Deep Context-Aware Feature Extraction for Anomaly Detection in Surveillance Videos

Anuja Phapale
Department of Computer Engineering & Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India
anuja.phapale@gmail.com (corresponding author)

Sukhada Bhingarkar
Department of Computer Engineering & Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India
sukhada.bhingarkar@mitwpu.edu.in

Received: 2 December 2024 | Revised: 29 December 2024 | Accepted: 7 January 2025

ABSTRACT

Surveillance video analysis plays a crucial role in ensuring public safety and security. Developing a context-aware framework for anomaly detection in surveillance videos is motivated by the need for enhanced security, safety, and efficiency in various domains. Context-aware anomaly detection depends on spatiotemporal features that help the model understand the context of anomalies in surveillance videos. This study aimed to provide a novel deep learning-based context-aware approach to feature extraction to detect anomalies in surveillance videos. The proposed method integrates ResNet50 for spatial feature extraction and 3D Convolutional Neural Network (CNN) for temporal feature extraction. This method identifies six anomalous activities, namely abuse, arrest, fighting, robbery, shooting, and road accidents, using the UCF-Crime dataset. The proposed integrated ResNet50 and 3D CNN model achieves promising accuracy for the six classes, such as 95% for abuse, 93% for arrest, 95.22% for fighting, 94.44% for robbery, 93% for shooting, and 94.22% for road accidents. By combining spatiotemporal features, the proposed model detects anomalies in behavior and unexpected movements, which makes it useful for security monitoring where deviations from normal behavior indicate anomalous events. This research contributes to advancing the capabilities of surveillance systems, enhancing public safety, and enabling proactive security measures in diverse urban environments.

Keywords-deep learning; ResNet50; surveillance video analysis; spatiotemporal analysis; UCF-crime dataset; 3D CNN

I. INTRODUCTION

Surveillance cameras are employed to monitor activities and recognize unexpected events in public areas, such as markets, schools, and hospitals, to improve security. Anomaly detection aims to identify events or behaviors that deviate from expected or normal patterns. Early detection of such abnormal events by surveillance cameras is beneficial for crime prevention, public safety management, and investigation. Different types of anomalies can be detected in surveillance videos, including theft, violence, erratic movements, unusual gestures, suspicious activities, traffic accidents, intrusion, vandalism, fire or smoke, and loitering.

As surveillance systems produce massive video data, the identification of anomalous events using human intervention is time-consuming and labor-intensive. To reduce time, human errors, and storage costs, it is necessary to develop an efficient and automated surveillance system to detect anomalies. Recent technological advances in deep learning have led to the development of new tools and techniques that can aid in anomaly detection. Classification and object detection have shown promising results in identifying suspicious activities to enhance security and safety. Anomaly detection in surveillance videos is an interesting research topic for security and monitoring applications. Several methods have been proposed for the detection of anomalous events. Models are trained on videos that exhibit regular patterns, and events that diverge are identified as abnormal. However, this is quite a challenging task due to the intensity of light while capturing events, environmental conditions, background noise, and handling real-time scenarios [1]. As an activity detected as anomalous in one scene might not be anomalous in another scene, since it depends on the context, there is a need for context-aware anomaly detection in surveillance videos. To understand the context of activity, prominent spatiotemporal features play a vital role in accurate anomaly detection.

The literature on anomaly detection in surveillance footage is mostly categorized into handcrafted feature extraction and machine learning or deep learning-based feature extraction approaches. Researchers have used handcrafted feature extraction approaches, such as Histogram of Oriented Gradients (HOG) [2-4], Scale-Invariant Feature Transform (SIFT) [5], and Local Binary Patterns (LBP) [6], for human activity and behavior identification in surveillance video. These methods extract features such as shape, edge, corner, color, and texture, and some low-level spatiotemporal features. Other low-level spatiotemporal feature extraction techniques have been devised based on HOG and Histogram Oriented Flow (HOF) to address issues associated with crowded scenes [3, 7]. In [8], a Markov Random Field (MRF) technique was proposed to extract spatiotemporal features from videos. The MRF graph was utilized to represent spatial-temporal local regions and local optical flow, modeled with a Mixture of Probabilistic Principal Component Analyzers (MPPCA). These handcrafted feature extraction techniques mostly rely on features that are manually generated, which requires a long time and a lot of resources to process the data. Moreover, these approaches perform poorly with real-life instances and do not scale well for different datasets. Although somewhat successful, these techniques frequently have trouble addressing changes in appearance and motion, as well as capturing intricate spatiotemporal patterns.

Traditional feature extraction methods are time-consuming and require feature engineering and domain knowledge. However, many researchers have attempted to create efficient machine learning and deep learning-based techniques to identify prominent features in unprocessed surveillance video data. Deep learning methods are capable of automatically processing raw video data for feature extraction, activity recognition, and classification. These techniques help to extract features that represent spatiotemporal dynamics, appearance variations, and motion patterns in surveillance videos for more reliable context-aware anomaly detection systems. Context-aware anomaly detection depends on spatiotemporal features that help to infer the context. Deep Convolution Neural Network (DCNN) and Recurrent Neural Network (RNN) based methods, discussed in [9-11], extract spatiotemporal features that have significant contributions to detecting unusual activities in surveillance video. The edges, textures, and patterns of the object correspond to spatial features, whereas the sequence of changes that occur on the object over time denotes temporal features of the detected object in the scene. In [12], a novel deep learning model was tailored for anomaly detection in surveillance videos, addressing the challenge of efficiently analyzing massive video data. A Lightweight CNN (LWCNN) was used to extract spatiotemporal features from salient shots, in which time-distributed 2D layers effectively collected information across frames. There is a growing need for anomaly detection in surveillance video systems for a variety of public safety and security applications. A CNN-based anomaly recognition framework tailored for smart city surveillance [13] and a system to avoid malpractices during exams in an educational institution [14] address the challenges of diverse and infrequent anomalies. Previously proposed methods take advantage of CNN, LSTM, EfficientNet, etc., for activity recognition and classification in real-time applications.

Although CNNs and RNNs are effective tools for detecting anomalies in surveillance videos, they face many difficulties in some applications. CNNs efficiently capture spatial features but have trouble with temporal dynamics. RNNs are better at capturing temporal dependencies but face challenges in handling long-term sequences, noisy data, and complicated interactions. Some of these constraints can be overcome by combining these strategies with other techniques, such as transformer models, attention mechanisms, or hybrid models that combine CNNs and RNNs and refine them for real-time processing. To further increase the discriminative power of extracted features, attention techniques have been incorporated into deep learning architectures to focus on pertinent spatiotemporal regions. In [15], an advanced video anomaly detection framework was proposed to interpret spatiotemporal visual surveillance data, based on a CNN and a vision transformer model. By employing a small portion of the video data, the transformer-based model enables accurate prediction and captures the complex spatiotemporal relationships between multiple frames. The vision transformer-based encoder-decoder introduced in [16] for anomaly identification and localization can identify anomalous images and regions. The encoder extracts a representation that includes most of the standard information, such as connections between the image patches, while the decoder maintains spatial information. Table I summarizes deep learning approaches for anomaly detection in surveillance videos along with extracted features and benchmark datasets used for evaluation.

TABLE I. METHODS FOR ANOMALY DETECTION IN SURVEILLANCE VIDEOS AND BENCHMARK DATASETS

Methods	Feature extracted	Reference	Datasets
CNN, Rolling prediction algorithm	Temporal	[1]	Vehicle Accident Image Dataset
Integration of ResNet-50/34/18 and SRU	Spatiotemporal	[9]	UCF-Crime
ResNet50, BiLSTM, PCA	Spatiotemporal	[10]	UCF-Crime
3D ResNet, DMIL	Appearance, motion features, spatiotemporal	[11]	UCF-Crime, ShanghaiTech
LWCNN, LSTM	Spatiotemporal	[12]	Avenue, UCSD Pedestrian1, UCSD Pedestrian2, UCF-Crime
MobileNetV2, Residual attention-based LSTM	Spatiotemporal	[13]	UCF-Crime, UMN [28], Avenue
CNN, Vision transformer	Spatiotemporal	[15]	Ped2, Avenue, ShanghaiTech
Auto-Encoder, Attention mechanism	Spatial	[17]	USCD Pedestrian1, USCD Pedestrian2 [20-22], CUHK Avenue [23, 24]
Autoencoder, SHAP	Appearance, Spatiotemporal	[18]	USCD Pedestrian1, USCD Pedestrian2
Deep Support Vector Data Description (DSVDD)	Spatiotemporal	[19]	CUHK Avenue ShanghaiTech Campus [27]

Spatial features help in understanding the static content of a scene, such as the appearance and locations of objects, whereas temporal features capture information about motions and changes over time. To recognize complex anomaly patterns in surveillance videos, a combination of spatial or temporal features plays an important role. Spatiotemporal features enable the detection of complex anomalies that involve both space and time dimensions. After studying previous research in this area, some gaps were determined that require more work. There is a need for movement towards self- and unsupervised learning methods that use big, unlabeled video datasets to learn feature representations. Still, there are some issues with handling occlusions and intricate interactions in crowded situations, as well as modeling long-term temporal connections efficiently. Changes in the environment, such as changes in lighting, weather conditions, or different times of the day, are some factors that affect the accuracy of anomaly detection models. There is an increasing need for automated systems that can process and extract relevant data from surveillance footage in an effective manner. This emphasizes the importance of advanced feature extraction techniques, such as deep context-aware approaches for video analysis to automatically extract complex spatiotemporal patterns.

The motivation for context-aware feature extraction in surveillance videos stems from the inherent complexity and dynamic nature of real-world environments. Traditional approaches often perform poorly in tasks such as object detection, tracking, and anomaly detection because they cannot fully extract the rich contextual information included in surveillance videos. This study aimed to improve the robustness and discriminative capacity of deep learning models by explicitly considering appearance, motion, temporal, and spatial signals in the context of their surrounding environment through the integration of context awareness into feature extraction. Through the extraction of more representative and useful features, events and activities caught in surveillance recordings can be understood and interpreted more accurately by integrating ResNet50 and a 3D CNN architecture. ResNet50 primarily extracts spatial features from individual frames of the input video. These spatial features capture the patterns, textures, and objects present within each frame. Although ResNet50 itself does not explicitly model temporal information, by processing each frame independently, it implicitly captures spatial features across different time steps. Therefore, the 3D CNN model is specifically designed to capture temporal information by operating directly on 3D volumes of data (i.e., spatial dimensions and time). The 3D CNN model convolves filters over both spatial and temporal dimensions, allowing it to extract features that capture motion and appearance changes over time. The extracted temporal features represent a more explicit modeling of the temporal dynamics and interactions between consecutive frames in the video sequence. So, the integration of ResNet50 and 3D CNN extracts spatiotemporal features to help detect anomalies in surveillance video. ResNet50 primarily extracts spatial features from individual frames, while the 3D CNN explicitly models temporal information by operating on 3D volumes. Integrating both models allows for a comprehensive representation of spatiotemporal features by combining the strengths of both approaches. Context-aware feature integration ultimately promises to increase the efficacy of surveillance systems in a variety of applications, such as traffic management, crowd analysis, and security monitoring.

II. THE PROPOSED METHOD

In today's data-driven environments, the integration of surveillance systems with anomaly detection mechanisms is crucial for protecting assets and data, as well as aiding well-informed decision-making. Context-aware anomaly detection in surveillance videos involves analyzing videos to identify unusual activities or behaviors that deviate from expected patterns. Figure 1 shows the proposed system for context-aware anomaly detection in videos.

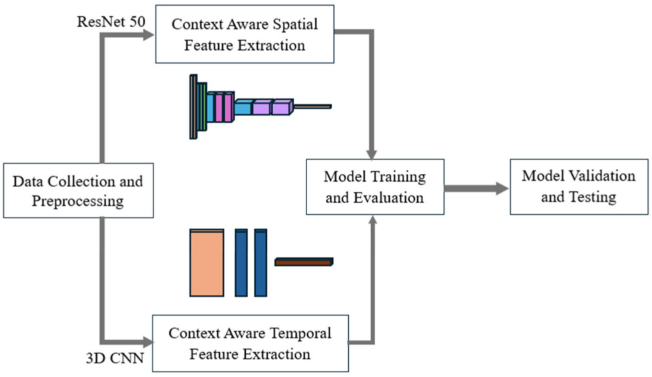
ResNet 50
Context Aware Spatial Feature Extraction
Data Collection and Preprocessing
Model Training and Evaluation
Model Validation and Testing
3D CNN
Context Aware Temporal Feature Extraction

Fig. 1. Integrated ResNet50 and 3DCNN anomaly detection method.

This study used a subset of the UCF-Crime dataset that contains images extracted from video. Six anomalous classes were chosen from the UCF dataset since entire training videos incur an extensive computational expense. For a context-aware anomaly detection system, both spatial and temporal features are essential. Spatial features provide static information, such as the shape, size, texture, pattern, and color of the detected object, and temporal features provide dynamic information, such as frame-to-frame changes in a video. ResNet50 is used to extract spatial features, whereas a 3D CNN is used to capture temporal features. The sequence of spatial features extracted from RescNet50 is fed into 3D CNN to capture temporal dependencies. By combining ResNet50 and 3D CNN, the system can establish a dynamic baseline for normal behavior and movements and quickly detect deviations that signify potential for detecting anomalies in surveillance videos.

A. Integrated Deep Learning Architecture for Anomaly Detection

In the proposed architecture, ResNet50 extracts spatial features and the 3D CNN extracts temporal features to create a unified feature representation. These features can then be passed to downstream tasks, such as action recognition, video classification, or other spatiotemporal analysis tasks. Additional layers, such as fully connected or recurrent layers, are added on top of the concatenated features for further processing or classification. Both ResNet50 and 3D CNN parameters were fine-tuned to achieve optimal performance.

The proposed anomaly detection approach in surveillance video for spatiotemporal feature extraction involves several steps. This approach leverages the spatial feature extraction capabilities of ResNet50 and the temporal dynamics captured by 3D CNNs. The proposed system is represented as:

- *X* denotes the input video frames.
- *FResNet50*(*X*) denotes the feature extraction performed by ResNet50.
- *F3DCNN*(*FResNet50*(*X*)) denotes the temporal feature extraction performed by the 3DCNN.
- *C* denotes the contextual information (if available).
- *Fcontext*(*F3DCNN*(*FResNet50*(*X*)), *C*) denotes the fusion of contextual information with the spatiotemporal features.

1) ResNet50 for Spatial Feature Extraction

Let *FResNet50(X)* represent the output feature map after passing the input *X* through the ResNet50 architecture represented as:

$$FResNet50(X) = ResNet50(X) \qquad (1)$$

2) 3D CNN for Temporal Feature Extraction

Let *F3DCNN*(*FResNet50*(*X*)) represent the output feature map after passing the features extracted by ResNet50 through the 3D CNN represented as:

$$F3DCNN\big(FResNet50(X)\big) = 3DCNN(FResNet50(X)) \qquad (2)$$

3) Context Fusion

Contextual information *C* is available, which can be fused with the spatiotemporal features obtained from the 3DCNN. Let *Fcontext*(*F3DCNN*(*FResNet50*(*X*)), *C*) represent the fused feature map given as given below:

$$Fcontext\big(F3DCNN\big(FResNet50(X)\big), C\big) = Fusion\big(F3DCNN\big(FResNet50(X)\big), C\big) \qquad (3)$$

4) Integrated ResNet50 and 3DCNN Architecture:

The final output of the integrated architecture is:

$$Y = Fcontext(F3DCNN(FResNet50(X)), C) \qquad (4)$$

III. RESULTS AND DISCUSSION

A. Dataset

The UCF-Crime dataset [25] consists of 128 hours of recordings with an average number of 7,247 frames per video. This dataset contains 13 actual anomalies, namely abuse, arrest, arson, assault, vehicle accident, explosion, fight, robbery, shooting, stealing, shoplifting, burglary, and vandalism, consisting of 1900 uncut actual surveillance recordings. It is a video-level labeled dataset, in which the training dataset consists of 1,610 videos out of which 800 are normal and 810 are anomalous, and the test dataset consists of 290 videos containing 150 normal and 140 anomalous. A subset of the dataset was prepared by extracting images from each video. Every tenth frame of every full-length video was extracted and added to each video in that class. There is a total of 1,377,653 images for the 13 anomalous classes and the normal class, of which the training dataset consists of 1,266,345 images and the test dataset consists of 111,308 images sized 64×64 in PNG format.

B. Results

A total of 2,310 images was considered, using 164 images in the test set and 2,146 images in the training set. The training and the test set contained six different anomaly classes, namely abuse, arrest, fighting, robbery, shooting, and road accidents. The accuracy and Area Under the Curve (AUC) scores were used to evaluate the efficacy of the proposed framework. The ResNet50 model is defined with ImageNet pre-trained weights. TensorFlow's Keras API was used for temporal feature extraction to define and assemble the 3D CNN model. This model was constructed in a stepwise manner and consists of three convolutional layers with 3×3×3 kernels and progressively more filters (32, 64, and 128). This method leverages the 3D CNN's capacity to detect temporal patterns and is frequently applied in tasks involving temporal data, such as video sequences or time-series data.

The proposed model was trained and tested using Keras in a Jupyter notebook. As the training loss dramatically decreased throughout epochs, the training accuracy increased steadily and reached 100%, demonstrating strong performance on the training set. The validation loss fluctuated but eventually declined, while the validation accuracy stayed steady at 18.9%, indicating that the model may be overfitting the training set without appreciably improving it on the validation set. This suggested that more varied validation data, regularization, or fine-tuning may be required.

Figure 2 shows the ResNet50 model's training and validation accuracy and loss plotted over 50 epochs. The training accuracy is represented by the orange line, which rapidly climbs and stabilizes near 1.0 to represent almost perfect accuracy. The validation loss, which measures performance on unseen data, is represented by the green line. It varies greatly, indicating the possibility of overfitting because the model may be learning the training data too well without properly generalizing. Finally, the validation accuracy is represented by the red line, which starts high and stays mostly stable but is marginally lower than the training accuracy. This indicates that although the model performs well on the validation set, it falls short of the nearly flawless performance observed during training. The plot's Y-axis displays accuracy and loss from 0 to 1 and the X-axis denotes the number of epochs. Figure 3 shows the 3D CNN model's training and validation accuracy and loss plotted over 50 epochs. The model effectively learned from the training data and reduced its error, as seen by the blue line representing the training loss that gradually dropped over time. Training accuracy is represented by the orange line, which rapidly increased to a value approaching 1.0, indicating that the model generated almost perfect predictions on the training set. The validation loss, represented by the green line, varies significantly, particularly after 30 epochs, indicating that the model's capacity to generalize to new data may not always be consistent. In the meantime, validation accuracy, shown by the red line, is roughly constant around 0.8-0.85 and somewhat lower than training accuracy. This suggests that the model performed well without experiencing severe overfitting and implies that there is space for improvement in stability due to the oscillations in validation loss.

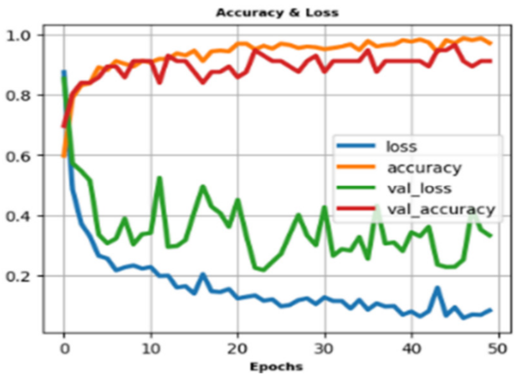

Fig. 2. Accuracy and loss plot for the Resnet50 model.

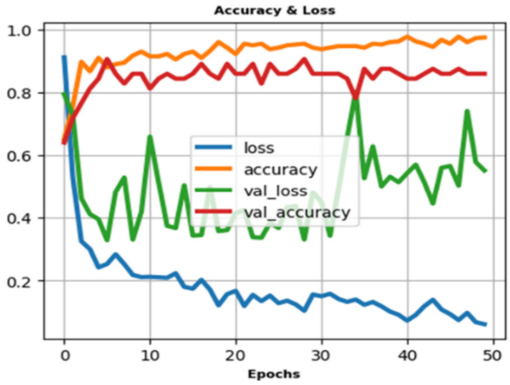

Fig. 3. Accuracy and loss plot for 3D CNN model.

The accuracy of the proposed model was compared with the pre-trained VGG16, MobileNetV2, VGG19, InceptionV3, and ResNet50 models, as shown in Table II, and Figure 4 shows their performance. The hardware requirements for building models using ResNet50 and 3D CNN on the UCF-Crime Dataset depend on various factors, including dataset size, model complexity, and the computing capacity needed for training and inference. The proposed model was implemented using Keras, TensorFlow, and OpenCV with Python 3.10.9 on Windows 11 Pro and Core$^{(TM)}$ i7-8665U setup with 32GB RAM.

The proposed integrated feature extraction model with ResNet50 and 3D CNN achieved promising accuracy for six classes, such as 95% for abuse, 93 % for arrest, 95.22% for fighting, 94.44% for robbery, 93% for shooting, and 94.22% for road accidents, compared to the baseline models VGG16, MobileNetV2, VGG19, InceptionV3, and ResNet50.

TABLE II. RESULTS

Model	Metrics (%)	Abuse	Arrest	Fighting	Robbery	Shooting	Road Accidents
Proposed model (Resnet50+ 3D CNN)	Accuracy	95	93	95.22	94.44	93	94.22
	AUC	95	93	92	91	89	78
VGG16	Accuracy	92	93	91	91	89	82
	AUC	94	94	90	91	87	77
MobileNetV2	Accuracy	92	89	88	90	77	78
	AUC	90	88	89	88	84	83
VGG19	Accuracy	89	89	84	86	79	78
	AUC	91	90	87	89	81	78
InceptionV3	Accuracy	85	83	82	79	73	71
	AUC	85	82	83	80	74	71
ResNet50	Accuracy	83	83	83	79	68	67
	AUC	81	78	81	80	76	71

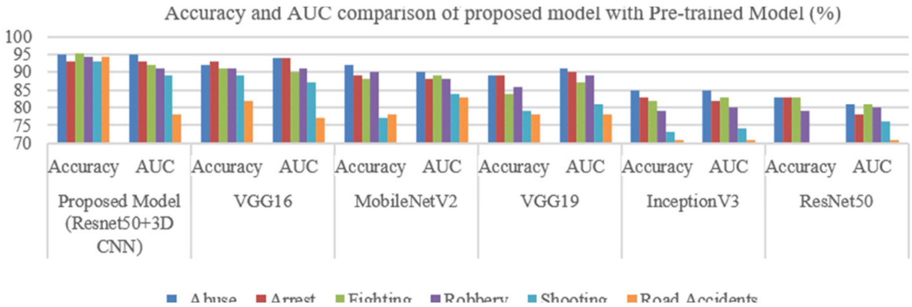

Fig. 4. Classwise accuracy and AUC comparison of the proposed with pretrained models.

IV. CONCLUSION

This study presents a pioneering approach to spatiotemporal feature analysis in surveillance videos by utilizing a deep context-aware feature extraction technique. Context-aware anomaly detection in video refers to the process of identifying abnormal events by considering the contextual information surrounding them. Spatiotemporal features are important to understand the context of scenes to detect anomalies. Many deep learning-based methods are available to help extract spatiotemporal features from video data and detect anomalies. In this work, integrating ResNet50 for spatial feature extraction and 3D CNN for temporal feature extraction demonstrated effectiveness in capturing both spatial context and temporal dynamics. By leveraging the UCF-Crime dataset, the proposed model achieved remarkable performance in extracting spatiotemporal features. These results demonstrate the potential of this feature extraction method to significantly increase surveillance system capabilities and consequently support public safety and security in a variety of urban settings. Further research and development in this area can lead to the adoption of more advanced techniques for proactive security measures and the prevention of criminal activities in real-world settings along with integrating them with automated systems.

REFERENCES

[1] S. W. Khan *et al.*, "Anomaly Detection in Traffic Surveillance Videos Using Deep Learning," *Sensors*, vol. 22, no. 17, Aug. 2022, Art. no. 6563, https://doi.org/10.3390/s22176563.

[2] K. P. Sanal Kumar and R. Bhavani, "Human activity recognition in egocentric video using HOG, GiST and color features," *Multimedia Tools and Applications*, vol. 79, no. 5–6, pp. 3543–3559, Feb. 2020, https://doi.org/10.1007/s11042-018-6034-1.

[3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, vol. 1, pp. 886–893, https://doi.org/10.1109/CVPR.2005.177.

[4] B. Sabzalian, H. Marvi, and A. Ahmadyfard, "Deep and Sparse features For Anomaly Detection and Localization in video," in *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, Tehran, Iran, Mar. 2019, pp. 173–178, https://doi.org/10.1109/PRIA.2019.8786007.

[5] A. Temizel, P. Güler, and T. T. Temizel, "Real-Time Global Anomaly Detection for Crowd Video Surveillance Using SIFT," in *5th International Conference on Imaging for Crime Detection and Prevention (ICDP 2013)*, London, UK, 2013, https://doi.org/10.1049/ic.2013.0263.

[6] K. Seemanthini, S. S. Manjunath, G. Srinivasa, B. Kiran, and P. Sowmyasree, "A Cognitive Semantic-Based Approach for Human Event Detection in Videos," in *Smart Trends in Computing and Communications*, 2020, pp. 243–253, https://doi.org/10.1007/978-981-15-0077-0_25.

[7] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," in *Computer Vision – ECCV 2006*, 2006, pp. 428–441, https://doi.org/10.1007/11744047_33.

[8] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, Jun. 2009, pp. 2921–2928, https://doi.org/10.1109/CVPR.2009.5206569.

[9] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," *Results in Engineering*, vol. 18, Jun. 2023, Art. no. 101026, https://doi.org/10.1016/j.rineng.2023.101026.

[10] Z. K. Abbas and A. A. Al-Ani, "An adaptive algorithm based on principal component analysis-deep learning for anomalous events detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, Jan. 2022, Art. no. 421, https://doi.org/10.11591/ijeecs.v29.i1.pp421-430.

[11] S. Dubey, A. Boragule, J. Gwak, and M. Jeon, "Anomalous Event Recognition in Videos Based on Joint Learning of Motion and Appearance with Multiple Ranking Measures," *Applied Sciences*, vol. 11, no. 3, Feb. 2021, Art. no. 1344, https://doi.org/10.3390/app11031344.

[12] S. Ul Amin *et al.*, "EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos," *Mathematics*, vol. 10, no. 9, May 2022, Art. no. 1555, https://doi.org/10.3390/math10091555.

[13] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos," *Sensors*, vol. 21, no. 8, Apr. 2021, Art. no. 2811, https://doi.org/10.3390/s21082811.

[14] N. Gupta and B. B. Agarwal, "Suspicious Activity Classification in Classrooms using Deep Learning," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12226–12230, Dec. 2023, https://doi.org/10.48084/etasr.6228.

[15] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Engineering Applications of Artificial Intelligence*, vol. 123, Aug. 2023, Art. no. 106173, https://doi.org/10.1016/j.engappai.2023.106173.

[16] Y. Lee and P. Kang, "AnoViT: Unsupervised Anomaly Detection and Localization With Vision Transformer-Based Encoder-Decoder," *IEEE Access*, vol. 10, pp. 46717–46724, 2022, https://doi.org/10.1109/ACCESS.2022.3171559.

[17] Q. Zhang, H. Wei, J. Chen, X. Du, and J. Yu, "Video Anomaly Detection Based on Attention Mechanism," *Symmetry*, vol. 15, no. 2, Feb. 2023, Art. no. 528, https://doi.org/10.3390/sym15020528.

[18] C. Wu, S. Shao, C. Tunc, P. Satam, and S. Hariri, "An explainable and efficient deep learning framework for video anomaly detection," *Cluster Computing*, vol. 25, no. 4, pp. 2715–2737, Aug. 2022, https://doi.org/10.1007/s10586-021-03439-5.

[19] B. Wang, C. Yang, and Y. Chen, "Detection Anomaly in Video Based on Deep Support Vector Data Description," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–6, May 2022, https://doi.org/10.1155/2022/5362093.

[20] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 2010, pp. 1975–1981, https://doi.org/10.1109/CVPR.2010.5539872.

[21] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly Detection and Localization in Crowded Scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, Jan. 2014, https://doi.org/10.1109/TPAMI.2013.111.

[22] "Anomaly Detection and Localization in Crowded Scenes." [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/.

[23] "Avenue Dataset." [Online]. Available: https://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html.

[24] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in *2013 IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 2013, pp. 2720–2727, https://doi.org/10.1109/ICCV.2013.338.

[25] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 6479–6488, https://doi.org/10.1109/CVPR.2018.00678.

[26] "UCF-Crime-database",[Online] Available: https://www.dropbox.com/sh/75v5ehq4cdg5g5g/AABvnJSwZI7zXb8_myBA0CLHa?dl=0

[27] W. Luo, W. Liu, and S. Gao, "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Oct. 2017, pp. 341–349, https://doi.org/10.1109/ICCV.2017.45.

[28] "Monitoring Human Activity - Detection of Events." [Online]. Available: https://mha.cs.umn.edu/proj_events.shtml.