IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Identification Of Anomalous Activities On Videos Using Time Distributed Auto-Encoder

[1]Suman Mondal, [2] Arindam Roy, [3] Sukumar Mondal
[1]Assistant Professor, [2]Associate Professor, [3]Associate Professor
[1]Department of Computer Science,
[1]Yogoda Satsanga Palpara Mahavidyalaya, Palpara, Purba Medinipur, India

Abstract: A major issue for video monitoring systems is the consecutive flow of unending data that may be investigated. However, a comprehension test makes no sense when using actual facts. Convolutional layers are promised by contemporary applications of convolutional neural networks for classifying and evaluating various objects. In order to be trained, neural networks need a number of classes that are grouped according to features. Here, we create a time-distributed auto-encoder initiation for the detection of anomalous occurrences in recordings of situations including natural mobility. A spatial feature illustration and an establishing temporal expansion of the spatial features are the two main modules of our introduction. The suggested method's precision is confirmed by empirical data from evaluations on the UCF-Crime Dataset and the UCSD Anomaly Dataset.

Index Terms - Convolutional Layer; Anomalous Occurrences; Time-Distributed; UCF; UCSD.

I. INTRODUCTION

With the huge volume of experimental data, the rapid rise of video data has shown rising interest in the recognition of visual objects and its potential applications in security and visual monitoring systems [Dee, Velastin, (2008), Hu, Tan, Wang, & Maybank, (2004) and Moeslund, Hilton, & Krüger, (2006)]. These methods have relied on human input to identify typical and abnormal events. One of the most difficult tasks in a constantly changing environment is to automatically send alarms on odd events without human intervention. From impulsive detection to video monitoring, finding such anomalies in videos is crucial.

Rarely have anomalous events—which could involve one or more personal objects—occurred in lengthy video sequences. The most difficult task for humans to perform has been to identify anomalies in videos. These promote the use of an automatic identification technique and segmentation of the video sequence's pattern in environments with complex layouts, such as when a car makes a U-turn on a one-way street or someone is jogging as people are strolling around them. Due to the vast spatial and chronological context they contain, these abnormalities are more challenging to spot. As a result, synchronous technology requires more layout work on the video in order to identify the heuristics [Brand, & Kettnaker, (2000), Loy, Xiang, & Gong, (2008) and Nguyen, Venkatesh, & Bui, (2006)], which makes it simple to recognize.

As a result of its high dimensionality, diversity of occurrences, and noise, video evidence presents a difficult set of examples for the model. Jogging in a crowded place would be an oddity, but running in a field would be expected. This is how contextually significant anomalies are. The term "anomaly" can, however, be clearly defined and is frequently found using template or trajectory matching [Bobick, & Davis, (2001) and Johnson, & Hogg, (1995)]. Imagine, for instance, that someone moving around on a train platform is usual, but that other people might anticipate it to be flagged as an abnormality because it might not be important. Due of these circumstances, it is difficult for machine learning approaches to recognize the patterns that give rise to anomalies in real-world scenarios.

There are many effective techniques for recognizing events [Tran, Bourdev, Fergus, Torresani, & Paluri, (2015), Karpathy, Toderici, Shetty, Leung, Sukthankar, & Li, (2014), Ji, Xu, Yang, & Yu, (2013) and Oneata, Verbeek, & Schmid, (2013)]. Furthermore, these methods can be used with labeled data that shows regions of interest but doesn't include a lot of blocky images, such as traffic data. It comes at a considerable cost to mark each type of common event, though. Therefore, there is no guarantee that previous and future occurrences will be included in the same way. Additionally, not all movements can be captured by this video clip because they are rarely captured on camera.

Though error-free and helpful, binary classification [Zhou, Shen, Zeng, Fang, Wei, & Zhang, (2016) and Mondal, Roy, & Mondal, (2023)] on a video series may miss a few unusual anomalous events. Numerous researchers, such as auto-encoders [Sabokrou, Fathy, Mojtaba, & Klette, (2015)], spatiotemporal features [Lu, Shi, & Jia, (2013) and Zhao, Li, & Xing, (2011)], etc., have shifted their focus to training without supervision. These techniques have been used on unlabeled data to find uncommon aberrations. The next section provides a brief explanation of various methods.

In this paper, a creative framework is shown for automatically extracting a set of video data features using deep learning methods. A neural network is made up of several time-distributed auto-encoders that work together to analyze video frames autonomously and extract spatial characteristics that combine to create the actual video. Then, in order to understand organized temporal patterns, this philosophy is given to temporal auto-encoders.

Our suggested model does not call for human involvement and is not connected to any particular domain. Specifically, demonstrate the efficiency of the pred model we used in the method to existing datasets and outperforms other cutting-edge approaches.

1.1 OUR ENDOWMENTS

The following are the main characteristics of our representation and the endowments of this investigation:

- We want to use a practical machine learning approach to enhance the research findings. Instead of using low-level handcrafted characteristics, these are employed as hierarchical features. Instead of modifying other aspects to fit our knowledge, we are effective in identifying the optimal attributes that enable us to approach the data's true outcomes.
- We substitute auto-encoders for the conventional coding method. Contrary to the existing features, there is no distinction between learning the model using the elements and extracting the characteristics. Additionally, the buried layers of auto-encoders enable to extract the hierarchical features.
- Unlike other auto-encoders, we employ time-distributed auto-encoder in this case to enhance the real results of all datasets. Time-distributed auto-encoder reflect the hierarchical spatiotemporal features among the video frames.

II. RELATED WORKS

It would be possible to anticipate when something deviates from the standard since almost all abnormalities are previously unspecified. Sadly, it is not realistic to be able to learn about every non-uniform event. So, the biggest issue for every researcher would be how to locate the true anomalies.

We are concentrating on training exclusively on conventional video patterns because, in most cases, picking aberrant does not make it more difficult to characterize the typical picture. After anomalies detect when they depart from the regular ones, researchers in this field generally take a broad strategy to picking up the normal behaviors first from the training set [Lu, Shi, & Jia, (2013), Zhao, Li, & Xing, (2011), Cong, Yuan, & Liu, (2011) and Li, Han, Ye, & Jiao, (2011)]. To build a model after the train, most research relies on local features [Mondal, Roy, & Mandal, (2021)] derived from moving frames.

In video analysis and anomaly detection, driving object directions are well-known [Li, Han, Ye, & Jiao, (2011), Zhou, Shen, Zeng, & Zhang, (2015), Piciarelli, Micheloni, & Foresti, (2008) and Mo, Monga, Bala, & Fan, (2014)]. The ability to take into account all of the trajectory's deviation in the training segment and apply a new assessment course to the testing segment is a crucial characteristic of trajectory-defined approaches. Furthermore, the trajectory-defined analysis is highly dependent on object tracking in difficult conditions, which is a major computer vision issue. Almost none of the scene's objects fit this kind of technique for spotting odd patterns in a complicated context.

There are now also various non-trajectory based methods for spotting spatial and temporal irregularities. These rely on the analysis of local visual aspects, such as direction flows utilizing histogram [Laptev, Marszalek, Schmid, & Rozenfeld, (2008)], HOG (Histogram Oriented Gradient) [Xiao, Zhang, Zha, & Wei,

(2014)], and the selection of attractive places [Reddy, Sanderson, & Lovell, (2011)] using spatial and temporal recordings. Local characteristics come together to form groups using affinity measures, creating a BOV or bag of visual words. Acceptance of these methods depends on their cheaper cost and ability to focus in unusual circumstances even in difficult conditions [Kratz, & Nishino, (2009)]. Another method is called sparse reconstruction, in which the freshly trained vocabulary is modelled using additional features that indicate common or rare events. These regard all one-time occurrences as commonplace.

However, when acquiring optical words, it completely eliminates all the pertinent information. Giving the number of clusters during the testing process is one of the major issues. It is impossible to search for a large space while simulating a real-world setting.

In a field like anomaly detection, enormous applications of deep learning techniques have achieved significant success. Understanding deep learning provides a learning method that uses hidden nodes in the network for hierarchical features, as stated methodologies are not necessary to characterize different features from the data. Neural networks, on the other hand, assist in extracting meaningful features from data with minimal preprocessing. Moreover, some of the constructive applications related deep learning are recognition of objects [Simonyan, & Zisserman, (2015)], activity recognition [Tran, Bourdev, Fergus, Torresani, & Paluri, (2015), and Simonyan, & Zisserman, (2014)], and human detection [Vu, Osokin, & Laptev, (2015)]. Hierarchical layers with a fully linked layer and a SoftMax classifier are part of deep understanding. A neural network called an auto-encoder replaces convolutional layers with a fully connected layer and a classifier. Anomalies in contracts are categorized using three-dimensional convolutional neural networks [Zhou, Shen, Zeng, Fang, Wei, & Zhang, (2016)], while other researchers employ an end-to-end auto-encoder [Hasan, Choi, Neumann, Roy-Chowdhury, & Davis, (2016)] to find anomalies in a video surveillance system. Their conclusion demonstrates the applicability of learned video representation using convolutional layers. On the other hand, the extended short-term memory model LSTM has gained popularity for learning with temporal features and forecasting time streaming data. In [Medel, (2016)], when material features are used for learning in video anomaly detection, a promising result has been found.

Despite clarity, those contemporary methods nevertheless have few limitations. Although a three-dimensional neural network can distinguish between abnormal and normal occurrences with ease thanks to its distinct properties, it cannot be used in real-time when there are no unusual events in the video clips. Convolutional and pooling methods that are spatially executed were proposed in [Hasan, Choi, Neumann, Roy-Chowdhury, & Davis, (2016)]. Following the complete loss of the initial convolutional temporal information, our suggested method uses a variety of frames as input. Other than this, the LSTM layers run by [Medel, (2016)] require a lot of memory, with tiny batches being used throughout training time and modest training and testing intervals.

III. MECHANISM

The method presented here is the foundation for the notion that the video's most recent frame is astonishingly independent of other structures. We were inspired by [Chong, & Tay, (2017)] to conduct our training using spatial characteristics and a temporal encoder-decoder that concurrently learns from the input objects' material composition. The goal of this model is to reduce the video reconstruction error between the input and output through training. In cases where the video contains an odd scenario, the model would anticipate a high reconstruction error even after it has been fully trained. By averaging the error in the input video, our method will find the anomaly when it happens.

Our implementation categorizes into three stages:

3.1 PRE-PROCEDURE

We discuss the allowed input for our implementation in this section. Frames from raw videos are first extracted and transformed to 280 × 280. Following then, each frame's pixel values vary from 0 to 1. The image then becomes grayscale. Then, training and testing are conducted using the processed images.

Ten continuous frames make up the input volume. For training, the system needs a lot of data. We rely on steps to create the enormous volume; for example, stride-2 has the frames {1,3,5,7,9}, and stride-4 contains the frames {1,5,9,13,17}. The input then prepares the model for training.

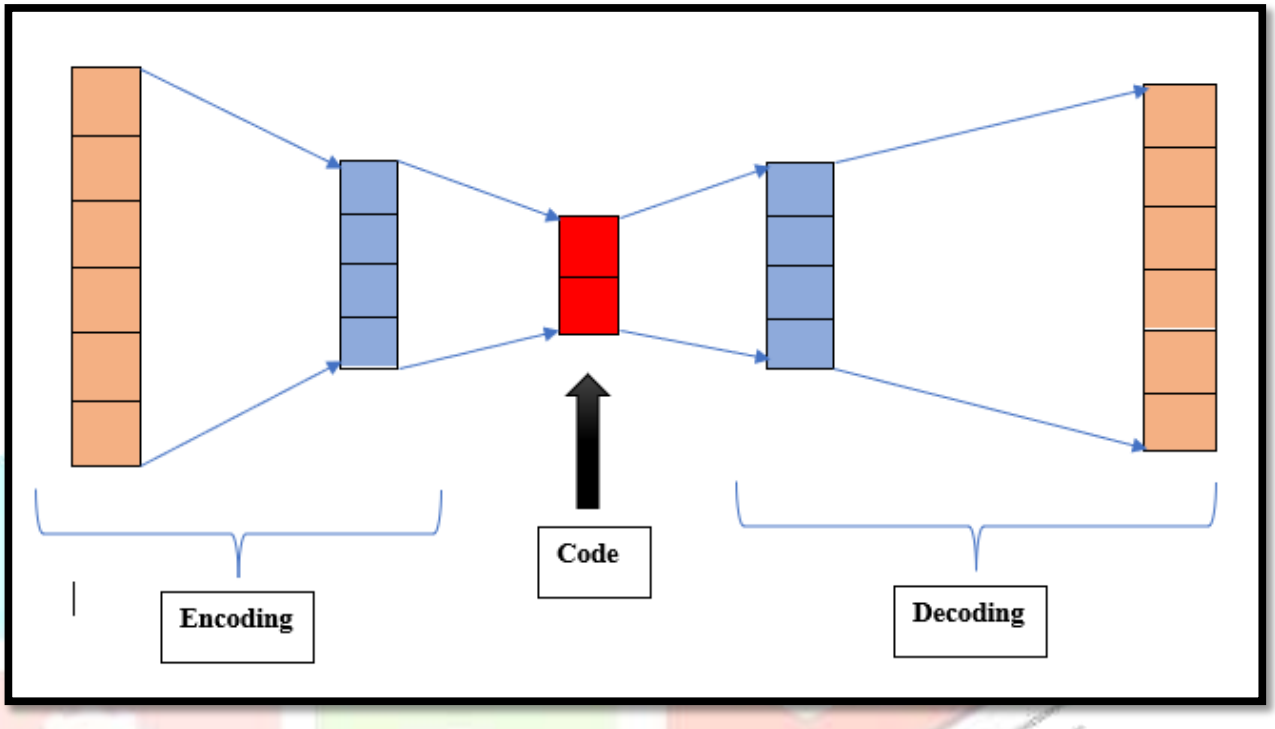

Fig. 1. *Structure of Auto-encoder*

3.2 FEATURE UNDERSTANDING

Encoder utilized to understand the training videos' input. Four pieces make up our defined implementation. To outperform other techniques, you must first comprehend the spatial structure of each video frame before learning its temporal rhythm. These processes must be repeated twice. The structure of the auto-encoder used in our implementation is shown in Figure 1. We have whole-cycle auto-encoder, one with convolutional layers and the other with deconvolutional layers.

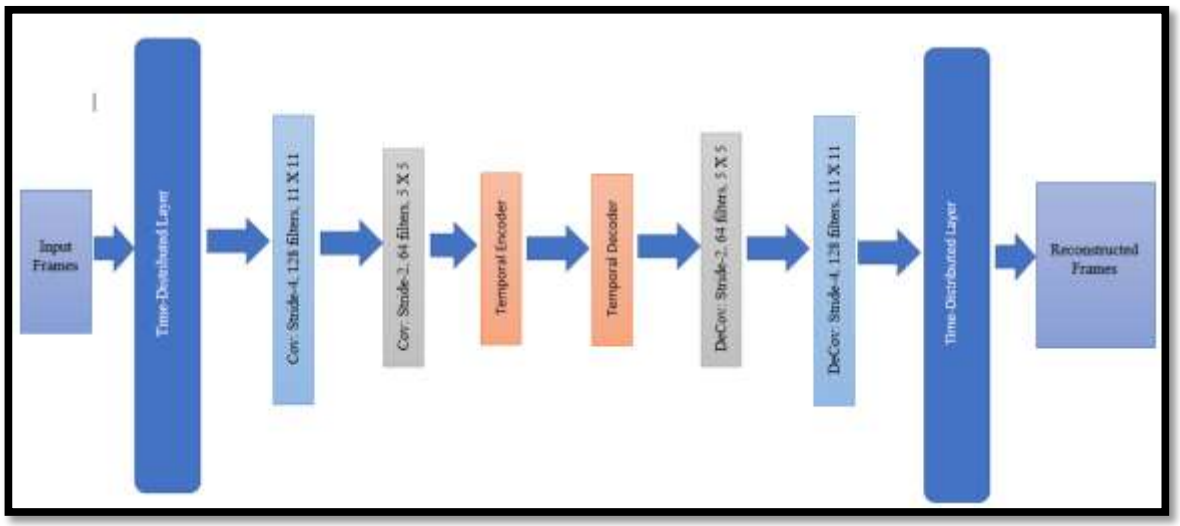

Fig. 2*. Proposed Architecture*

The three layers of the long short-term memory (LSTM) model are introduced in our model's proposed design, which is shown in Figure 2. These layers are used for better per-performance in video object detection and recognition; at this time, LSTM accepts a wide range of applications, including picture description, speech recognition, handwriting recognition, etc.

3.3 EXPLANATION OF AUTO-ENCODER

Encoding and decoding are the two stages of an auto-encoder. For each input, it uses to reduce the image dimension first. Second, after reconstructing the original data, the model is trained using back-propagation, which reduces error in an unsupervised manner. A nonlinear activation function like PCA is used by the auto-encoder to create features.

3.4 CONVOLUTION

Convolution is mostly used to extract useful characteristics from the input photos. The pixels' spatial in-interconnection is preserved. Convolution refers to the dot product between pixels and the specified kernel. If we employ a kernel (W) k × k and a p-input channel, the convolutional output dimension will be (p-k+1) × (p-k+1). At training, the convolutional algorithm uses a filter and provides additional parameters, such as the size and number of filters. We can extract better characteristics from the input frames if we increase the number of filters. But we need not bring up the high number of filters in relation to the performance problem.

3.5 RECURRENT NEURAL NETWORK (RNN)

Unlike a typical feed-forward neural network, it also takes into account the whole history of the input sequence that was given to the network. Compared to the conventional neural network, these enhance the network's performance. It can extract data from a lengthy input sequence in order to comprehend an unusual condition in the input video.

3.6 LONG SHORT-TERM MEMORY (LSTM)

We use the RNN variant known as long short-term memory LSTM to address this challenge. It employs a gate known as the forget gate. Its redesigned structure also stops the back-propagation fault. Additionally, it holds the higher-order information better when the input sequence is long. The general structure of an LSTM unit is shown in Figure 3.

One of the variations of the standard LSTM is called Convolutional Long Short-Term Memory (ConvLSTM). Convolutional processing is used in place of straightforward matrix multiplication to improve the spatial feature extracted from the input sequence. Additionally, due to its relationship between the input to hidden and hidden to obscure, ConvLSTM requires less weights.

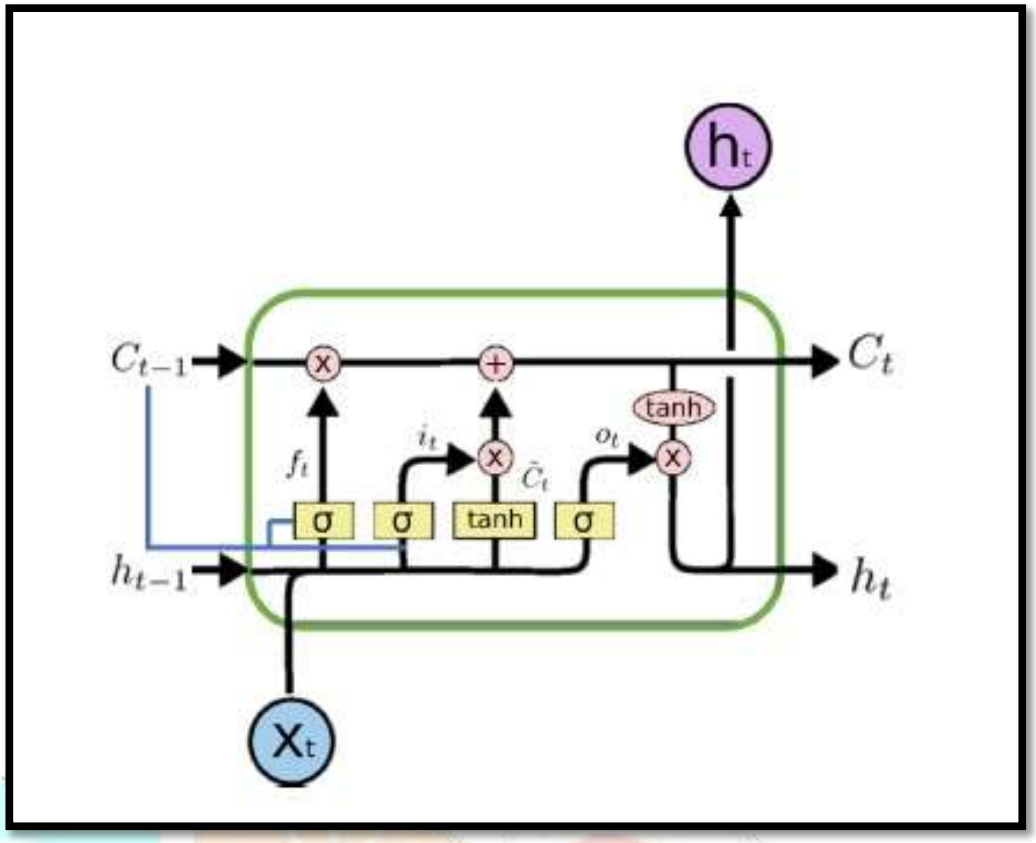

Fig. 3*. The LSTM Unit.* $C_t, h_t, x_t, tanh$ *are cell state, hidden state, input vector, and activation function respectively*

3.7 SCORE CALCULATION

After the model has been trained, we feed it testing data so that it can recognize anomalous occurrences while minimizing false alarms. To begin building the error, we measured the Euclidian distance between the input frame and the finished frame.

$$e(k) = \sqrt{p(k)^2 - f_w\big(p(k)\big)^2} \qquad (1)$$

where p(k) is the input frame pixel value, e(k) is an error calculation metric, and f_w (p(k)) is the reconstructed frame pixel value. We determine the abnormality score s_a (k), which is scaled from 0 to 1. Next, the regularity score s_r is determined (k).

$$s_a(k) = \frac{e(k) - e(k)_{min}}{e(k)_{max}} \qquad (2)$$

$$s_r(k) = 1 - s_a(k) \qquad (3)$$

IV. EXPERIMENTS

4.1 DATASET

The UCF-Crime dataset consists of 128 hours of movies and is quite extensive. 1900 uncut, lengthy real-world surveillance movies are included, together with 13 actual anomalies such as abuse, arson, assault, road accidents, burglaries, explosions, fights, robberies, shootings, shoplifting, and vandalism. The reason these anomalies were chosen is that they significantly affect public safety. Two tasks can be accomplished using this dataset. First, a general analysis of anomalies is performed, taking into account all anomalies in one group and all regular activities in another. Second, for identifying each of the 13 odd behaviors.

Fig. 4. *Frames from Training Set UCF Crime Dataset*

Fig. 5. *Frames from Test Set UCF Crime Dataset*

Pedestrians appear in normal events in the UCSD pedestrian dataset, but motorbikes, lorries, etc. appear in abnormal events. The industry standard for detecting anomalous events is this dataset. The UCSD dataset is split into the Ped1 and Ped2 subsets. Due to Ped1's low frame resolution of 158 × 240, which would have made using our framework for object proposal difficult, we decided to adopt Ped2. We determined that this was not a significant weakness in our architecture because low-cost, higher resolution cameras are already widely accessible on the market. Figures 6 and 7 show a few examples of datasets from the UCSD pedestrian dataset.

Fig. 6*. Frames from Train Set of UCSD Anomaly Dataset*

Fig. 7. *Frames from Test Set of UCSD Anomaly Dataset*

We used industry-standard metrics, such as the area under the curve (AUC), and equal error rate (EER), as defined in Li et al. [Kratz, & Nishino, (2009)], for both frame-level and pixel-level detection.

4.2 USED PARAMETERS

By lowering the input sequence's regularity error, we train our model. To set the learning parameter, we use the Adam optimizer to automatically update the weight history. We use mini-batches with a size of 128 and 50 epochs, or we run the batch until the reconstruction error decreases over time.

Here we use hyperbolic tangent for spatial auto-encoder. To ensure the regularity of the auto-encoder, we skip using the rectified linear unit because it has no upper bound.

4.3 ANOMALY DETECTION

Depending on the reconstruction error, we display the outcome. We determine the regularity score using this error. If our regularity score is low, the video input contains an aberrant event frame.

4.4 RESULTS ON UCF-CRIME AND UCSD DATASET

The AUC/EER on UCSD Ped2 datasets are summarized in Table 1, which shows that our system outperformed all other approaches on all benchmarks. Particularly, the AUC of Ped2 was 90.3%, outperforming the technique (66.5% in [Li, Mahadevan, & Vasconcelos, (2014)]) by a substantial margin. The AUC on UCF-Crime Dataset is summarized in Table 2.

Since our approach was based on a time-distributed auto-encoder and used Spatiotemporal characteristics to capture object-level spatial structure. Additionally, the UCF-Crime dataset included other activities that were absent from the auto-encoder training data. This demonstrated that the detection of untrained categories was generalized by spatiotemporal characteristics using a time-distributed auto-encoder. It should be noted that although other methods used motion features based on optical flows, our method performed better when using spatiotemporal data. Our future research will need to focus on learning auto-encoded C3D features.

Additionally, due to the poor resolution problem mentioned above, which should be fixed in the future, our performance on Ped1 is significantly inferior than state-of-the-art (70.8/34.8 in AUC/EER).

Method	***Ped2 (Frame)***	***Ped2 (Pixel)***
LMH [Adam, Rivlin, Shimshoni, & Reinitz, (2008)]	69.3 / 30	15.9 / 77.9
Social force [Mehran, Oyama, & Shah, (2009)]	55.6 / 42	21.7 / 72.4
MPPCA [Kim, & Grauman, (2009)]	69.3 / 30	22.2 / 77.6
MDT [Li, Mahadevan, & Vasconcelos, (2014)]	82.9 / 27.9	66.5 / 29.9
VideoParsing [Antic, & Ommer, (2011)]	92 / -	--
AMDN [Xu, Ricci, Yan, Song, & Sebe, (2015)]	90.8 / 17	--
Conv-AE [Hasan, Choi, Neumann, Roy-Chowdhury, & Davis, (2016)]	90.0 / 21.7	--
Spare 150fps	--	--
Our Method	92.4 / 12.8	90.3 / 14.8

Table 1: *Abnormal Event Detection Metrics AUC/EER (%) on UCSD Dataset*

Method	***UCF-Crime (AUC)***
Binary Classifier	50.0
Lu et al. [Lu, Shi, & Jia, (2013)]	65.51
Hasan et al. [Hasan, Choi, Neumann, Roy-Chowdhury, & Davis, (2016)]	50.6
Sultani et al. [Sultani, Chen, & Shah, (2018)]	75.41
Our Method	76.01

Table 2: *Abnormal Event Detection Metrics AUC (%) on UCF-Crime Dataset*

V. FUTURE SCOPE

In this study, we used a time-distributed auto-encoder along with a spatial-temporal feature extracted from the existing movie. The auto-encoded C3D feature will be used in the future to enhance our learning model while also fine-tuning the hyperparameters to extract better spatiotemporal characteristics from the video. Therefore, enhancing the training model yields an improvement over the prior model.

VI. Conclusion

In this experiment, we used a neural network to detect anomalies in videos. The spatiotemporal time-distributed auto-encoders system's spatial feature is used in this situation. ConvLSTM also attempts to improve the spatial quality of each frame from a video clip. A model that detects anomalous video occurrences from beginning to end is also available. This unsupervised method, which does not require labeling of the video sequences, can effectively discern between common and uncommon bizarre events in every video series. Nevertheless, depending on the noisier or more complicated environment of the frame, this method can produce false alarms. In this research, we attempted to create a supervised system for categorizing the data using the established methodology. We need enough video data for labeling in the study in order to accomplish this.

VII. CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

VIII. FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 555–560.

[2] Antic, B., & Ommer, B. (2011). Video parsing for abnormality detection. IEEE International Conference on Computer Vision (ICCV), pp. 2415-2422.

[3] Bobick, A., & Davis, J.W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 257–267.

[4] Brand, M., & Kettnaker, V. (2000). Discovery and segmentation of activities in video. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 844–851.

[5] Chong, Y.S., Tay, & Y.H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder, 25th IEEE International Conference on Image Processing (ICIP), pp. 189–196.

[6] Cong, Y., Yuan, J., & Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3449–3456.

[7] Dee, H., & Velastin, S. (2008). How close are we to solving the problem of automated visual surveillance ? : A review of real-world surveillance, scientific progress and evaluative mechanisms. Machine Vision and Applications, pp. 329–343.

[8] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K. & Davis, L. S. (2016). Learning temporal regularity in video sequences. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733-742.

[9] Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, pp. 334–352.

[10] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell., pp. 221–231.

[11] Johnson, N., & Hogg, D. (1995). Learning the distribution of object trajectories for event recognition. Image and Vision Computing.

[12] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. (2014). Large-scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

[13] Kim, J., & Grauman, K. (2009). Observe locally, infer globally a space-time mrf for detecting abnormal activities with incremental updates. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 915-930.

[14] Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1446–1453.

[15] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

[16] Li, C., Han, Z., Ye, Q., & Jiao, J. (2011). Abnormal behavior detection via sparse reconstruction analysis of trajectory. Proceedings - 6th International Conference on Image and Graphics, pp. 807–810.

[17] Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 18–32.

[18] Loy, C.C., Xiang, T., & Gong, S. (2008). From local temporal correlation to global anomaly detection. The 1st International Workshop on Machine Learning for Vision-based Motion Analysis - MLVMA'08, pp. 1–13.

[19] Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 fps in matlab. International Conference on Computer Vision (ICCV), pp. 2720–2727.

[20] Medel, J.R. (2016). Anomaly detection using predictive convolutional long short-term memory units. master's thesis, pp. 1–84

[21] Mehran, R., Oyama, A. & Shah, M. (2009). Abnormal crowd behaviour detection using social force model. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 591-600.

[22] Mo, X., Monga, V., Bala, R., & Fan, Z. (2014). Adaptive sparse representations for video anomaly detection, IEEE Trans. Circuits Syst. Video Techn., pp. 631–645.

[23] Moeslund, T.B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision based human motion capture and analysis. Computer Vision and Image Understanding, pp. 90–126.

[24] Mondal, S., Roy, A. & Mondal, S. (2023). Identification of Trajectory Anomalies on Video Surveillance Systems. International Journal of Digital Literacy and Digital Competence, pp. 1-18.

[25] Mondal, S., Roy, A., & Mandal, S. (2021). A supervised trajectory anomaly detection using velocity and path deviation. In: Proceedings of International Conference on Frontiers in Computing and Systems, pp. 777–784.

[26] Nguyen, N., Venkatesh, S., & Bui, H. (2006). Recognising behaviours of multiple people with hierarchical probabilistic model and statistical data association. Proceedings of the British Machine Vision Conference, pp. 1239–1248.

[27] Oneata, D., Verbeek, J., & Schmid, C. (2013). Action and event recognition with fisher vectors on a compact feature set. International Conference on Computer Vision (ICCV), pp. 1817–1824.

[28] Piciarelli, C., Micheloni, C., & Foresti, G.L. (2008). Trajectory-based anomalous event detection. Circuits and Systems for Video Technology. IEEE Transactions on Circuits and Systems for Video Technology, pp. 1544–1554.

[29] Reddy, V., Sanderson, C., & Lovell, B. (2011). Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 55–61.

[30] Sabokrou, M., Fathy, M., Mojtaba, H., & Klette, R. (2015). Real-time anomaly detection and localization in crowded scene. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 56–62.

[31] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 568–576.

[32] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for largescale image recognition. International Conference on Learning Representations, pp. 1–10.

[33] Sultani, W., Chen, C., & Shah, M. (2018). Real-world Anomaly Detection in Surveillance Videos. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9.

[34] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497.

[35] Vu, T.-H., Osokin, A., & Laptev, I. (2015). Context-aware cnns for person head detection. International Conference on Computer Vision (ICCV), pp. 2893–2901.

[36] Xiao, T., Zhang, C., Zha, H., & Wei, F. (2014). Anomaly detection via local coordinate factorization and spatio-temporal pyramid. Asian Conference on Computer Vision, pp. 66–82.

[37] Xu, D., Ricci, E., Yan, Y., Song, J., & Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. Computer Vision and Image Understanding, pp. 117-127.

[38] Zhao, B., Li, F.-F., & Xing, E. (2011). Online detection of unusual events in videos via dynamic sparse coding. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3313–3320.

[39] Zhou, S., Shen, W., Zeng, D., & Zhang, Z. (2015). Unusual event detection in crowded scenes by trajectory analysis. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 1300–1304.

[40] Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., & Zhang, Z. (2016). Spatialtemporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Processing: Image Communication, pp. 358–368.