Anomalous Event Detection Using LSTM Methodology

Abhishek Kumar[1], Sneha Singh[2], Sandeep Kumar[3]
[1,2,3]*Department Computer Science Engineering, Sharda University, Greater Noida, Uttar Pradesh, India*

Abstract- Anomaly detection is a significant issue that has been studied in a variety of study fields and application domains. It refers to the extraordinary occurrences, events, or observations that significantly deviate from the majority of the data and do not fit into a predetermined description of typical behavior. Anomalous events may be categorized as, driving vehicles on footpath, people running on road, snatching people's purse in public, etc. Such events need to be reported immediately in fear of getting too late to report such events if they become a big issue in future. Wide-spread uses for anomaly detection include military surveillance for enemy activities, insurance, or healthcare, detecting intrusions to improve cyber security, identifying faults in safety-critical systems, and fraud detection for credit cards, insurance, or healthcare. The process of anomaly identification can be challenging when applied to the analysis of event sequence data because the sequential and temporal character of such data gives rise to a variety of definitions and adaptable types of abnormalities. This in turn makes it more challenging to interpret abnormalities that are discovered.

However, this paper outlines an effective strategy for spotting irregularities in videos. Recent convolutional neural network applications, particularly in image recognition, have shown promise for convolutional layers. Contrarily, convolutional neural networks require labels as learning signals and are under supervision. Spatiotemporal architecture is thereby put forth for finding anomalies in films with packed situations. The two key parts of our architecture are one for representing spatial information and the other for understanding how the spatial features change over time. After the analysis of the anomalies, the conclusion of the test if the any abnormal event is detected is sent to nearby police station with fraction of seconds. So, if there is an inappropriate action then the police can take action in time.

Important Key Words: Anomaly, Anomalous events, abnormalities, convolutional neural network, spatiotemporal architecture.

1. INTRODUCTION

With the exponential development of video data, there is a rising demand to not only recognizes objects and behavior analysis, but also for identifying peculiar objects or suspect behavior in a sea of otherwise unremarkable data. For uses such as automated quality control and visual monitoring, finding such anomalies in videos is fundamental [1]. In order to maintain the security and safety for the general public, anomalous activity identification in video surveillance is a crucial task. [2] An autonomous system is desperately needed for recognizing suspicious occurrences that could pose a possible threat because the vast amount of video data makes it difficult for people to monitor and find anomalous events. A visible action or state change in a video stream that would be crucial for security management is referred to as an anomalous video even the duration of an anomalous video event can vary substantially. [1][2][3]. Recently, academics have paid increasing attention to anomalous event identification in intelligent video surveillance and industry worlds because to its connections to visual saliency, interestingness forecasting, recognizing dominating behavior, and other computer vision-related subjects. This task has grown in importance for intelligent video surveillance. [4] Due to the ambiguity of the criteria of regularity and irregularity, and their Depending on the backdrop circumstances, detecting aberrant events in video sequences is a difficult process. [4] Currently, one important component for identifying anomalous events in models already in use is feature extraction. The two primary categories of anomalous event detection models, in terms of feature representation, are models with extensive features and models with intricate characteristics. [5]. Considering the occurrences and interactions with great dimensionality, noise, and variety, video data presents a demanding representation and modelling problem [5]. Anomalies also have a strong contextual component; for instance, running in a dining establishment unusual, whereas taking a park run

would be typical. Furthermore, what constitutes an abnormality is sometimes imprecise. [6] Some people may believe that wandering around on a train platform should be reported as an anomaly since it may be suspicious, despite the fact that one may believe it to be typical. Due to these difficulties, machine learning techniques have difficulty spotting visual patterns that result in anomalies in practical applications. [7][8]

In the related field of action recognition, there are numerous successful examples. These techniques, however, are only appropriate for labelled video footages in a lucid defined important occasions and without heavily occluded areas, like crowded scenes. [9] Additionally, identifying each sort of occurrence is highly expensive. However, it cannot be without fail include all recent and upcoming events. The length of the video recordings that was captured is probably insufficient to capture all kinds of actions, particularly unusual ones that happened infrequently or not at all. [10]. Contemporary study on finding alterations by categorizing the assignment as either normal or abnormal has shown it to be accurate and useful, however the applicability of such a strategy is constrained because footages of anomalous events are hard to come by because they are uncommon. As a result, many academics are now focusing on models like autoencoder, dictionary education, and spatiotemporal features that might be trained with minimal to no oversight. [9] Contrary to supervised approaches, these procedures just call for video with no labels footages that contain not many or few anomalous incidents and are easy to impart in practical implementations. The next section goes into a discussion of various approaches as well as their drawbacks. [10]

The methodology presented in this research uses deep learning to automatically infer a set of general features from a lot of video data to represent video data. In particular, video frames were processed in an unsupervised manner by an intricate neural network made of collection of convolutional autoencoders, capturing the spatial data structures that, when gathered together, make up the video representation. Then, in order to discover the regular temporal patterns, Convolutional temporal autoencoders are fed this representation in a stack.

Suggested approach may be simply adapted to many settings and is domain unlocked (i.e., unrelated to a particular task, necessitating no domain specialist). It also requires no additional human effort. Application of suggested method to real-world datasets to demonstrate its efficacy, and thus demonstrate that it regularly beats competing procedures along with preserving a quick running time.

There are few primary attributes of methodology and the contributions made by this study like It will help to lessen the time-consuming effort involved in feature engineering to produce a data representation that supports effective machine learning. This can be accomplished by substituting learned hierarchical features for low-level handmade features. Instead of creating appropriate features based on our expertise, we are able to find representative features with the aid of autoencoders. Also, Autoencoders are used in place of conventional sparse coding techniques. In contrast to current methods, learning a model of features and extracting feature representation from videos are done simultaneously. Through this paper, Hierarchical feature learning can also be accomplished by using an autoencoder with numerous layers of hidden units. The time required to report a crime or event will be shortened because the last phase of this project will concentrate on alerting relevant parties if any anomalous event is discovered.

2. RELATED WORK

The majority of these aberrant occurrences are typically unknown beforehand because doing so would require foreseeing every possible way that anything could go awry. As a result, learning a model for any situation is simply not possible. abnormality or irregularity. But without knowing what to search for, how can we find an anomaly?

In video analysis and anomaly identification, trajectory analysis has long been used. Training phase theoretical object trajectory deviations and assessment phase comparisons of brand-new test trajectory deviations from the theoretical classes are common features of trajectory-based techniques. An anomaly is defined as a substantial amount of data departure across all courses. [11] Path analysis, however, significantly depends on monitoring, which is still an important factor of difficulty for visual computing,

especially in complicated scenarios. While spotting-based methods work well in scenarios with few items, they are not practical for spotting unusual patterns in busy or complicated environments. [12]

There are also non-spotting methods which concentrate on the multi-scale irregularities in videos. Using spatiotemporal video volumes, they mostly rely on collecting and examining local low-level visual data such as optical flow histograms, histograms of directed gradients, and optical bow (selecting interest points or using dense sampling). Then, using similarity criteria, these local features are categorized into clusters, or bags of visual words (BOV). [13]. Their low computing cost and capacity to detect anomalous behavior even in densely populated situations account for their popularity. [14] Sparse reconstruction is a related method. Any updated design that depicts typical or abnormal event may be thought of as being combined linearly for feature representations in a trained dictionary. This is the basic fundamental presumption of these techniques. This is based on the supposition that all previously recorded incidents were typical.

Deep learning techniques have become increasingly popular in anomaly detection as a result of their success in a variety of applications. Contrary to earlier methods, Deep learning techniques require extremely minimal preprocessing to extract the important characteristics from the input. [12] [13] As a result, no particular set of characteristics must be defined in order to extract data from the dataset. Convolutional neural networks (ConvNet) in particular have demonstrated their efficacy in a variety of applications counting object identification, human detection, and activity recognition. [1][3] [25] In essence, a convolutional auto encoder is a ConvNet with a mirror image of the classifier and fully-connected layer. convolutional layer stacks with a Soft ax classifier and a fully connected layer make up ConvNet. [20] [23] [24] LSTM models, on the other hand, are widely known for having the ability to recognize temporal patterns and forecasting data over time. has just suggested using convolutional LSTMs to understand the predictable temporal patterns in videos, and his research shows that deep neural networks can learn some really interesting things. [19]

However, because traditional BOV methods group comparable volumes, they obliterate entire compositional data in the course of organizing pictural words. Additionally, the number of clusters must be predetermined; this can only be done during testing period by trial and error. Codebook models are also problematic for real-time anomaly detection since they demand searching over a wide area even during testing. [25]

These freshly presented methods still have certain drawbacks despite how straightforward they are. The lack of video segments with aberrant occurrences makes it impractical to use 3D ConvNet in real-world applications, despite its remarkable ability to distinguish between anomalies and regular occurrences using distinguishing characteristics. [20] Operations like as convolution and pooling are only performed spatially in the convolutional autoencoder presented by, despite the fact that the envisioned network accepts a number of input frames. This is because 2D convolutions totally collapse temporal information after the rest convolution layer. [16] Additionally, because the convolutional LSTM layers utilized by must be trained in tiny mini-batch sizes since they need a lot of memory., which prolongs training and testing time. [17]

3.METHODOLOGY

The approach presented here is predicated on the idea that, in the event of an anomalous event, the video's most recent frames will change significantly from its earlier frames. A temporal encoder-decoder and a spatial feature extractor work together to learn the time trends of the number of frames supplied as part of a complete model that we train in the spirit of. The model is trained using only regular scene-containing video volumes with the goal of lowering the discrepancy in how well the learned model reconstructed the input and output video volumes. once the model has undergone proper training, it is anticipated that normal video volume will possess minimal reconstruction error while unusual scene-based video volume will possess an elevated Reconstruction blunder. Our system thresholds the error generated by each testing input volume. can determine whether an abnormal event occurs. [17] [20] [2] [23]

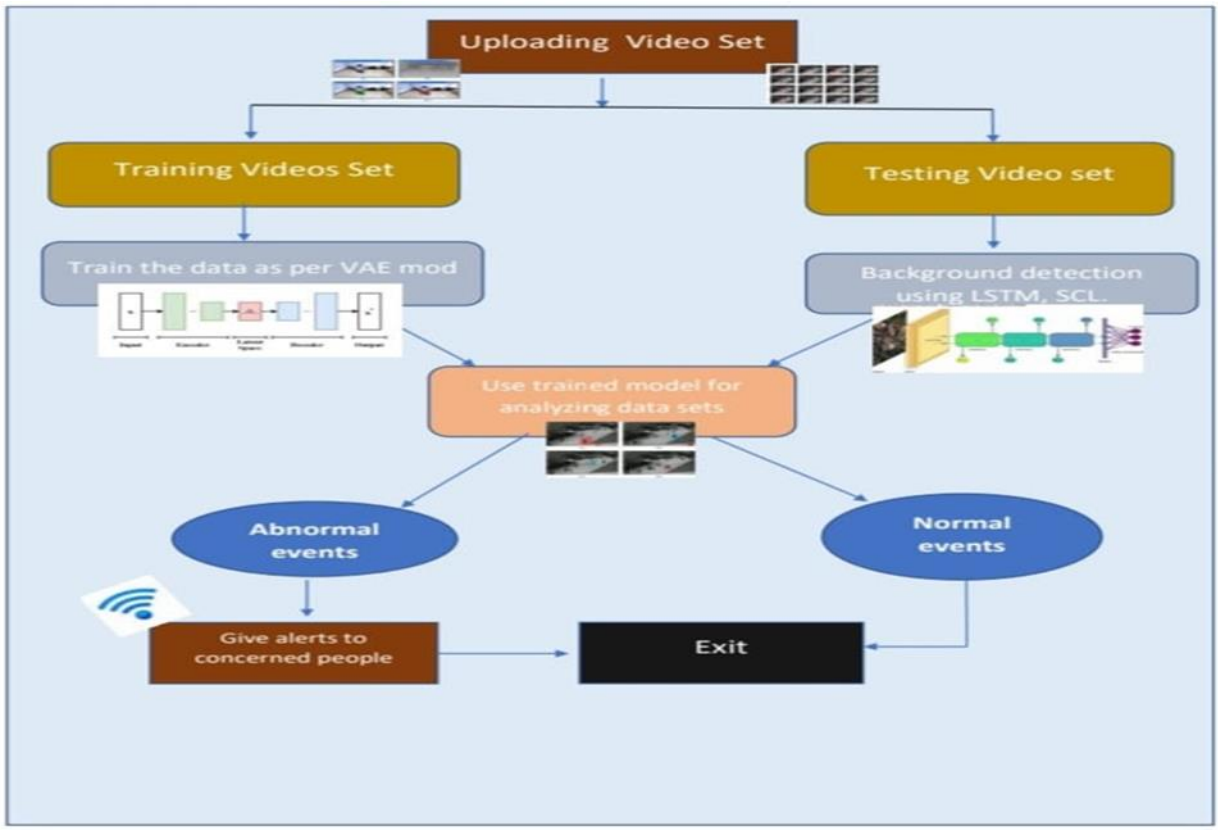

Figure 1:Methodology of anomalous event detection

A. Preprocessing

During this phase, raw data will be transformed into a model input that is aligned and acceptable. From the raw video, each frame is taken out and scaled to 227 X 227. The pixel values are scaled between 0 and 1 and subtracted from each frame's global mean picture for normalization in order to guarantee that the input images are all on the same scale. The training dataset's pixel values are used to produce the mean picture. averaged across all locations within each frame. To reduce dimensionality, the photos are then changed to grayscale. The images are processed and then normalized to have a mean value and variance value of 0 and 1, respectively. Video volumes are used as the model's input; each volume contains 10 consecutive frames that have varying skip lengths. This model has a lot of parameters, so huge amount of training data is needed. To expand the amount of the training dataset, we carry out data augmentation in the temporal dimension.

B. Feature learning

To discover the recurring patterns in the training films, we suggest using a convolutional spatiotemporal autoencoder. The two parts of the architecture that are proposed are a to learn the spatial patterns of the spatial structures encoded, a temporal encoder-decoder is used, along with a spatial auto encoder to learn the spatial structures of each video frame. The spatial encoder and decoder each have two convolutional and DE convolutional layers, as opposed to the temporal encoder, which has a three-layer convolutional long short term memory (LSTM).) model.[22] While LSTM models are frequently used for time-series modelling and sequence learning and have proven their efficacy in applications like speech recognition and handwriting recognition, The outstanding performance of convolutional layers in the area of object identification is widely recognized.

- Autoencoder

Encoding and decoding are the two phases of autoencoders, as the name suggests. By making the encoder output units fewer than the input units, it was originally employed to minimize dimensionality. Back-propagation is typically used to train models in an unsupervised fashion, reducing the decoding outputs' reconstruction error from the original inputs.

More usable features than other popular linear transformation techniques, such as PCA can be extracted by an autoencoder, if the non-linearity of the activation function is intended.

➢ Spatial Convolution

In the instance of a convolutional network, convolution's primary purpose is to take the input image's characteristics and extract them. By extracting visual information from small squares of input data, convolution retains the spatial relationship between pixels. The values of these filters are learned by a convolutional network on its own during the training process, albeit we still need to define the parameters., just like the quantity, size, and number of filters it had before training. The more filters we have, the more picture attributes can be recovered, and the network gets better at identifying patterns in previously unseen images. We must maintain equilibrium by not using an excessive number of filters because more filters will lengthen calculation times and deplete memory more quickly.

➢ Long Short-Term Memory (LSTM)

A RNN version termed the LSTM model, which includes a forget gate recurrent gate, is introduced to address this issue. With the new structure, long sequences may be processed using LSTMs, which can also be stacked together to gather more comprehensive data because backpropagated errors are prevented from disappearing or bursting.

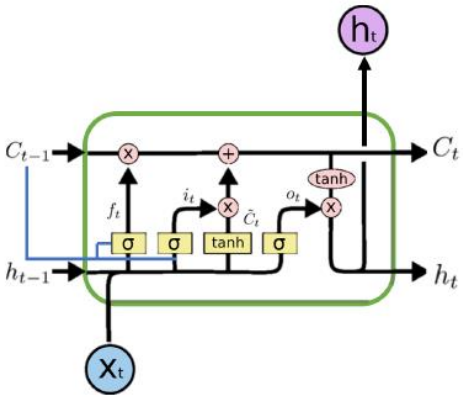

Figure 2: Structure of LSTM unit [15]

➢ Variational autoencoders (VAEs)

Latent representations can be learned using the deep learning method known as variational autoencoders (VAEs). Additionally, they have been utilized to interpolate between phrases, produce state-of-the-art semi-supervised learning outcomes, and draw pictures. On VAEs, there are numerous online tutorials. The primary distinction between the architectures of the VAE and the Autoencoder is that the input is encoded into two vectors rather than just one. The latent representation of the input is then derived from the normal distribution defined by these two vectors.

C. Regularity score

After the model has been trained, testing data may be fed into it to see how well it performs in terms of recognizing abnormal occurrences while preserving a low rate of false alarms to calculate the regularity score across all frames, we utilized the same technique to more accurately compare with, with the trained model being of a different type.

Anomaly detection

Thresholding

Identifying a normal or abnormal video frame is a simple process. Each frame's reconstruction mistake affects whether or not it qualifies as an unusual frame. The detecting system's responsiveness is controlled by the threshold. For instance, if the threshold is too low, the system is more responsive to events taking place in the scene, resulting in added alarms being generated. To determine the ratios of true positives and false positives are calculated using various error thresholds based on the area under the ROC curve (AUC). In

order to achieve the same error rate (EER), there are an equal number of false positives and false negatives.

Event count

With a fixed temporal window of 50 frames, we used the Persistence1D approach to group local minima in order to decrease the number of noisy and meaningless minima in the regularity score. Local minima that show up repeatedly within 50 frames are thought to represent related anomalies. Here is an appropriate duration for the temporal window because an anomalous occurrence needs to last at least 2-3 seconds to be significant.

4.RESULTS

The primary benefit of this approach is how effectively it extracts optical flow features in tandem with the integration of a regularisation term for compactness during training. As compared to state-of-the-art approaches, this method has very high-performance experimental results and shows promise for the detection and localisation of anomalies via surveillance. By examining and condensing the DL techniques used in AD for video streaming, this review aims to make a substantial research contribution to the study of DL in the intelligent surveillance area. The first category looked at how many frames were required for detection, while the second one at how many anomalies were present in a scene. In addition, our study classified prominent DL algorithms for anomaly detection into groups based on the network type and architectural style in order to examine their effectiveness for doing so. Also, a thorough list of the benchmark datasets and performance indicators used to assess the efficacy of DL techniques was provided. Also, our work emphasised the main problems and applicability of DL-based AD techniques. Furthermore, the paper is certain that this study will be beneficial to the community researchers working on this issue in comprehending this important field of research. Our primary objective was to inspire researchers to conduct further study in this field so that it can advance in the near future.

Figure 3: Anomaly detection in public [4]

Table 2: Anomalous event and false alarm count detected by different methods on various event type in Avenue dataset.

	Run	Loiter	Throw	Opposite Direction	False Alarm
Groundtruth	12	8	19	8	0
Ours	10	8	19	7	12

.

Table 3: Anomalous event and false alarm count detected by different methods on various event type in Subway Entrance dataset. WD: wrong direction; NP: no payment; LT: loitering; II: irregular interaction; Misc.: miscellaneous.

	WD	NP	LT	II	Misc.	False Alarm
Groundtruth	26	13	14	4	9	0
Ours	24	10	14	4	9	9

Table 4: Anomalous event and false alarm count detected by different methods on various event type in Subway Exit dataset. WD: wrong direction; LT: loitering; Misc.: miscellaneous.

	WD	LT	Misc.	False Alarm
Groundtruth	9	3	7	0
Ours	8	3	7	10

Table 5: Details of run-time during testing

	Time (in sec)			
	Preprocessing	Representation	Classifying	Total
CPU	0.0010	0.2015	0.0002	0.2027 (∼5fps)
GPU	0.0010	0.0058	0.0002	0.0070 (∼143fps)

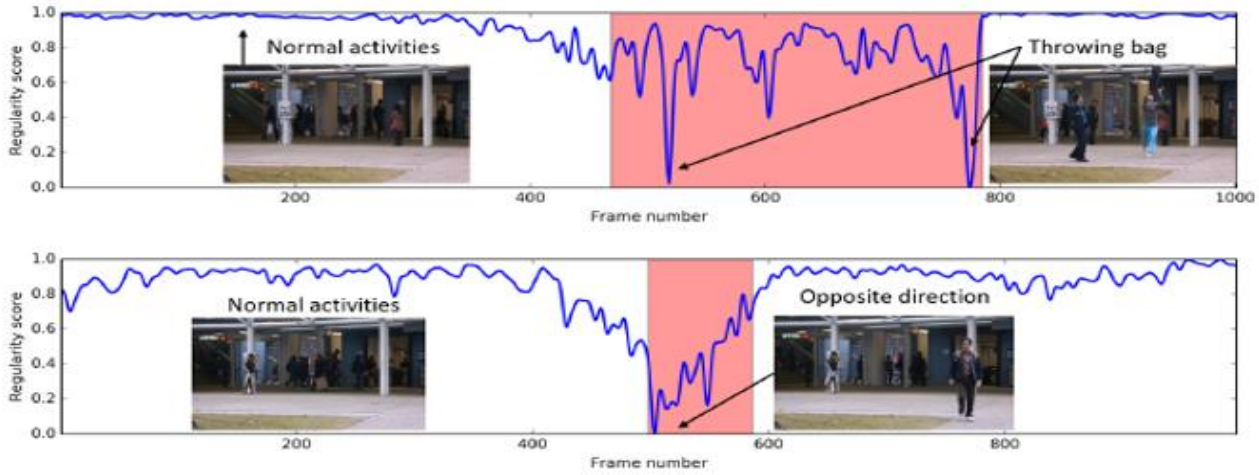

Figure 4: Regularity score of videos #5 (top) and #15 (bottom) from the Avenue dataset.

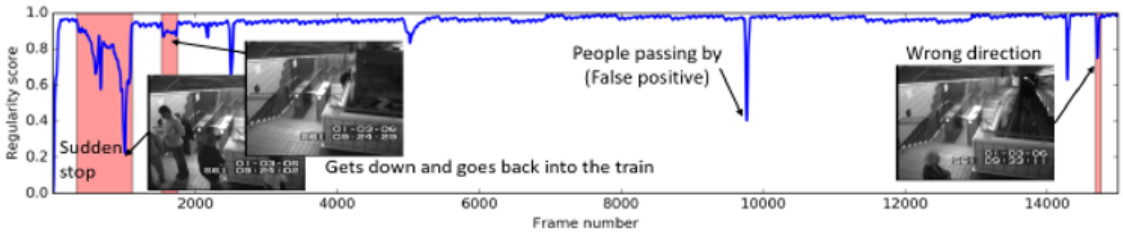

Figure 5: Regularity score of frames 22500-37500 from the Subway Entrance video.

The output of the suggested system is shown in Figures 4, 5, for samples from the Avenue dataset, Subway entry scenes, and departure scenes, respectively. Our technique accurately finds abnormalities in these circumstances even in busy scenarios. A low regularity score is indicated by the fact that almost all anomalies result in significant downward spikes.

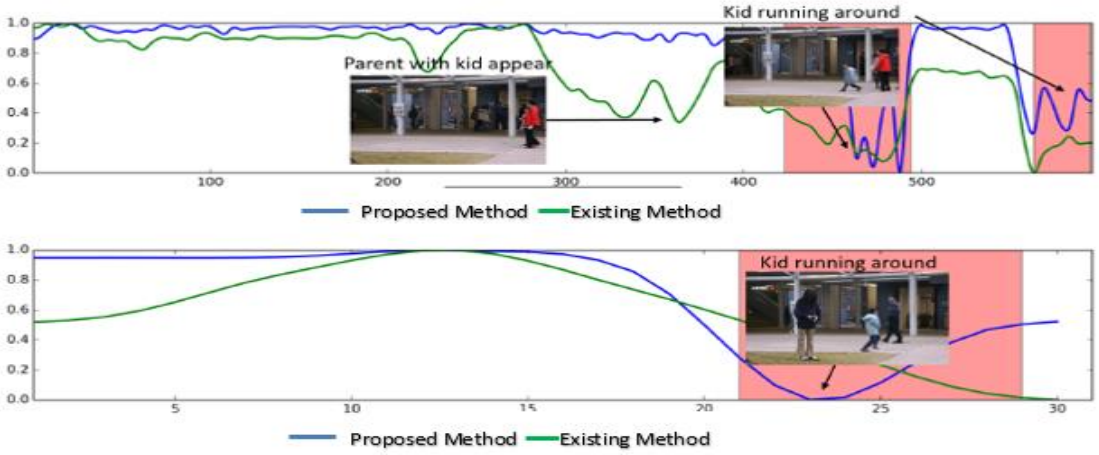

Figure 6: Comparing our method with ConvAE on Avenue dataset video #7 (top) and #8 (bottom).

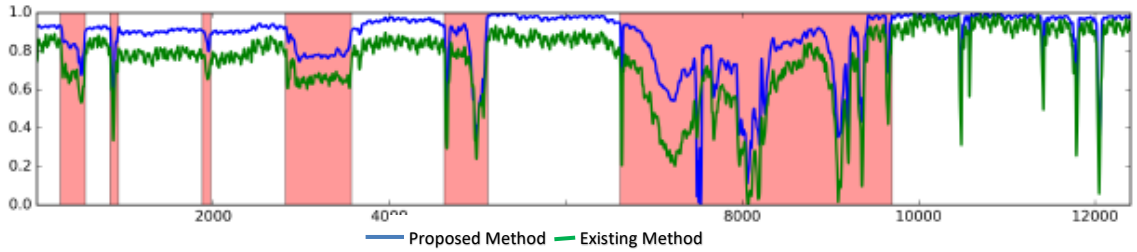

Figure 7: Comparing our method with ConvAE on Subway Exit video frames 10000-22500.

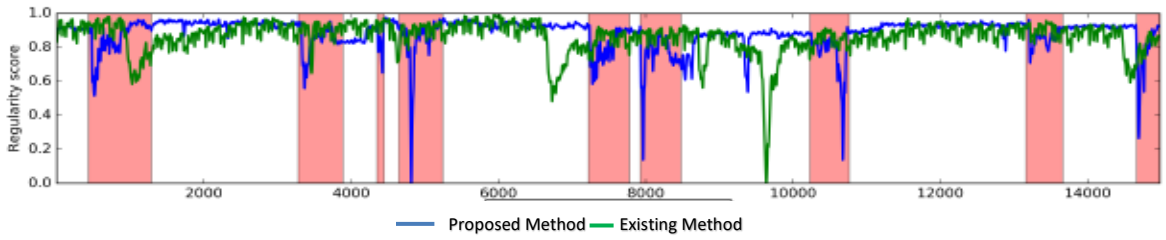

Figure 8: Comparing our method with ConvAE on Subway Entrance video frames 120000-144000.

From figure 6 and 7, It is evident that, in comparison to convAE our strategy has found more aberrant occurrences while generating less false alarms as depicted by more number of downward spikes and lower regularity score. From figure 8 it is depicted that the system is capable of producing greater regularity ratings during typical activities and lower scores when there are irregularities as can be seen by downward spikes.

5.FUTURE SCOPE & CONCLUSIONS

A video sequence is used to extract spatiotemporal characteristics via the convolutional LSTM autoencoder (unsupervised), which then uses this

information to learn regularity. Depending on the camera configuration, it can identify unusual occurrences and activities, but because it is unsupervised, it can be educated over time and get better on its own. It works well with stationary camera setups and busy locations. In semi-supervised hybrid classification, the aforementioned encoder performs as a high recalled, and anomalies are routed through a false positive reduction model. a neural network with deep learning good accuracy and recall results from this combination In this study, a sparse combination learning-based approach to anomalous event detection is proposed. With this method, sparse combinations are learned, which dramatically speeds up testing without sacrificing efficacy. To handle massive amounts of data, an online solution is also available. In several datasets, our technique produces cutting-edge outcomes. It is similar to yet significantly different from conventional subspace clustering. In addition to providing a fundamentally new knowledge of video structures, we think it will have a significant positive impact on numerous applications. It supports different types of cameras. It is anticipated to be helpful in the quick identification of incidences of minor crimes and other unusual occurrences.

Considering one of the most effective uses of computer vision is for automated industrial anomaly detection, the advancements made possible by Patch-Core can be of particular interest to experts in this field. Negative societal impact is constrained because our work is primarily focused on industrial anomaly identification. Furthermore, even though the fundamental strategy may be used for detection systems in more contentious fields, we don't think our advances are substantial enough to alter the way such systems are used in society.

REFERENCES

[1] Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, A Large-Scale Video Benchmark for Human Activity Understanding, (2015).

[2] Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P. The Kinetics Human Action Video Dataset (2017).

[3] Luo, W.; Liu, W.; Gao, S. A Revisit of Sparse Coding Based Anomaly Detection in Stacked Rnn Framework (2017).

[4] Yong Shean Chong Yong Haur Tay, Abnormal Event Detection in Videos using Spatiotemporal Autoencoder, (2017)

[5] Waqas Sultani, Chen Chen, Mubarak Shah, Real-world Anomaly Detection in Surveillance Videos, (2017)

[6] Sultani, W.; Chen, C.; Shah, M. Real-World Anomaly Detection in Surveillance Videos, (2018)

[7] Xing Hu, Yingping Huang*, Qianqian Duan, Wenyan Ci, Jian Dai1 and Haima Yang, Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor, (2018)

[8] S. Arif Ahmed, D. Prosad Dogra, S. Kar, and P. Pratim Roy, Trajectory-based surveillance analysis: A survey, (2019).

[9] S. Lee, H. G. Kim, and R. M. Ro, BMAN: bidirectional multiscale Aggregation networks for abnormal event detection, (2019).

[10] T. Gupta, V. Nunavath, and S. Roy, CrowdVAS-net: A deep-CNN based framework to detect abnormal crowd-motion behavior in videos for predicting crowd disaster, (2019).

[11] Soliman, M.M.; Kamal, M.H.; Nashed, M.A.E.-M.; Mostafa, Y.M.; Chawky, B.S.; Khattab, D. Violence Recognition from VideosUsing Deep Learning Techniques, (2019).

[12] Lindemann, B.; Maschler, B.; Sahlab, N.; Weyrich, M. A Survey on Anomaly Detection for Technical Systems Using LSTM Networks, (2019).

[13] Tian Wang, Zichen Miao, Yuxin Chena, Yi Zhoub, Guangcun Shan*, Hichem Snoussid, Aed-net: An abnormal event detection network, (2019).

[14] Wu, P.; Liu, J.; Shen, F. A Deep One-Class Neural Network for Anomalous Event Detection in Complex Scenes, (2019).

[15] Cewu Lu, Jianping Shi, Weiming Wang, Jiaya Jia, Fast Abnormal Event Detection, (2019).

[16] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos, (2020).

[17] O. Ye, J. Deng, Z. Yu, T. Liu, and L. Dong, Abnormal event detection via feature expectation subgraph calibrating classification in video surveillance scenes, (2020).

[18] Mandal, M.; Kumar, L.K.; Vipparthi, S.K. Mor-Uav: A Benchmark Dataset and Baselines for Moving Object Recognition in UavVideos, (2020).
[19] Ou Ye, Iun Deng, Zhenhua Yu, Tao Liu, and Lihong Dong, Abnormal Event Detection via Feature Expectation Subgraph Calibrating Classification in Video Surveillance Scenes, (2020).
[20] Dimitris Tsiktsiris, Nikolaos Dimitriou, Antonios Lalas, Minas Dasygenis, Konstantinos Votis 1 and Dimitrios Tzovaras, Real-Time Abnormal Event Detection for Enhanced Security in Autonomous Shuttles, (2020).
[21] Maschler, B.; Weyrich, M. Deep Transfer Learning for Industrial Automation: A Review and Discussion of New Techniques for Data-Driven Machine Learning, (2021).
[22] Mandal, M.; Vipparthi, S.K. An Empirical Review of Deep Learning Frameworks for Change Detection: Model Design, Experimental Frameworks, Challenges and Research Needs. (2021).
[23] Qinmin Ma, Abnormal Event Detection in Videos Based on Deep Neural Networks, (2021).
[24] Weichao Zhang, Guanjun Wang, Mengxing Huang, Hongyu Wang, Shaoping Wen, Generative Adversarial \Networks for Abnormal Event Detection in Videos Based on Self-Attention Mechanism, (2021).
[25] Jongmin Yu, Minkyung Kim, Hyeontaek Oh, Jinhong Yang, Real time abnormal insider event detection on enterprise resource planning systems via predictive auto-regression model, (2021)