate normal frames with little to no mistakes, but aberrant frames with significant errors

Anomaly Recognition. In study [51] the authors recommended the Anomaly Net neural network as a unique neural network for anomaly recognition because it combines feature learning, sparse representation, and dictionary learning in three joint neural processing blocks. To address the shortcomings of existing sparse coding optimizers, the researchers developed a special RNN to learn sparse representation with a sparse representation dictionary. Numerous trials demonstrate the method's cutting-edge performance in the task of detecting abnormal occurrences.

In study [52] the authors suggested that to monitor and respond appropriately in the event of tragic incidents, real-world traffic surveillance recordings need constant oversight. However, it is time-consuming and error-prone to oversee them continually with humans. As a result, a DL method for automatically detecting and localizing traffic accidents has been suggested by redefining the issue as anomaly finding. The technique uses sequence-to-sequence LSTM autoencoder and spatiotemporal autoencoder to model spatial and temporal representations in the video. Additionally, it employs a one-class categorization scheme. Real-world video traffic surveillance datasets are being used to apply the methodology, and both subjectively and numerically useful outcomes were achieved.

Paper [53] offered that the Deep Residual Spatiotemporal Translation Network (DR-STN) is a unique unsupervised Deep Residual Conditional Generative Adversarial Network (DR-cGAN) system that employs an online hard negative mining (OHNM) technique. It expands the network available for finding a mapping from spatial to temporal memories thus raising the perceived calibre of artificially created images. It has thoroughly tested against publicly accessible benchmarks and has outperformed other cutting-edge techniques. The difference in AUC between DR-STN and cutting-edge techniques at the frame level is 7.6%.

In study [54] the authors suggested a cheating detection method to deal with plagiarism and other types of academic dishonesty. During examinations, the system employs video monitoring to keep an eye on student behavior, particularly unusual behavior. The system employs three distinct methods: calculating the direction of the students' heads as they turn from their starting orientation, seeing pupil movement, and recognizing the point at which a student's hands come into touch with their faces. An automated alarm that informs the appropriate authority when any of these are found helps to reduce the possibility of error that may result from manual monitoring.

In paper [55] the authors proposed a weakly supervised deep temporal encoding-decoding approach employing multiple instances of learning for anomaly detection in surveillance videos. The proposed approach makes use of a deep temporal encoding-decoding network to record the spatiotemporal evolution of video instances across time while training using both abnormal and typical video clips. Low false alarm rates are produced by the suggested loss function, which optimizes the mean separation between predictions for normal and abnormal instance types. On the UCF-crime dataset, the suggested technique obtains a state-of-the-art false alert rate when compared to cutting-edge procedures.

Study [56] adopts a geographical branch and a temporal branch in a single network, effectively using both geographic and temporal information. It has a residual auto-encoder structure that is comprised of a deep CNN-powered encoder and a multi-stage channel attention-based decoder. System performance is estimated by utilizing three standard benchmark datasets: UCSD Ped2 (97.4%), CUHK Avenue (86.7%), and ShanghaiTech (73.6%).

Paper [57] demonstrated the creation of an autonomous security mechanism that can quickly assist victims in recognizing suspicious activity in real time. It utilizes an adaptive method based on DL (DL), PCA, and machine learning (ML). The suggested method has an experimented AUC and detection accuracy of 88.46% on the UCF-crime dataset. When compared to previously constructed systems, the suggested solution has proven to be accurate and resilient.

Study [58] introduces Ancilia, an end-to-end scalable, intelligent video surveillance solution for the IoT. Ancilia utilizes cutting-edge artificial intelligence for practical surveillance applications while upholding moral considerations and executing complex cognitive operations in real time. By bringing more efficient, intelligent, and fair security to the field, Ancilia hopes to change the surveillance environment and create safer and more secure communities without asking individuals to give up their right to privacy.

In paper [59] the deep feature space of video clips was divided using isolation tree-based unsupervised grouping. A pseudo anomaly score is produced by the RGB- -stream, while a pseudo dynamicity score, is produced by the flow stream. The majority voting method is employed to combine these scores and generate initial bags of beneficial and negative parts. Both scores are refined using a segment re-mapping and cross-branch feed-forward network refinement approach. According to experimental findings, the suggested framework

outperforms cutting-edge video anomaly identification techniques in terms of accuracy.

Paper [60] offered that in surveillance footage, finding frames that differed noticeably from the norm was the aim of identifying anomalies. The investigators used bi-directional frame interpolation to develop a brand-new model for video anomaly identification to address this issue. The proposed system includes a unique dynamic memory technique to balance memory scarcity with normality presentation variance, along with an optical flow estimating network with an interpolation system that has both been collaboratively optimized end-to-end. Numerous tests on widely used benchmarks show how much better the proposed framework is than existing solutions.

In common, the computationally and time-intensive phase of feature representation through video violence recognition acts as a substantial impediment to the deployment of violence detection in practical applications. The vast number of technologies as given in Table 5 is now in use to detect violent material in video and thus have substantial time and space complexity costs. These methods are therefore inappropriate for usage in real-world applications. As a result, denser and deeper DNN models are needed for better feature extraction and description. There is also a need for quicker, easier, and more accurate ways to identify violence. The high dimensional structure and a non-local shift among frames, nevertheless, make it more challenging for the methods employed to identify aberrant video.

3. Summary. After studying different paradigms of human behavior and activity recognition in video surveillance systems the following conclusions arrive which must be concentrated on in future research:

– The review of the literature reveals that certain methodologies aim to ignore the background and concentrate exclusively on foreground characteristics for anomaly identification. We believe that background knowledge might be helpful to simulate potential event-causing situations.

– The potential of human activity and behavior identification employing CNN, Deep Learning, LSTM, and GAN is bright since they provide more accuracy in most surveillance settings.

– In addition, the performance of the anomaly detection technique depends on the crowd density; as the crowd grows, its effectiveness declines, and it performs best in sparse crowds.

– In addition, while only benchmark data sets could be used for comparison, they might not be adequate to account for all real-world events.

– Edge computing is a potential strategy for delay-sensitive applications like intelligent surveillance and anomaly recognition. Since the data is handled on the device itself, it provides greater privacy and security. The burden is distributed through job offloading and ongoing improvement in edge devices, increasing total efficiency. Edge computing and human behavior and activity identification together will open up new opportunities for computer vision.

– From the research, we can find that more concentration must be provided in the arena of human action identification and anomaly detection which is the most crucial need in the current video surveillance.

– Although every group may be employed in a supervised or unsupervised way, the assessment reveals that the majority of investigators employed unsupervised learning to address the challenge of recognizing human conduct and activities since there was a dearth of huge datasets.

Table 5 compares various deep learning techniques used in detecting human activities across different applications. Convolutional Neural Networks (CNNs) excel in identifying faces, leveraging robust feature extraction and spatial hierarchies. Recurrent Neural Networks (RNNs) are pivotal in emotion identification, adept at modeling sequential data and capturing temporal dependencies. Long Short-Term Memory Networks (LSTMs) shine in action identification, effectively handling long-range dependencies in sequences. Autoencoders prove valuable in anomaly identification by self-supervised learning, although they may be sensitive to hyperparameters. Generative Adversarial Networks (GANs) show promise in generating synthetic data for anomaly detection, but their training stability can be challenging. Capsule Networks (CapsNets) offer improved handling of spatial hierarchies and resistance to certain adversarial attacks, though they're still underexplored. Lastly, Transfer Learning is a versatile approach applicable across all subcategories, leveraging pre-trained models to reduce the need for extensive data and accelerate training, but it may require task-specific fine-tuning for optimal performance.

Table 5. Comparison chart based on various deep learning methods

Technique	**Used for**	**Signifi-cance**	**Advantages**	**Limitations**	**Future Recommen-dations**
Convolutional Neural Networks (CNNs)	Identifying faces	Highly effective in image recognition tasks	Robust feature extraction, spatial hierarchies	Limited to fixed-size inputs, may struggle with occlusions	Investigate multi-scale architectures for better adaptability
Recurrent Neural Networks (RNNs)	Emotion identification	Sequential data modeling	Captures temporal dependencies, variable-length sequences	Prone to vanishing / exploding gradients, computationally intensive	Explore variants like LSTMs, and GRUs for improved efficiency
Long Short-Term Memory Networks (LSTMs)	Action identification	Handling sequential data	Effective for modeling long-range dependencies, avoids vanishing gradient	Computationally expensive, harder to interpret	Investigate attention mechanisms for better context modeling
Autoencoders	Anomaly identification	Anomaly detection in unlabeled data	Self-supervised learning, robust to noisy data	Sensitive to choice of hyperparameters, may require large datasets	Explore unsupervised pre-training for improved anomaly detection
Generative Adversarial Networks (GANs)	Anomaly identification	Generate synthetic data for anomaly detection	Effective in generating realistic data distributions	Training instability, mode collapse	Investigate techniques for stable GAN training, utilize in semi-supervised setups
Capsule Networks (CapsNets)	Identifying faces, Emotion identification	Improved handling of spatial hierarchies	Resistant to certain types of adversarial attacks	Limited adoption, computationally intensive	Investigate hybrid architectures with CNNs for improved performance
Transfer Learning	Across all subcategories	Utilizing pre-trained models for specific tasks	Reduces the need for large datasets, faster training	May not always transfer well, task-specific fine-tuning needed	Investigate techniques for better model adaptation in transfer learning scenarios

4. Future recommendation. Advancements in CNNs, Deep Learning, LSTM, and GANs hold great promise for improving accuracy in surveillance. Further exploration of novel architectures and techniques, such as attention mechanisms and multi-modal integration, could significantly enhance human activity identification. Some more recommendations are given below based on the state of the art:

– Future research could explore hybrid approaches that incorporate both foreground and background information, leveraging the potential benefits of simulating event-causing situations with background knowledge.

– Investigate novel architectures and techniques within CNNs, Deep Learning, LSTM, and GANs to further enhance accuracy in surveillance settings, possibly by incorporating multi-modal information or attention mechanisms.

– Develop adaptive anomaly detection techniques that dynamically adjust their sensitivity based on crowd density, potentially utilizing reinforcement learning or adaptive thresholding mechanisms.

– Encourage the creation of more diverse and realistic benchmark datasets that capture a broader range of real-world events, potentially through crowdsourcing or incorporating simulated data augmentation techniques.

– Investigate techniques for optimizing and accelerating anomaly detection algorithms specifically for edge computing environments, possibly through model compression, quantization, or specialized hardware acceleration.

– Allocate research efforts towards developing specialized models and algorithms dedicated to human action identification and anomaly detection, possibly exploring novel architectures or incorporating domain-specific knowledge.

– Encourage the creation of larger annotated datasets to support the application of supervised learning approaches. Additionally, explore techniques for semi-supervised learning that leverage limited labeled data with a larger pool of unlabeled data.

– These future directions aim to address specific areas of improvement and expansion within the field of human activity identification and anomaly detection, ultimately advancing the capabilities and effectiveness of surveillance systems.

5. Conclusion. This article examines DL-based methods for video surveillance that span a range of techniques and approaches for recognizing human behavior and activities. Readers should ideally be able to appreciate not just the justification for employing a particular technique, but also to

compare several approaches, generate a comparative analysis, and suggest a strategy after reading a complete overview of human behavior and activity identification. First, we divided the methods into four groups, based on how well they could identify faces, emotions, actions, and anomalies. Additionally, we listed every category's advantages and disadvantages in accordance. Future work on the DL model should focus on studying human action and emotion identification, which can improve situational knowledge for targeted video surveillance.

References

1. Zhang J., Zi L., Hou Y., Wang M., Jiang W., Deng D. A DL-based approach to enable action recognition for construction equipment. Advances in Civil Engineering. 2020. pp. 1–14.
2. Wang X., Che Z., Jiang B., Xiao N., Yang K., Tang J., Ye J., Wang J., Qi Q. Robust unsupervised video anomaly detection by multipath frame prediction. IEEE transactions on neural networks and learning systems. 2021. vol. 33. no. 6. pp. 2301–2312.
3. Zhang H.B., Zhang Y.X., Zhong B., Lei Q., Yang L., Du J.X., Chen D.S. A comprehensive survey of vision-based human action recognition methods. Sensors. 2019. vol. 19(5). no. 1005.
4. Pervaiz M., Jalal A., Kim K. A hybrid algorithm for multi-people counting and tracking for smart surveillance. International Bhurban conference on applied sciences and technologies (IBCAST). 2021. pp. 530–535.
5. Kong Y., Fu Y. Human action recognition and prediction: A survey. International Journal of Computer Vision. 2022. vol. 130(5). pp. 1366–1401.
6. Franco A., Magnani A., Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognition Letters. 2020. vol. 131. pp. 293–299.
7. Wang L., Huynh D.Q., Koniusz P. A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing. 2019. vol. 29. pp. 15–28.
8. Zhou X., Liang W., Kevin I., Wang K., Wang H., Yang L.T., Jin Q. Deep-learning-enhanced human activity recognition for the Internet of Healthcare things. IEEE Internet of Things Journal. 2020. vol. 7(7). pp. 6429–6438.
9. Qiu Z., Yao T., Ngo C.W., Tian X., Mei T. Learning spatio-temporal representation with local and global diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 12056–12065.
10. Sreenu G., Durai S. Intelligent video surveillance: a review through DL techniques for crowd analysis. Journal of Big Data. 2019. vol. 6(1). pp. 1–27.
11. Elharrouss O., Almaadeed N., Al-Maadeed S., Bouridane A., Beghdadi A. A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence. 2021. vol. 51. pp. 690–712.
12. Jaouedi N., Boujnah N., Bouhlel M.S. A new hybrid DL model for human action recognition. Journal of King Saud University – Computer and Information Sciences. 2020. vol. 32. no. 4. pp. 447–453.
13. Dang L.M., Min K., Wang H., Piran M.J., Lee C.H., Moon H. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition. 2020. vol. 108. no. 107561.

14. Saeed A., Ozcelebi T., Lukkien J. Multi-task self-supervised learning for human activity detection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2019. vol. 3(2). pp. 1–30.
15. Fu B., Damer N., Kirchbuchner F., Kuijper A. Sensing technology for human activity recognition: A comprehensive survey. IEEE Access. 2020. vol. 8. pp. 83791–83820.
16. du Toit J., du Toit T, Kruger H. Heuristic Data Augmentation for Improved Human Activity Recognition. Proceedings of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC). 2019. pp. 264–269.
17. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on DL-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2021. pp. 1–17.
18. Concone F., Re G.L., Morana M. A fog-based application for human activity recognition using personal smart devices. ACM Transactions on Internet Technology (TOIT). 2019. vol. 19(2). pp. 1–20.
19. He J.Y., Wu X., Cheng Z.Q., Yuan Z., Jiang Y.G. DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. Neurocomputing. 2021. vol. 444. pp. 319–331.
20. Beddiar D.R., Nini B., Sabokrou M., Hadid A. Vision-based human activity recognition: a survey. Multimedia Tools and Applications. 2020. vol. 79. no. 41-42. pp. 30509–30555.
21. Chen J., Li K., Deng Q., Li K., Philip S.Y. Distributed DL model for intelligent video surveillance systems with edge computing. IEEE Transactions on Industrial Informatics. 2019. DOI: 10.1109/TII.2019.2909473.
22. Zhao Y., Shen X., Jin Z., Lu H., Hua X.S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 4913–4922.
23. Kaur G., Sinha R., Tiwari P.K., Yadav S.K., Pandey P., Raj R., Vashisth A., Rakhra M. Face mask recognition system using CNN model. Neuroscience Informatics. 2021. vol. 2(3). no. 100035. DOI:10.1016/j.neuri.2021.100035.
24. Wang Y., Yue Y., Lin Y., Jiang H., Lai Z., Kulikov V., Huang G. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). 2022. pp. 20030–20040.
25. Goyal H., Sidana K., Singh C., Jain A., Jindal S. A real-time face mask detection system using a convolutional neural network. Multimedia Tools and Applications. 2022. vol. 81(11). pp. 14999–15015.
26. Sayeed A., Srizon A.Y., Hasan M.M., Shin J., Hasan M.A.M., Mahmud M.R. A Hybrid Campus Security System Combined Face, Number-Plate, and Voice Recognition. International Conference on Recent Trends in Image Processing and Pattern Recognition. 2022. pp. 356–368.
27. Kumar B.A., Bansal M. Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning. Applied Sciences. 2023. vol. 13(2). no. 935.
28. Kamyab T., Daealhaq H., Ghahfarokhi A.M., Beheshtinejad F., Salajegheh E. Combination of Genetic Algorithm and Neural Network to Select Facial Features in Face Recognition Technique. International Journal of Robotics and Control Systems. 2023. vol. 3(1). pp. 50–58.
29. Singh A., Bhatt S., Nayak V., Shah M. Automation of surveillance systems using DL and facial recognition. International Journal of System Assurance Engineering and Management. 2023. vol. 14. pp. 236–245.

30. Terhorst P., Ihlefeld M., Huber M., Damer N., Kirchbuchner F., Raja K., Kuijper A. Qmagface: Simple and accurate quality-aware face recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. 3484–3494.
31. Wang K., Peng X., Yang J., Meng D., Qiao Y. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 4057–4069.
32. Hossain M.S., Muhammad G. Emotion recognition using DL approach from audio–visual emotional big data. Information Fusion. 2019. vol. 49. pp. 69–78.
33. Kanjo E., Younis E.M., Ang C.S. DL analysis of mobile physiological, environmental, and location sensor data for emotion detection. Information Fusion. 2019. vol. 49. pp. 46–56.
34. Wang K., Peng X., Yang J., Lu S., Qiao Y. Suppressing uncertainties for large-scale facial expression recognition. Proceedings of the IEEE/CVF computer vision and pattern recognition. 2020. pp. 6897–6906.
35. Minaee S., Minaei, M., Abdolrashidi A. Deep-emotion: Facial expression recognition using the attentional convolutional network. Sensors. 2021. vol. 21(9). no. 3046.
36. Umer S., Rout R.K., Pero C., Nappi M. Facial expression recognition with trade-offs between data augmentation and DL features. Journal of Ambient Intelligence and Humanized Computing. 2022. pp. 1–15.
37. Zhang W., Qiu F., Wang S., Zeng H., Zhang Z., An R., Ma B., Ding Y. Transformer-based multimodal information fusion for facial expression analysis. Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition. 2022. pp. 2428–2437.
38. Zhu X., Li Z., Sun J. Expression recognition method combining convolutional features and Transformer. Mathematical Foundations of Computing. 2023. vol. 6. no. 2. pp. 203–217.
39. Bapat M.M., Patil C.H., Mali S.M. Database Development and Recognition of Facial Expression using DL. 2023. 20 p. DOI: 10.21203/rs.3.rs-2477808/v1.
40. Mukhiddinov M., Djuraev O., Akhmedov F., Mukhamadiyev A., Cho J. Masked Face Emotion Recognition Based on Facial Landmarks and DL Approaches for Visually Impaired People. Sensors. 2023. vol. 23(3). no. 1080.
41. Xia K., Huang J., Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access. 2020. vol. 8. pp. 56855–56866.
42. Dhiman C., Vishwakarma D.K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 3835–3844.
43. Paoletti G., Cavazza J., Beyan C., Del Bue A. Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance. 2022. arXiv preprint arXiv:2204.10312.
44. Sanchez-Caballero A., de Lopez-Diz S., Fuentes-Jimenez D., Losada-Gutiérrez C., Marrón-Romera M., Casillas-Perez D., Sarker M.I. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. Multimedia Tools and Applications. 2022. vol. 81. no. 17. pp. 24119–24143.
45. Khan I.U., Afzal S., Lee J.W. Human activity recognition via hybrid DL-based model. Sensors. 2022. vol. 22(1). no. 323.
46. Yadav S.K., Tiwari K., Pandey H.M., Akbar S.A. Skeleton-based human activity recognition using Conv LSTM and guided feature learning. Soft Computing. 2022. pp. 1–14.
47. Zhu Q., Deng H. Spatial adaptive graph convolutional network for skeleton-based action recognition. Applied Intelligence. 2023. pp. 1–13.
48. Singh G., Choutas V., Saha S., Yu F., Van Gool L. Spatio-Temporal Action Detection under Large Motion. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 6009–6018.

49. Ahn D., Kim S., Hong H., Ko B.C. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 3330–3339.
50. Peng K., Roitberg A., Yang K., Zhang J., Stiefelhagen R. Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions. IEEE Transactions on Multimedia. 2023. arXiv preprint arXiv:2202.11423v3.
51. Zhou J.T., Du J., Zhu H., Peng X., Liu Y., Goh R.S.M. Anomalynet: An anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security. 2019. vol. 14(10). pp. 2537–2550.
52. Pawar K., Attar V. DL-based detection and localization of road accidents from traffic surveillance videos. ICT Express. 2022. vol. 8. no. 3. pp. 379–387.
53. Ganokratanaa T., Aramvith S., Sebe N. Video anomaly detection using deep residual-spatiotemporal translation network. Pattern Recognition Letters. 2022. vol. 155. pp. 143–150.
54. Roa'a M., Aljazaery I.A., ALRikabi H.T.S., Alaidi A.H.M. Automated Cheating Detection Based on Video Surveillance in the Examination Classes. iJIM. 2022. vol. 16(08). no. 125.
55. Kamoona A.M., Gostar A.K., Bab-Hadiashar A., Hoseinnezhad R. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. Expert Systems with Applications. 2023. vol. 214. no. 119079. DOI: 10.1016/j.eswa.2022.119079.
56. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
57. Abbas Z.K., Al-Ani A.A. An adaptive algorithm based on principal component analysis-DL for anomalous events detection. Indonesian Journal of Electrical Engineering and Computer Science. 2023. vol. 29(1). pp. 421–430.
58. Pazho A.D., Neff C., Noghre G.A., Ardabili B.R., Yao S., Baharani M., Tabkhi H. Ancilia: Scalable Intelligent Video Surveillance for the Artificial Intelligence of Things. 2023. arXiv preprint arXiv:2301.03561.
59. Thakare K.V., Raghuwanshi Y., Dogra D.P., Choi H., Kim I.J. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 5541–5550.
60. Deng H., Zhang Z., Zou S., Li X. Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 2634–2643.

Nukala Sujata Gupta — Research scholar, Department of computer science and engineering, Koneru Lakshmaiah Education Foundation. Research interests: science and engineering. gsuji29@gmail.com; Green Fields, Vaddeswaram, 522302, Guntur, Andhra Pradesh, India; office phone: +91(8645)350-0200.

Ramya K. Ruth — Associate professor, Department of computer science and engineering, Koneru Lakshmaiah Education Foundation. Research interests: science and engineering. The number of publications — 12. ramya_cse@kluniversity.in; Green Fields, Vaddeswaram, 522302, Guntur, Andhra Pradesh, India; office phone: +91(8645)350-0200.

Karnati Ramesh — Associate professor, Department of computer science and engineering, Vardhaman College of Engineering. Research interests: data mining, machine learning, artificial intelligence, IoT. The number of publications — 17. ramesh.krnt@vardhaman.org; Kacharam, Shamshabad, 501218, Hyderabad, Telangana, India; office phone: +91(8688)901-557.

УДК 004.7 **DOI** 10.15622/ia.23.2.5

Н. Суджата Гупта, К.Р. Рамья, Р. Карнати

РАСПОЗНАВАНИЕ ДЕЙСТВИЙ ЧЕЛОВЕКА В СИСТЕМАХ ВИДЕОНАБЛЮДЕНИЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ – ОБЗОР

Суджата Гупта Н., Рамья К.Р., Карнати Р. **Распознавание действий человека в системах видеонаблюдения с использованием методов глубокого обучения – обзор.**

Аннотация. Несмотря на широкое применение во многих областях, точная и эффективная идентификация деятельности человека продолжает оставаться интересной исследовательской проблемой в области компьютерного зрения. В настоящее время проводится много исследований по таким темам, как распознавание активности пешеходов и способы распознавания движений людей с использованием данных глубины, трехмерных скелетных данных, данных неподвижных изображений или стратегий, использующих пространственно-временные точки интереса. Это исследование направлено на изучение и оценку подходов DL для обнаружения человеческой активности на видео. Основное внимание было уделено нескольким структурам для обнаружения действий человека, которые используют DL в качестве своей основной стратегии. В зависимости от приложения, включая идентификацию лиц, идентификацию эмоций, идентификацию действий и идентификацию аномалий, прогнозы появления людей разделены на четыре различные подкатегории. В литературе было проведено несколько исследований, основанных на этих распознаваниях для прогнозирования поведения и активности человека в приложениях видеонаблюдения. Сравнивается современное состояние методов DL для четырех различных приложений. В этой статье также представлены области применения, научные проблемы и потенциальные цели в области распознавания человеческого поведения и активности на основе DL.

Ключевые слова: распознавание лиц, распознавание эмоций, распознавание действий, распознавание аномалий, DL, распознавание человеческого поведения и активности /обнаружение.

Литература

1. Zhang J., Zi L., Hou Y., Wang M., Jiang W., Deng D. A DL-based approach to enable action recognition for construction equipment. Advances in Civil Engineering. 2020. pp. 1–14.
2. Wang X., Che Z., Jiang B., Xiao N., Yang K., Tang J., Ye J., Wang J., Qi Q. Robust unsupervised video anomaly detection by multipath frame prediction. IEEE transactions on neural networks and learning systems. 2021. vol. 33. no. 6. pp. 2301–2312.
3. Zhang H.B., Zhang Y.X., Zhong B., Lei Q., Yang L., Du J.X., Chen D.S. A comprehensive survey of vision-based human action recognition methods. Sensors. 2019. vol. 19(5). no. 1005.
4. Pervaiz M., Jalal A., Kim K. A hybrid algorithm for multi-people counting and tracking for smart surveillance. International Bhurban conference on applied sciences and technologies (IBCAST). 2021. pp. 530–535.
5. Kong Y., Fu Y. Human action recognition and prediction: A survey. International Journal of Computer Vision. 2022. vol. 130(5). pp. 1366–1401.

6. Franco A., Magnani A., Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognition Letters. 2020. vol. 131. pp. 293–299.
7. Wang L., Huynh D.Q., Koniusz P. A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing. 2019. vol. 29. pp. 15–28.
8. Zhou X., Liang W., Kevin I., Wang K., Wang H., Yang L.T., Jin Q. Deep-learning-enhanced human activity recognition for the Internet of Healthcare things. IEEE Internet of Things Journal. 2020. vol. 7(7). pp. 6429–6438.
9. Qiu Z., Yao T., Ngo C.W., Tian X., Mei T. Learning spatio-temporal representation with local and global diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 12056–12065.
10. Sreenu G., Durai S. Intelligent video surveillance: a review through DL techniques for crowd analysis. Journal of Big Data. 2019. vol. 6(1). pp. 1–27.
11. Elharrouss O., Almaadeed N., Al-Maadeed S., Bouridane A., Beghdadi A. A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence. 2021. vol. 51. pp. 690–712.
12. Jaouedi N., Boujnah N., Bouhlel M.S. A new hybrid DL model for human action recognition. Journal of King Saud University – Computer and Information Sciences. 2020. vol. 32. no. 4. pp. 447–453.
13. Dang L.M., Min K., Wang H., Piran M.J., Lee C.H., Moon H. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition. 2020. vol. 108. no. 107561.
14. Saeed A., Ozcelebi T., Lukkien J. Multi-task self-supervised learning for human activity detection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2019. vol. 3(2). pp. 1–30.
15. Fu B., Damer N., Kirchbuchner F., Kuijper A. Sensing technology for human activity recognition: A comprehensive survey. IEEE Access. 2020. vol. 8. pp. 83791–83820.
16. du Toit J., du Toit T, Kruger H. Heuristic Data Augmentation for Improved Human Activity Recognition. Proceedings of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC). 2019. pp. 264–269.
17. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on DL-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2021. pp. 1–17.
18. Concone F., Re G.L., Morana M. A fog-based application for human activity recognition using personal smart devices. ACM Transactions on Internet Technology (TOIT). 2019. vol. 19(2). pp. 1–20.
19. He J.Y., Wu X., Cheng Z.Q., Yuan Z., Jiang Y.G. DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. Neurocomputing. 2021. vol. 444. pp. 319–331.
20. Beddiar D.R., Nini B., Sabokrou M., Hadid A. Vision-based human activity recognition: a survey. Multimedia Tools and Applications. 2020. vol. 79. no. 41-42. pp. 30509–30555.
21. Chen J., Li K., Deng Q., Li K., Philip S.Y. Distributed DL model for intelligent video surveillance systems with edge computing. IEEE Transactions on Industrial Informatics. 2019. DOI: 10.1109/TII.2019.2909473.
22. Zhao Y., Shen X., Jin Z., Lu H., Hua X.S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 4913–4922.

23. Kaur G., Sinha R., Tiwari P.K., Yadav S.K., Pandey P., Raj R., Vashisth A., Rakhra M. Face mask recognition system using CNN model. Neuroscience Informatics. 2021. vol. 2(3). no. 100035. DOI:10.1016/j.neuri.2021.100035.
24. Wang Y., Yue Y., Lin Y., Jiang H., Lai Z., Kulikov V., Huang G. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). 2022. pp. 20030–20040.
25. Goyal H., Sidana K., Singh C., Jain A., Jindal S. A real-time face mask detection system using a convolutional neural network. Multimedia Tools and Applications. 2022. vol. 81(11). pp. 14999–15015.
26. Sayeed A., Srizon A.Y., Hasan M.M., Shin J., Hasan M.A.M., Mahmud M.R. A Hybrid Campus Security System Combined Face, Number-Plate, and Voice Recognition. International Conference on Recent Trends in Image Processing and Pattern Recognition. 2022. pp. 356–368.
27. Kumar B.A., Bansal M. Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning. Applied Sciences. 2023. vol. 13(2). no. 935.
28. Kamyab T., Daealhaq H., Ghahfarokhi A.M., Beheshtinejad F., Salajegheh E. Combination of Genetic Algorithm and Neural Network to Select Facial Features in Face Recognition Technique. International Journal of Robotics and Control Systems. 2023. vol. 3(1). pp. 50–58.
29. Singh A., Bhatt S., Nayak V., Shah M. Automation of surveillance systems using DL and facial recognition. International Journal of System Assurance Engineering and Management. 2023. vol. 14. pp. 236–245.
30. Terhorst P., Ihlefeld M., Huber M., Damer N., Kirchbuchner F., Raja K., Kuijper A. Qmagface: Simple and accurate quality-aware face recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. 3484–3494.
31. Wang K., Peng X., Yang J., Meng D., Qiao Y. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 4057–4069.
32. Hossain M.S., Muhammad G. Emotion recognition using DL approach from audio–visual emotional big data. Information Fusion. 2019. vol. 49. pp. 69–78.
33. Kanjo E., Younis E.M., Ang C.S. DL analysis of mobile physiological, environmental, and location sensor data for emotion detection. Information Fusion. 2019. vol. 49. pp. 46–56.
34. Wang K., Peng X., Yang J., Lu S., Qiao Y. Suppressing uncertainties for large-scale facial expression recognition. Proceedings of the IEEE/CVF computer vision and pattern recognition. 2020. pp. 6897–6906.
35. Minaee S., Minaei, M., Abdolrashidi A. Deep-emotion: Facial expression recognition using the attentional convolutional network. Sensors. 2021. vol. 21(9). no. 3046.
36. Umer S., Rout R.K., Pero C., Nappi M. Facial expression recognition with trade-offs between data augmentation and DL features. Journal of Ambient Intelligence and Humanized Computing. 2022. pp. 1–15.
37. Zhang W., Qiu F., Wang S., Zeng H., Zhang Z., An R., Ma B., Ding Y. Transformer-based multimodal information fusion for facial expression analysis. Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition. 2022. pp. 2428–2437.
38. Zhu X., Li Z., Sun J. Expression recognition method combining convolutional features and Transformer. Mathematical Foundations of Computing. 2023. vol. 6. no. 2. pp. 203–217.
39. Bapat M.M., Patil C.H., Mali S.M. Database Development and Recognition of Facial Expression using DL. 2023. 20 p. DOI: 10.21203/rs.3.rs-2477808/v1.

40. Mukhiddinov M., Djuraev O., Akhmedov F., Mukhamadiyev A., Cho J. Masked Face Emotion Recognition Based on Facial Landmarks and DL Approaches for Visually Impaired People. Sensors. 2023. vol. 23(3). no. 1080.
41. Xia K., Huang J., Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access. 2020. vol. 8. pp. 56855–56866.
42. Dhiman C., Vishwakarma D.K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 3835–3844.
43. Paoletti G., Cavazza J., Beyan C., Del Bue A. Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance. 2022. arXiv preprint arXiv:2204.10312.
44. Sanchez-Caballero A., de Lopez-Diz S., Fuentes-Jimenez D., Losada-Gutiérrez C., Marrón-Romera M., Casillas-Perez D., Sarker M.I. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. Multimedia Tools and Applications. 2022. vol. 81. no. 17. pp. 24119–24143.
45. Khan I.U., Afzal S., Lee J.W. Human activity recognition via hybrid DL-based model. Sensors. 2022. vol. 22(1). no. 323.
46. Yadav S.K., Tiwari K., Pandey H.M., Akbar S.A. Skeleton-based human activity recognition using Conv LSTM and guided feature learning. Soft Computing. 2022. pp. 1–14.
47. Zhu Q., Deng H. Spatial adaptive graph convolutional network for skeleton-based action recognition. Applied Intelligence. 2023. pp. 1–13.
48. Singh G., Choutas V., Saha S., Yu F., Van Gool L. Spatio-Temporal Action Detection under Large Motion. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 6009–6018.
49. Ahn D., Kim S., Hong H., Ko B.C. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 3330–3339.
50. Peng K., Roitberg A., Yang K., Zhang J., Stiefelhagen R. Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions. IEEE Transactions on Multimedia. 2023. arXiv preprint arXiv:2202.11423v3.
51. Zhou J.T., Du J., Zhu H., Peng X., Liu Y., Goh R.S.M. Anomalynet: An anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security. 2019. vol. 14(10). pp. 2537–2550.
52. Pawar K., Attar V. DL-based detection and localization of road accidents from traffic surveillance videos. ICT Express. 2022. vol. 8. no. 3. pp. 379–387.
53. Ganokratanaa T., Aramvith S., Sebe N. Video anomaly detection using deep residual-spatiotemporal translation network. Pattern Recognition Letters. 2022. vol. 155. pp. 143–150.
54. Roa'a M., Aljazaery I.A., ALRikabi H.T.S., Alaidi A.H.M. Automated Cheating Detection Based on Video Surveillance in the Examination Classes. iJIM. 2022. vol. 16(08). no. 125.
55. Kamoona A.M., Gostar A.K., Bab-Hadiashar A., Hoseinnezhad R. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. Expert Systems with Applications. 2023. vol. 214. no. 119079. DOI: 10.1016/j.eswa.2022.119079.
56. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
57. Abbas Z.K., Al-Ani A.A. An adaptive algorithm based on principal component analysis-DL for anomalous events detection. Indonesian Journal of Electrical Engineering and Computer Science. 2023. vol. 29(1). pp. 421–430.

58. Pazho A.D., Neff C., Noghre G.A., Ardabili B.R., Yao S., Baharani M., Tabkhi H. Ancilia: Scalable Intelligent Video Surveillance for the Artificial Intelligence of Things. 2023. arXiv preprint arXiv:2301.03561.
59. Thakare K.V., Raghuwanshi Y., Dogra D.P., Choi H., Kim I.J. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 5541–5550.
60. Deng H., Zhang Z., Zou S., Li X. Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 2634–2643.

Суджата Гупта Нукала — научный сотрудник, факультет компьютерных наук и инженерии, Образовательный фонд Конеру Лакшмайи. Область научных интересов: наука и техника. gsuji29@gmail.com; Зеленые поля, Ваддесварам, 522302, Гунтур, Андхра-Прадеш, Индия; р.т.: +91(8645)350-0200.

Рамья К. Рут — доцент, факультет компьютерных наук и инженерии, Образовательный фонд Конеру Лакшмайи. Область научных интересов: наука и техника. Число научных публикаций — 12. ramya_cse@kluniversity.in; Зеленые поля, Ваддесварам, 522302, Гунтур, Андхра-Прадеш, Индия; р.т.: +91(8645)350-0200.

Карнати Рамеш — доцент, факультет компьютерных наук и инженерии, Инженерный колледж Вардхамана. Область научных интересов: интеллектуальный анализ данных, машинное обучение, искусственный интеллект, интернет вещей. Число научных публикаций — 17. ramesh.krnt@vardhaman.org; Качарам, Шамшабад, 501218, Хайдарабад, Телангана, Индия; р.т.: +91(8688)901-557.