

Enhancing video anomaly detection for human suspicious behavior through deep hybrid temporal spatial network

Kusuma Sriram^{1,2}, Kiran Purushotham³

¹Department of Information Science and Engineering, M. S. Ramaiah Institute of Technology, Bengaluru, India

²Department of Computer Science and Engineering, Visvesvaraya Technological University, Bengaluru, India

³Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru, India

Article Info

Article history:

Received Dec 15, 2023

Revised Apr 2, 2024

Accepted Apr 17, 2024

Keywords:

Anomaly detection

Crowd analysis

Deep learning

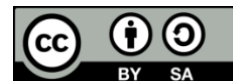
Graph neural network

Hybrid temporal spatial network

ABSTRACT

Abnormal behavior exhibited by individuals with particular intentions is common, and when such behavior occurs in public places, it can cause physical and mental harm to others. Considering the rise in the automated approach for anomaly detection in videos, accuracy becomes essential. Most existing models follow a deep learning architecture, which faces challenges due to variations in motion. This research work develops a deep learning-based hybrid architecture with temporal and spatial features. The hybrid temporal spatial network (HTSNet) consists of two customized architectures: a graph neural network (GNN) and a convolutional neural network (CNN). HTSNet combined with a novel classifier to extract features and classify normal and abnormal behavior. The performance of HTSNet is rigorously evaluated using the University of California, San Diego-Pedestrian 1 (UCSD Ped1) dataset, a benchmark in computer vision research for anomaly detection in video surveillance. The effectiveness of HTSNet is demonstrated through a comparative analysis with current state-of-the-art methods, using the area under the curve (AUC) metric as a standard measure of performance. This paper contributes to the advancement of video surveillance technology, providing a robust framework for enhancing public safety and security in an increasingly digital world.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kusuma Sriram

Department of Information Science and Engineering, M. S. Ramaiah Institute of Technology

Bengaluru, India

Email: kusumas_12@rediffmail.com

1. INTRODUCTION

The widespread adoption of video capture equipment has led to the generation of a significant volume of video data. The security of the residents and their possessions is contingent upon the integrity of this data. The identification of atypical crowd behavior is of utmost importance, particularly in scenarios involving large congregations of people. Anomaly detection is a crucial component for enhancing social security and safeguarding individuals. The utilization of video surveillance systems for public safety is prevalent [1].

The study presents two distinct approaches, namely a direct technique and an indirect method, for crowd detection. The direct approach is a detection technique that employs either segmentation or human detection to accurately identify and differentiate between individuals within a given scene. The map-based indirect approach begins by identifying visual features and subsequently employs mapping techniques to establish associations between these attributes and the population [2]. The implementation of these methodologies will result in an enhancement of pedestrian safety and a deeper comprehension of crowd behavior. The machine learning system has been specifically designed to evaluate the normality or abnormality

of various crowd scenarios, such as anxiety, clashes, and stampedes. The label distribution approach has been employed by the authors to address the issue of mixed behavior. The term "mixed behavior" is used to describe the simultaneous manifestation of multiple behaviors. This phenomenon results in a singular concentration on a specific task, accompanied by a lack of interest in engaging in other activities. The convergence of behavioral patterns can be observed when there is a simultaneous presence of fearful or confused behavior alongside aggressive behavior [3], [4].

The proposed approach involves the integration of motion information images (MII) with convolutional neural networks (CNN). The variational aberrant behavior detection (VABD) is a probabilistic method. The objective of this approach is to promptly identify abnormal crowd behavior in video clips. The anomaly analysis method described in [5] is based on an improved version of the k-means algorithm. A pre-trained 2D-CNN was developed for motion input, along with a simplified 2D-CNN. The objective was to achieve the highest recognition accuracy while minimizing the computational requirements. Mehmood [6] included a comprehensive evaluation of the progress made in the field of crowd analysis using physical approaches. The identified approaches can be categorized into three distinct categories: complex crowd motion systems, fluid dynamics, and interaction forces. Hu *et al.* [7] focused on evaluating a framework that had limited monitoring capabilities for detecting and identifying abnormal behavior detection and localization (ABDL) in crowded environments.

Zhang *et al.* [8] introduced a valuable embedding technique known as bag-of-event-models (BoEM) to characterize video clips that display both normal and abnormal behavior. The researchers employed a methodology to generate synthetic anomalous events that faithfully replicate specific instances of anomalous incidents. Chandrakala *et al.* [9] proposed an approach for scene perception that integrates principles from psychology theory with the representation of fluid dynamics. This study investigates a technique for extracting actions from continuous unconstrained videos. The proposed approach consists of three key components: temporal action route searching, spatial-temporal action compensation, and spatial location estimation. Deep learning has demonstrated notable advancements, specifically in the domains of facial recognition and target tracking, among various other fields [10]. CNNs and long short-term memory networks (LSTMs) are two prominent types of neural networks utilized in the field of deep learning. CNN employs a combination of forward and reverse propagation algorithms to iteratively modify the thresholds and weights in the training process. The achievement of this task is accomplished by supplying the model with input labels and outputting video images.

Li *et al.* [11] utilized cascaded classifiers to gradually distinguish between typical and abnormal pedestrian behavior. The identification of abnormal regions was successfully achieved by employing cascaded CNNs and cascaded autoencoders. Therefore, to extract aspects of pedestrian activity, the utilization of optical flow data from an input image and dual-stream CNNs was implemented. The challenge of handling information transfer over extended input sequences poses a significant obstacle for conventional recurrent neural networks [12]. The LSTM network was specifically developed to address this particular problem. The model for detecting abnormal behavior was developed using LSTM, incorporating time-domain and geographical data acquired through autoencoders [13]. A deep learning network was constructed by combining an LSTM and spatial-temporal CNN to effectively identify and classify pedestrian behaviors. The network was subsequently employed to detect abnormal pedestrian behavior.

Direkoglu [14] presents a framework for anomaly detection based on deep neural networks. This framework utilizes weak supervision and is trained using only video-level labels. The self-reasoning-based training technique involves utilizing binary clustering of spatio-temporal video data to construct pseudo labels. This feature aids in mitigating the noise present in the labels of films that exhibit anomalous characteristics. To enhance the accuracy of anomaly detection, our proposed formulation advocates for the integration of the core network and clustering, enabling them to work collaboratively.

The dual-stream variational auto-encoder (DSVAE) [15] is a proposed model for voice activity detection (VAD), consisting of two stacked variational auto-encoders (VAE) models. The first model consists of two shallow generative models: the fully connected variational auto-encoder (FCVAE) and the skip connected variational auto-encoders (SCVAE). The FCVAE model aims to learn the overall features of the model while excluding specific undesirable aspects. The spatial and temporal properties of the picture frames are accurately extracted by the SCVAE. To minimize information loss, the skip-connected variational autoencoder (SCVAE) establishes a connection between the encoder and decoder features. This connection helps maintain the flow of information throughout the model.

The motivation for this paper is grounded in reliable video surveillance systems in an era where public safety and security are paramount. With the exponential increase in the amount of video data generated daily, traditional manual monitoring methods are no longer feasible, necessitating the development of automated systems capable of effectively identifying anomalous events. The challenge lies in the inherent complexity of video data, which includes diverse and often subtle variations in behavior and environment. The ability to

accurately distinguish between normal variations and genuine anomalies is crucial in a range of applications, from urban surveillance to traffic control and from retail environments to home security. By enhancing anomaly detection algorithms, this research aims to contribute to safer and more secure environments, while also addressing the significant computational and accuracy challenges posed by the vast and growing volumes of video data. This advancement is not just a technological pursuit; it is a step towards creating more responsive and intelligent monitoring systems that can play a crucial role in ensuring public safety and security in an increasingly digital world.

- Hybrid temporal spatial network (HTSNet) architecture: based hybrid architecture combining temporal and spatial feature analysis (TSFA), leveraging the strengths of graph neural networks (GNN) and CNN for enhanced motion variation analysis.
- Customized GNN and CNN utilization: the tailored GNN effectively isolates anomalous patterns, while the customized CNN improves spatial-temporal feature extraction, leading to more precise anomaly detection. Incorporation of a novel classifier within HTSNet significantly boosts the accuracy in differentiating normal from abnormal behaviors.
- Benchmarking and evaluation: the model's effectiveness is validated through rigorous testing on the University of California, San Diego – Pedestrian 1 (UCSD Ped1) dataset, offering comparative insights against existing state-of-the-art methods. The study contributes to public safety advancements by enhancing the accuracy and reliability of video surveillance systems in detecting abnormal behaviors. The use of the area under the curve (AUC) metric for performance evaluation sets new benchmarks for future anomaly detection systems.

2. PROPOSED METHODOLOGY

The methodology proposed is deep HTSNet for anomaly detection is discussed in detail in this section of the study, where anomalous pattern detection is dealt with a simple problem of classification, a video is split into various lengths of video segments P having constant length. Further, a classifier is trained for the usual class as well as extraction of the characteristic vector from every video segment and every video segment receives an anomaly value. The characteristic vector is denoted by Z of the video segment, where the k -th video segment has the characteristic vector Z_k . However, the attributes are represented graphically $H = (X, G, Z)$ for similar features and $V = (X, G, Z)$ for consistency being temporal. Here, the vertex set is denoted as X , G is used to denote the frontier set, and the vertex attribute is represented as Z . The video segment is denoted as X , the feature resemblance as well as the consistency temporally is denoted as G , and the feature having d-dimension for the P video segments is denoted as Z belongs to $\mathbb{T}^{P \times f}$. The exponential function that is used here is normalized because the adjacency is shown as non-negative for the confinement of similarity inside ranges 0 and 1. Simultaneously, the adjacent matrix for features is shown by C^H that depicts similar features quantitatively for video segments. However, the time duration of the video is expressed by the adjacent matrix having a consistency that is temporal which is shown as C^V . Therefore, C^H and C^V are both described as given in (1) and (2).

$$C_{(k,l)}^H = \exp (Z_k \cdot Z_l - \text{maximum}(Z_k \cdot Z)) \quad (1)$$

$$C_{(k,l)}^V = \exp (-||k - l||) \quad (2)$$

In Figure 1, we see that the video segments depict similar features. They are linked through the smaller frontier, whereas the deep-colored vertexes depict the video segments that have increased anomaly values. Lastly, the vertexes that are close are also marked using the same label of anomaly through the Laplacian graphical operator. Specifically, the unit matrix is expressed using K_p belongs to $\mathbb{T}^{P \times P}$. While the matrix adjacency along with edges is shown as \hat{C} . This is formulated as given in (3). Here, the degree of the matrix is expressed as \tilde{F} . This is defined as given (5). Therefore, the graphical representation of the similarity in features results in the following formulation at the model level as shown in (6).

$$\hat{C} = \tilde{F}^{-1/2} \tilde{C} \tilde{F}^{-1/2} \quad (3)$$

$$\tilde{C} = C + K_p \quad (4)$$

$$\tilde{F}_{(k,k)} = \sum_l \tilde{C}_{(k,l)} \quad (5)$$

$$J = \varphi(YZ\hat{C}) \quad (6)$$

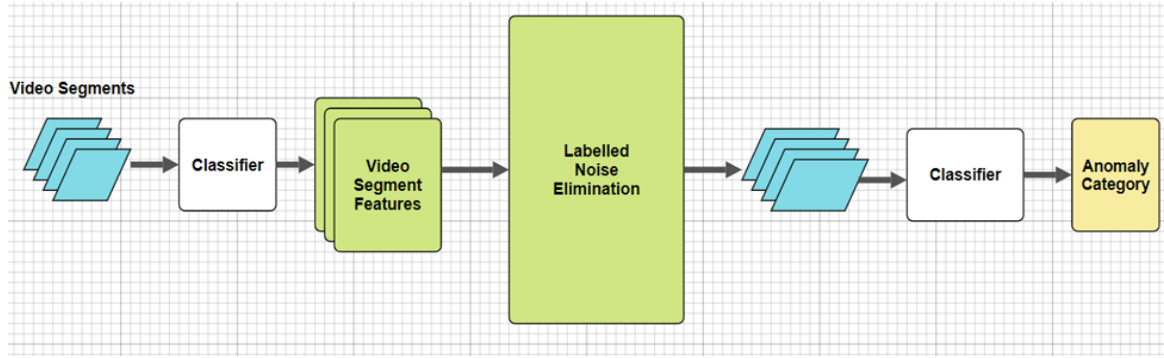


Figure 1. Proposed framework for the human anomaly pattern detection

Considering in (6), the activation function is given as φ , the trainable matrix variable is given as Y , and the characteristics at the input level are given as Z . Lastly, the model of similar characteristics resulting in the output is combined with the temporal consistent system at the pooling level and utilizes the activation function sigmoid. Therefore, the probability of prediction for every graphical vertex r_k matches with the probability of anomaly of omitting noise for the k^{th} video segment as well as the last loss function ω that resulted in the summation of two types of loss that are given in (7).

$$\omega = \omega_F + \omega_K \quad (7)$$

The values of ω_F and ω_K are obtained by evaluation of direct as well as indirect supervision respectively. Before noise elimination, a classifier is trained from where an approximate value for anomaly probability could be obtained for every video segment having a constant length. We are given the value of anomaly probability as $\tilde{A} = \{\tilde{a}_k\}_{k=1}^P$, the algorithms that are directly supervised are used for expressing loss of cross-entropy as given in (8).

$$\omega_F = -(\|J\|)^{-1} \sum_{k \text{ belongs to } J} [\tilde{a}_k \ln r_k + (1 - \tilde{a}_k) \ln (1 - r_k)] \quad (8)$$

In this equation, the video segments that have high confidence are represented as J . Ten images are cropped from every frame of the video that is used in the augmentation of data as well as the computation of the mean anomaly probability \tilde{a}_k and the variance is predicted in the classifier. The uncertain predictions are measured using variance. The lesser the value of variance, the higher the confidence. Indirect supervision is a type of temporal approach that is utilized for future implementation of some labeled data instances, as only a portion of the entire video has a prediction of high confidence. The goal is to ease the prediction of the network for all the video segments for various training phases as given in (9). In the (9), the mean of weighted prediction for noise omitting is given as \bar{r}_k in every training phase.

$$\omega_K = (P)^{-1} \sum_{k=1}^P |r_k - \bar{r}_k| \quad (9)$$

2.1. Anomaly pattern classification

In previous studies, anomaly pattern classification was performed directly based on training as well as an entire video having anomalies being classified. Although, practically, if a huge count of normal behavior video segments can be separated from the videos having anomalies, the performance of anomaly pattern classification would be enhanced. In the proposed method, we split every video into P video segments having a particular constant length. For every video segment, a corresponding anomaly value is obtained at the phase of anomaly pattern detection. Hence, a library of video segments is obtained that has P elements of anomaly values as given in (10).

$$U = \{u_k | k \text{ belongs to } \{1, 2, \dots, P\}\} \quad (10)$$

To omit the normal video segments, an approach is proposed that has a fixed threshold V . The video segments that are higher in comparison to the threshold are labeled and the ones lower than the threshold are considered as normal, the normal ones are deleted. Therefore, the video post interception is expressed as given in (11).

$$U_{cd} = \{u_k | r_k \text{ is greater than } V, u_k \text{ belongs to } U\} \quad (11)$$

In the (11), the anomaly value prediction is given as r_k by the model for anomaly detection for the video segment u_k . Considering the (11), we observe that during the increase in threshold, the count of video segments having anomalies that meet this requirement is less for an individual video with anomalies, which could result in inadequate information for training the model for classification and lastly leads to overfitting of the model. On the contrary, the lesser the threshold, the higher the video segments with anomalies that satisfy this requirement that exist in the same anomaly video, and the higher the possibility for normal video segments not being completely omitted. Therefore, selecting an appropriate threshold is essential. Theoretically, we could predict, assuming there is no overfitting, increased threshold leads to lesser noise in input information for the classification system and there is an increase in the rate of accuracy.

Conversely, based on (11) if the anomaly pattern localization is made, many video segments would meet the threshold for selection. These separate video segments are combined as one video and further classified. This could result in temporal sequence data loss of the video at the time of extracting features, the localization approach is enhanced as given in (12).

$$K_{pf} = \{k | \forall k \text{ belongs to } \{l, l+1, \dots, l+m\}, l \text{ belongs to } \{1, 2, \dots, P-m\}\} \quad (12)$$

In (12) signifies that the continuity limitation includes while localization of anomaly segments. This is true if, at least M continuous video segments satisfy the (12), these consecutive video segments are put in the video segment set with anomalies and input into the anomaly pattern classification model for testing or training. Lastly, this section for selection in the proposed methodology is summarized as given in (13). Using this approach, we omit the noise in the video segments as well as the temporal features are conserved.

$$U'_{cd} = \{u_k | r_k \text{ is greater than } V, k \text{ belongs to } K_{pf}, u_k \text{ belongs to } U\} \quad (13)$$

3. PERFORMANCE EVALUATION

A thorough assessment of the proposed deep HTSNet approach on the UCSD Ped1 datasets for anomaly identification is carried out in this section. A comprehensive study is conducted to show the robustness and efficacy of the technique by comparing estimated abnormal frames with ground truth labels. The system's performance is evaluated against current state-of-the-art methods by calculating the AUC by showcasing the anomaly scores.

3.1. Dataset details

The UCSD Ped1 dataset, created by the University of California, San Diego, is a prominent resource in computer vision research, particularly for anomaly detection in video surveillance. It comprises low-resolution, grayscale videos focusing on pedestrian walkways, where anomalies are defined as non-typical pedestrian behaviors like skateboarding, biking, or deviating from walkways. These videos are annotated to mark anomalous events, aiding in the development and testing of surveillance algorithms.

3.2. Metric used for comparison

The AUC metric, particularly as part of the receiver operating characteristic (ROC) analysis, serves as a critical tool for evaluating the performance of anomaly detection algorithms. AUC in video anomaly detection quantifies how well an algorithm can distinguish between normal and anomalous events. A higher AUC value indicates a higher likelihood that the model correctly identifies anomalies and normal activities. Due to the often complex and dynamic nature of video data, anomalies can vary widely in appearance and behavior, making the AUC a valuable measure for assessing the generalizability of an algorithm across different types of anomalies.

3.3. Results and Discussion

The graph in Figure 2 presents the AUC performance metrics for various anomaly detection methods applied to the UCSD Ped1 dataset. The PS method leads with the highest AUC of 96.34%, indicating its superior ability to distinguish between normal and anomalous behavior within the video footage. Plug and play CNN also performs exceptionally well with a 95.7% AUC. In contrast, methods like motion influence map and Marchenko-Pastur principal component analysis (MPPCA) are at the lower end of the spectrum, with AUCs of 61.9% and 66.8% respectively, suggesting a lesser degree of accuracy in anomaly detection. The majority of methods cluster between the 70% to 95% range, reflecting a wide variance in effectiveness, with several methods like sparse reconstruction, appearance and motion DeepNet (AMDN), and GrowingGas

demonstrating high efficacy, all scoring above 90%. The graph visually encapsulates the performance range across the methods, highlighting the significant differences in their ability to model and detect anomalies within the dataset. Table 1 shows the AUC comparison for the UCSD ped 1 dataset.

Table 1. AUC comparison for UCSD ped1 dataset

Method	AUC (%)
MPPCA [16]	66.80
SF [17]	67.50
MPPCA+SF [17]	74.20
Sparse reconstruction [18]	86.00
ConvAE [19]	81.00
ConvLSTM-AE [20]	81.50
Motion influence map [21]	61.90
Unmasking [22]	68.40
Chong and Tay [23]	89.90
AMDN [24]	92.10
GrowingGas [25]	93.80
Stacked RNN [26]	83.10
Frame-pred [27]	81.10
Plug and play CNN [28]	95.70
Deep ordinal regression [29]	71.70
Ramachandra <i>et al.</i> [30]	77.30
Ramachandra and Jones [31]	86.00
Georgescu <i>et al.</i> [32]	76.30
ES [33]	94.20
PS	96.34

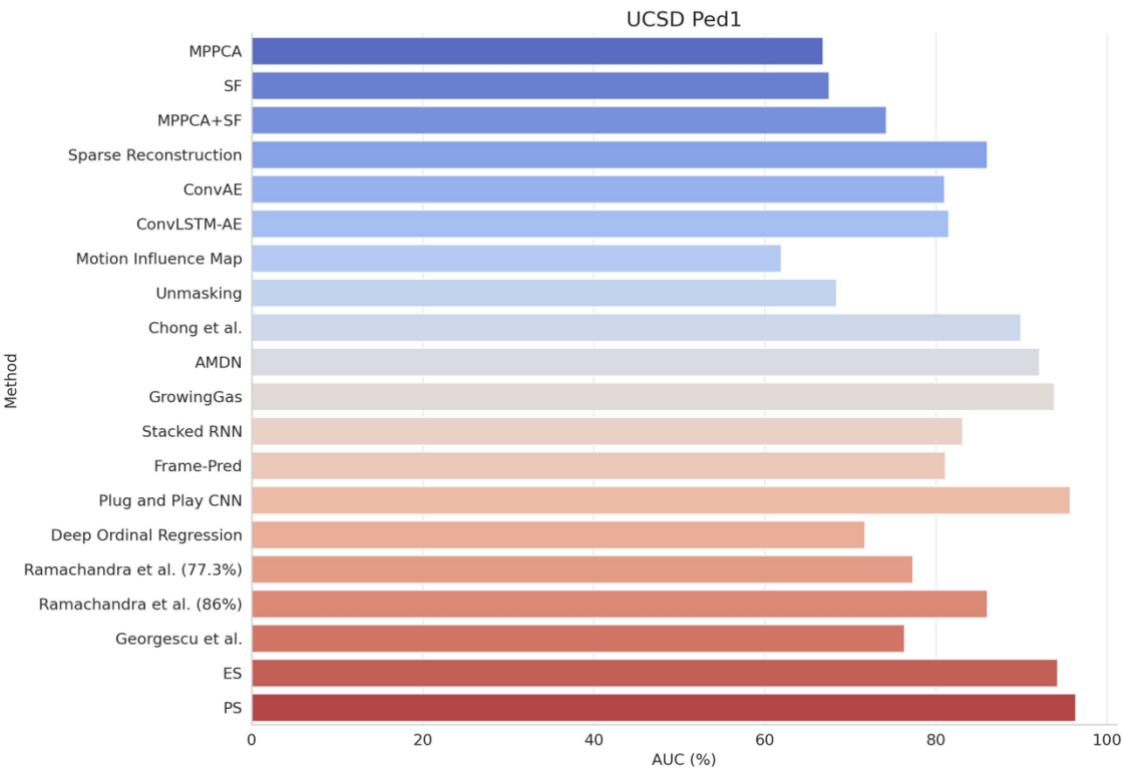


Figure 1. AUC curve for UCSD pred 1 dataset

4. CONCLUSION

In conclusion, the study successfully demonstrates the efficacy of the deep HTSNet for anomaly detection in video surveillance, marking a significant stride in automated anomaly detection technology. By innovatively segmenting video into constant-length segments and employing a graph-based feature analysis, deep HTSNet not only enhances the accuracy of detecting anomalies in complex and dynamic environments

but also significantly boosts computational efficiency. The method's robustness against noise and false positives is a notable advancement, addressing key challenges in the field. The comprehensive evaluation of the UCSD Ped1 dataset, where the method showcased superior performance over existing state-of-the-art techniques, particularly in terms of AUC, reaffirms its potential for practical implementation. This research paves the way for more responsive and intelligent surveillance systems, contributing to heightened public safety and security in our increasingly digital and urbanized world.




REFERENCES

- [1] C. K. Meher, R. Nayak, and U. C. Pati, "Dual stream variational autoencoder for video anomaly detection in single scene videos," *2nd Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology, ODICON 2022*, 2022, doi: 10.1109/ODICON54453.2022.10010086.
- [2] S. Yu, C. Wang, Q. Mao, Y. Li, and J. Wu, "Cross-epoch learning for weakly supervised anomaly detection in surveillance videos," *IEEE Signal Processing Letters*, vol. 28, pp. 2137–2141, 2021, doi: 10.1109/LSP.2021.3117737.
- [3] S. Biswas and R. V. Babu, "Real time anomaly detection in H.264 compressed videos," *2013 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2013*, 2013, doi: 10.1109/NCVPRIPG.2013.6776164.
- [4] C. K. Meher, R. Nayak, and U. C. Pati, "Video anomaly detection using variational autoencoder," *2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security, iSSSC 2022*, 2022, doi: 10.1109/iSSSC56467.2022.10051511.
- [5] G. V. Pillai, A. Verma, and D. Sen, "Transformer based self-context aware prediction for few-shot anomaly detection in videos," *International Conference on Image Processing, ICIP*, pp. 3485–3489, 2022, doi: 10.1109/ICIP46576.2022.9897615.
- [6] A. Mehmood, "Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks," *IEEE Access*, vol. 9, pp. 138283–138295, 2021, doi: 10.1109/ACCESS.2021.3118009.
- [7] X. Hu *et al.*, "A weakly supervised framework for abnormal behavior detection and localization in crowded scenes," *Neurocomputing*, vol. 383, pp. 270–281, 2020, doi: 10.1016/j.neucom.2019.11.087.
- [8] X. Zhang, Q. Yu, and H. Yu, "Physics inspired methods for crowd video surveillance and analysis: A survey," *IEEE Access*, vol. 6, pp. 66816–66830, 2018, doi: 10.1109/ACCESS.2018.2878733.
- [9] S. Chandrakala, K. Deepak, and V. L.K.P., "Bag-of-event-models based embeddings for detecting anomalies in surveillance videos," *Expert Systems with Applications*, vol. 190, 2022, doi: 10.1016/j.eswa.2021.116168.
- [10] W. Lin, J. Gao, Q. Wang, and X. Li, "Learning to detect anomaly events in crowd scenes from synthetic data," *Neurocomputing*, vol. 436, pp. 248–259, 2021, doi: 10.1016/j.neucom.2021.01.031.
- [11] N. Li, Y. B. Hou, and Z. Q. Huang, "Implementation of a real-time fall detection algorithm based on body's acceleration," *Journal of Chinese Computer Systems*, vol. 33, no. 11, pp. 2410–2413, 2012. [Online]. Available: <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XXWX201211019&DbName=CJFQ2012>
- [12] D. Pan, H. Liu, D. Qu, and Z. Zhang, "Human falling detection algorithm based on multisensor data fusion with SVM," *Mobile Information Systems*, vol. 2020, 2020, doi: 10.1155/2020/8826088.
- [13] Y. Zhong, X. Chen, Y. Hu, P. Tang, and F. Ren, "Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8285–8296, 2022, doi: 10.1109/TCSVT.2022.3190539.
- [14] C. Direkoglu, "Abnormal crowd behavior detection using motion information images and convolutional neural networks," *IEEE Access*, vol. 8, pp. 80408–80416, 2020, doi: 10.1109/ACCESS.2020.2990355.
- [15] J. Li, Q. Huang, Y. Du, X. Zhen, S. Chen, and L. Shao, "Variational abnormal behavior detection with motion consistency," *IEEE Transactions on Image Processing*, vol. 31, pp. 275–286, 2022, doi: 10.1109/TIP.2021.3130545.
- [16] S. Guo, Q. Bai, S. Gao, Y. Zhang, and A. Li, "An analysis method of crowd abnormal behavior for video service robot," *IEEE Access*, vol. 7, pp. 169577–169585, 2019, doi: 10.1109/ACCESS.2019.2954544.
- [17] Z. Ma, Y. Luo, C. B. Yun, H. P. Wan, and Y. Shen, "An MPPCA-based approach for anomaly detection of structures under multiple operational conditions and missing data," *Structural Health Monitoring*, vol. 22, no. 2, pp. 1069–1089, 2023, doi: 10.1177/14759217221100708.
- [18] T. A. Mangoli, C. Sujatha, and U. Mudanagudi, "Anomaly detection in surveillance video using motion-direction model," *Proceedings of 2018 2nd International Conference on Advances in Electronics, Computers and Communications, ICAECC 2018*, 2018, doi: 10.1109/ICAEECC.2018.8479508.
- [19] M. Hasan, J. Choi, J. Neumann, A. K. R. -Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 733–742, 2016, doi: 10.1109/CVPR.2016.86.
- [20] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," *IEEE International Conference on Multimedia and Expo*, pp. 439–444, 2017, doi: 10.1109/ICME.2017.8019325.
- [21] D. G. Lee, H. I. Suk, S. K. Park, and S. W. Lee, "Motion influence map for unusual human activity detection and localization in crowded scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1612–1623, 2015, doi: 10.1109/TCSVT.2015.2395752.
- [22] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2914–2922, 2017, doi: 10.1109/ICCV.2017.315.
- [23] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," *Advances in Neural Networks - ISNN 2017*, pp. 189–196, 2017, doi: 10.1007/978-3-319-59081-3_23.
- [24] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017, doi: 10.1016/j.cviu.2016.10.010.
- [25] T. Harada and H. Liu, "Online growing neural gas for anomaly detection in changing surveillance scenes," *Pattern Recognition*, vol. 64, pp. 187–201, 2017, doi: 10.1016/j.patcog.2016.09.016.
- [26] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding-based anomaly detection in stacked RNN framework," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 341–349, 2017, doi: 10.1109/ICCV.2017.45.
- [27] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection - a new baseline," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018, doi: 10.1109/CVPR.2018.00684.




- [28] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2405–2415, 2019, doi: 10.1109/TCSVT.2018.2864148.
- [29] G. Pang, C. Yan, C. Shen, A. V. D. Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12170–12179, 2020, doi: 10.1109/CVPR42600.2020.01219.
- [30] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pp. 2587–2596, 2020, doi: 10.1109/WACV45572.2020.9093417.
- [31] B. Ramachandra and M. J. Jones, "Street scene: a new dataset and evaluation protocol for video anomaly detection," *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pp. 2558–2567, 2020, doi: 10.1109/WACV45572.2020.9093457.
- [32] M. I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12737–12747, 2021, doi: 10.1109/CVPR46437.2021.01255.
- [33] S. Zhang *et al.*, "Influence-aware attention networks for anomaly detection in surveillance videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5427–5437, 2022, doi: 10.1109/TCSVT.2022.3148392.

BIOGRAPHIES OF AUTHOR



Mrs. Kusuma Sriram    received B.E. degree under VTU in 2004, and M.Tech. from RVCE in the year 2010. She is an active academician having 18 + years of teaching experience. Currently she is working at Ramaiah Institute of Technology, Bengaluru, as an Assistant Professor in the Department of Information Science and Engineering. She is pursuing Ph.D. in the area of video processing/machine learning. She is a trained Faculty of Infosys InfyTQ to motivate students for infosys job opportunities. She has worked as trainer for infosys campus connect program. She can be contacted at email: kusumas_12@rediffmail.com.



Dr. Kiran Purusotham    currently working as HOD, Department of Computer Science and Engineering. with total experience of 20 years. He has done research on privacy preserving data mining with focus on detection of sensitivity patterns and was awarded Ph.D. from VTU in 2014. His research interests include cryptography, randomization, anonymization methods in generalization, indexing techniques and design patterns. He has guided several M.Tech. and B.E. projects & internships and is currently guiding four research scholars at VTU. He can be contacted at email: kiranpmys@gmail.com.