Proceedings of 8th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2020 Month: June Volume: 8

Incremental Spatiotemporal Learner Model for Anomaly Detection and Localization for Video Surveillance

Sarita Sharma [a], Sanjeeb Prasad Panday [b], Sajjan Acharya [c]

[a, b] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU, Nepal*
Corresponding Email: [a] 074msice016.sarita@pcampus.edu.np, [b] sanjeeb@ioe.edu.np,
[c]074msmse012.sajjan@pcampus.edu.np

Abstract

This research presents an efficient method for anomaly detection in video. ISTL is an unsupervised deep learning approach that utilizes active learning with fuzzy aggregation, to continuously update and distinguish between new anomalies and normality that evolve over time. Hence, a spatiotemporal autoencoder architecture is unsupervised and used for anomaly detection in videos including crowded scenes. This architecture includes two main components, one spatial autoencoder for learning feature representation, and other temporal autoencoder for learning the temporal patterns of the spatial features. During training, the model is trained with only normal scenes, with the objective to minimize the reconstruction error between the input video frames and the output video frames reconstructed by the learned model. After the model is trained, normal video volume is expected to have low reconstruction error, whereas abnormal video volume is expected to have a high reconstruction error. By means of error threshold produced during each testing input volumes, the system will be able to detect when an abnormal event occurs. Experiments have done in three most common benchmark dataset Avenue, UCSD Ped1 and UCSD Ped2.

Keywords

Unsupervised Learning, Anomaly Detection, Anomaly Localization, Deep Learning, Active Learning, Spatiotemporal Analysis

1. Introduction

An anomaly is defined as something that deviates from what is standard, normal or expected. That means any activity that does not fit the learned model is considered as an anomaly. Anomaly detection is a complex task as the anomalies to be detected are not known prio, imposing difficulties even for a human observer[1].Anomaly detection is an unsupervised learning task where the goal is to identify abnormal patterns or motions in data that are by definition infrequent or rare events. Unsupervised anomaly detection system is able to detect abnormal behaviors without any prior knowledge of data labels [2]. Incremental Spatio-Temporal Learner (ISTL) is an anomaly detection approach for video surveillance that actively learns spatiotemporal patterns of normal behavior as it evolves over time. Inspired by the human brain, ISTL learns from immediately available information to distinguish between normal (safe) and anomalous (unsafe) behaviors, and continuously refines this understanding as the surroundings change and new information becomes available [3]. Active learning is mainly used for refinement and validation in ISTL, where a human observer contributes to the learning process for improved learning outcomes across iterations. The paradigm of active learning has been widely used in industrial image and video analysis applications such as character reading, facial recognition, autonomous vehicles and e-commerce.

2. Related Work

Anomaly detection techniques for abnormal event detection in videos broadly ranges across two areas of research i.e handcrafted features based and deep learning-based approaches.

2.1 Handcrafted Approaches

Techniques that utilize hand-crafted features, trajectories and spatiotemporal changes. Handcrafted approaches extract different types of motion information such as trajectories, optical flow and histogram gradient. The trajetory-based models generally learn the normal pattern of trajectories and describe the dynamic information of moving objects.Visual tracking algorithms are employed to obtain the motion of moving objects in the videos. Hu et al[4]. proposed a system that automatically learns motion patterns of moving objects and uses a fuzzy k-means based tracking algorithm. Xie et al.[5] proposed a motion instability based anomaly detection framework that discriminates anomalous behaviour based on the direction randomness and motion intensity. Wu et al. [6] proposed an approach in which objects are classified as anomalous based on how they follow the learned normal trajectory. Zhao et al. [7] utilize histogram of gradient (HoG) and histogram of optical flow (HoF) along spatial and temporal dimensions to encode an event and learn the normality upon dynamic sparse coding. As a result, trajectory-based approaches have good performance when the scenes are sparse because detecting and tracking moving objects are easy in sparse scenes. On the opposite, the success of the trajectory-based approaches have good performance when the scenes are sparse because detecting and tracking moving objects are easy in sparse scenes. On the opposite, the success of the trajectory-based methods degrades in crowded scenes due to the difficulty of detecting and tracking objects.

2.2 Deep Learning Approaches

With the advancements of deep learning, convolution neural networks (CNN), autoencoders and recurrent neural networks (RNN) have been utilized for video anomaly detection [8]. Convolutional neural network was initially used for anomaly detection to obtain both spatial and temporal features. However, CNN were not built to learn temporal features and are not a natural fit for videos. Convolutional Autoencoder is a good alternative to CNNs. Hasan et al. [9] proposed autoencoder, as an input he used state-of-the-art hand crafted motion features. Convolution and pooling operations are performed only spatially, eventhough the multiple frames as an input , because of the 2D convolutions, after the first convolution layer, temporal information is collapsed completely.On the other hand, long short term memory (LSTM) model is well-known for learning temporal patterns and predicting time series data. Convolutional LSTMs for learning the regular temporal patterns in videos. Luo.[10] and his findings show great promise of what deep neural network can learn. He attempted to detect anomalies by leveraging a CNN for appearance encoding and a convolutional long short-term memory (ConvLSTM) for remembering history of the motion information. Vu et al.[11] proposed an anomaly detection approach using a deep generative network.Chong et al.[2] proposed an abnormal event detection in videos using convolutional Autoencoder in which abnormal events are detected based on predefined data set and unaware of the continuous learning in videos survelliance domain.

3. Methodology

In this paper regularity score for anomaly detection and the concept for active learning is applied. At first ISTL is trained with a pre-identified normal behavior. If reconstruction error of the input cuboid is above the anomaly threshold, the input cuboid is classified as an anomaly. The classified frames are then sent to a human observer for verification. The objective of human observer feedback is to actively feed the learning model with dynamically evolving normality behavior. Therefore, if a detected video frame is an incorrect detection (false positive), then the human observer can mark the video frame as 'normal', which is used in the continuous learning phase. Subsequent to human observer feedback, the video frames that were marked as normal is used to continuously train the ISTL model, updating its knowledge of normality.

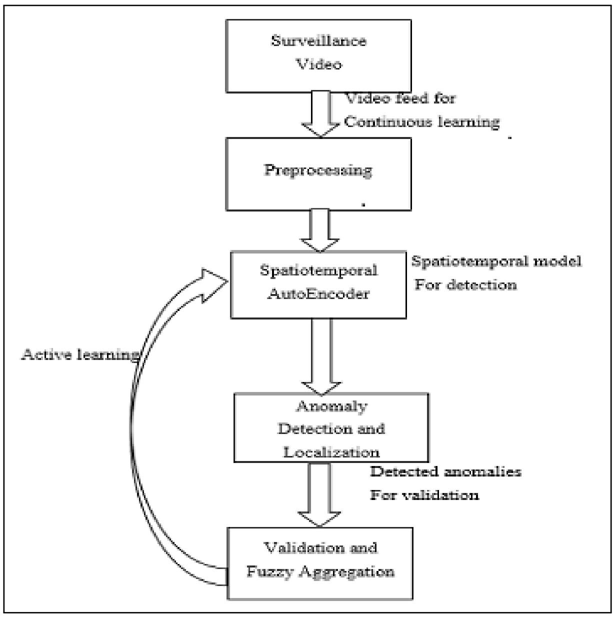
Surveillance Video
Video feed for Continuous learning
Preprocessing
Spatiotemporal AutoEncoder
Spatiotemporal model For detection
Active learning
Anomaly Detection and Localization
Detected anomalies For validation
Validation and Fuzzy Aggregation

Figure 1: Proposed System Block diagram

3.1 Preprocessing:

At first, the video data has been extracted as consecutive frames and then it has been converted into grayscale to reduce the dimensions after that resized to 256 x 256 pixels and to ensure that the input images are all on the same scale and then normalize pixel values by scaling between 0 and 1. The input to the spatiotemporal autoencoder model is a temporal cuboid of video frames, which is extracted using a sliding window of length T without any feature transformation. The consecutive frames of length T are stacked together to construct the input temporal cuboid. To generate more volumes data augmentation has been done. i.e concatenated frames with stride-1, stride-2, and stride-3. Now the input is ready for model training.

3.2 Feature Learning:

The ISTL model is composed of a convolutional spatiotemporal autoencoder to learn the regular patterns in the training videos. The model consists of a spatial decoder and an encoder with convolutional LSTM layers. Proposed architecture consists of two parts: spatial autoencoder for learning spatial structures of each video frame, and temporal encoder-decoder for learning temporal patterns of the encoded spatial structures. In the proposed architecture, the spatiotemporal autoencoder consists of a series of CNN layers to learn the spatial representation and a series of ConvLSTM layers to learn the temporal representation, which are described below.

3.2.1 Autoencoder:

Here, autoencoder is employed to learn regularity in video sequences. It is expected that the trained autoencoder will reconstruct regular video sequences with low error but will not accurately reconstruct motions in irregular video sequences. The encoder accepts as input a sequence of frames in chronological order, and it consists of two parts: the spatial encoder and the temporal encoder. The encoded features of the sequence that comes out of the spatial encoder are fed into the temporal encoder for motion encoding. The decoder mirrors the encoder to reconstruct the video sequence, so our autoencoder looks like a sandwich. In figure 2 below, T: Depth of temporal cuboid, F: Number of filters, K: Kernel size, S: strides, *: Multiplication.

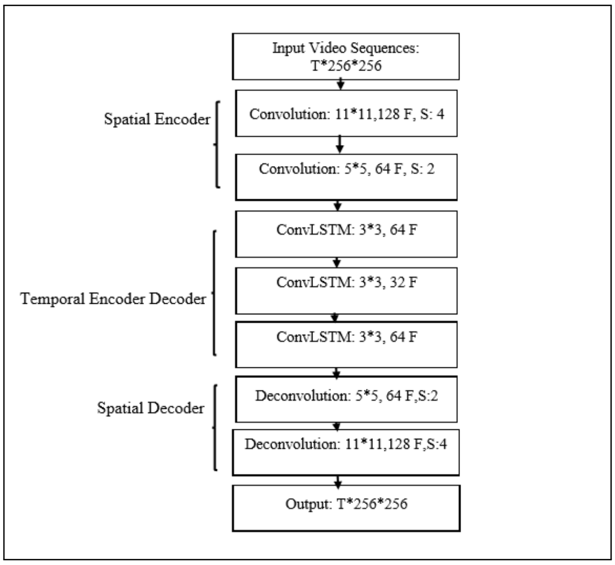
Input Video Sequences: T*256*256
Spatial Encoder
Convolution: 11*11,128 F, S: 4
Convolution: 5*5, 64 F, S: 2
Temporal Encoder Decoder
ConvLSTM: 3*3, 64 F
ConvLSTM: 3*3, 32 F
ConvLSTM: 3*3, 64 F
Spatial Decoder
Deconvolution: 5*5, 64 F,S:2
Deconvolution: 11*11,128 F,S:4
Output: T*256*256

Figure 2: Spatiotemporal Autoencoder Architecture

3.2.2 Spatial Convolution:

In the proposed model CNN layers is used to learn the spatial representations within the input frames by using filters, whose values are learned during the training process. Spatial convolution maintains the spatial correlation between image areas by using square filters.Let us assume that we are given nxn square and mxm filter, the convolutional layer output will be (n-m+1) x (n-m+1).

3.2.3 Convolutional LSTM Layers (ConvLSTM)

ConvLSTM is an extension of FC-LSTM. It has convolutional structures in both the input-to-state and state-to-state transitions. The ConvLSTM overcomes the drawback of LSTM by designing its inputs, hidden states, gates and cell outputs as 3D tensors, whose last dimension will be the spatial dimension. Further, the matrix operations in its inputs and gates are replaced with convolution operator. With these modifications, the ConvLSTM is able to capture the spatiotemporal features from the input frame sequences. The ConvLSTM model is represented in the equations (1) to (5) below. Hence, preserving the spatial relations.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci}^{o} C_{T-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf}^{o} C_{T-1} + b_f) \quad (2)$$

$$C_t = f_t^{o} C_{t-1} + i_t^{o} \tan h(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3)$$

$$O_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W^o_{c_o} C_{t-1} + b_o) \quad (4)$$

$$H_t = O^o_t \tan h(C_t) \quad (5)$$

In the euqations ,* and o represents convolution operation and hadamard product respectively. Inputs are represented by x_i, ...x_t, the cell states are represented by c_i , ... c_t, the hidden states are represented by h_i, h_t, and the gates i_t, f_t and o_t are all 3D tensors. σ is the sigmoid function and W_x and w_h are 2D convolution kernels in the ConvLSTM. Proposed ISTL model consists of three ConvLSTM layers.

3.3 Anomaly Detection and Localization

The proposed model is expected to learn regular spatiotemporal patterns from long training video, and during testing, detect those instances which donot fit into this model as anomalies. The reconstruction error represents the score for each temporal cuboid defining the anomaly. The reconstrucion error of each frame determines whether the frame is classified as anomalous. During the reconstruction cost analysis. If the reconstruction cost is greater than reconstruction threshold, the video sequence is considered as the abnormal whereas for reconstruction threshold less than threshold value is considered as the normal clip.

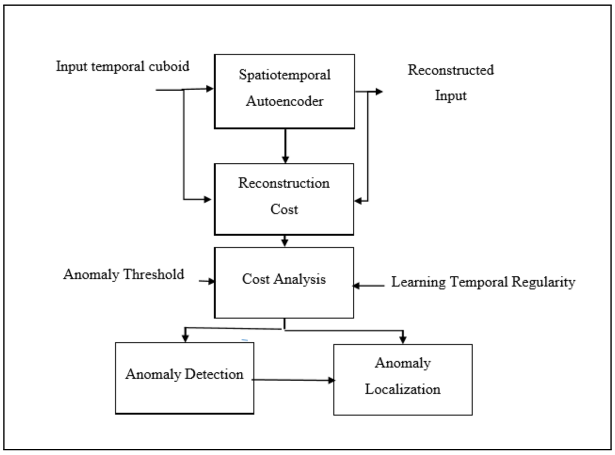

Figure 3: Anomaly detection and localization approach

The threshold is used to distinuish between normal behaviour and anomalies, named anomaly threshold. In practical Video survelliance applications, the human obseerver can select a value for μ based on the sensitivity required for the surveillance application. Figure 3 illustrates in more detail about anomaly detection and localization approach based on the reconstruction error. Anomaly loaclization locates the specific area of the video frame, where an anomaly has occured.

3.4 Validation and active learning with Fuzzy Aggregation

The purpose of the active learning in practical video surveillance context is to enable anomaly detection of dynamically evolving environments. The learning model is trained to identify accepted normal behavior provided at the beginning. When new normal behavior that has not been anticipated is considered abnormal, which is itself evolving normal behavior. The detection system evolves with capabilities for detecting such new scenarios.The continuous learning of the ISTL model is enriched by fuzzy aggregation of video frames, in order to retain stability across iterations of learning. At the detection phase, all the video frames being evaluated has been tagged with a fuzzy measure 'gλ' based on its reconstruction error and grouped into finite number (n) of sets based on 'gλ'. Subsequently, in the continuous learning phase, the algorithm selects the 'k' video frame cuboids that contain highest 'gλ ' from each set of fuzzy measures (S) to train the ISTL model. The parameters k and n will be defined at initiation based on the duration of video for continuous learning.The scene selection is defined by the equation4; ∀s ∈ S , where, S = s_1, s_2, ... s_n and d is the indexes of the selected temporal cuboids that will be included in the continuous training dataset. x_s

$$d = \Sigma_{i=0}^{n} \quad \underset{j=[1,K]}{max(S_i)} \quad (6)$$

As shown in figure 4, below the dataset for continuous training iteration includes

1. False positive detection verified by the human observer,
2. Temporal cuboids selected across normal behavior using the fuzzy aggregation.

This ensures the continuous training update the detection model's capability to capture novel normal behavior while remaining stable for previously known normal behavior.

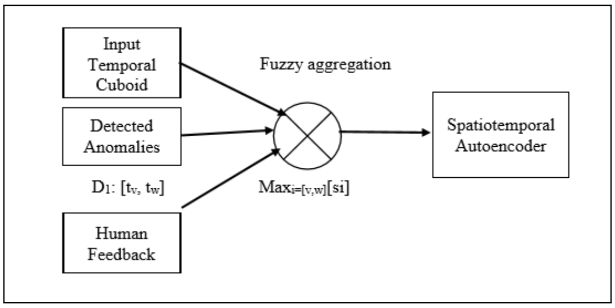

Figure 4: Active Learning of spatiotemporal model

3.5 Reconstruction Error and Regularity Score

The reconstruction error of all pixel values I in frame t of the video sequence is taken as the Euclidean distance between the input frame and the reconstructed frame:

$$e(t) = ||x(t) - fW(x(t))||^2 \tag{7}$$

Where, fW denotes the weights in the spatiotemporal model. Once the model is trained, performance of the model can be evaluate by feeding in testing data and check whether it is capable of detecting abnormal events while keeping false alarm rate low. Based on the reconstruction error, we then compute the abnormality score sa(t) by scaling the error values between 0 and 1.

$$s_a(t) = \frac{e(t) - e(t)_{min}}{e(t)_{max}} \tag{8}$$

Where, e(t) is the sequence reconstruction cost.Regularity score sr(t) can be simply derived by subtracting abnormality score from 1.

$$s_r(t) = 1 - s_a(t) \tag{9}$$

4. Experiment and Discussion

4.1 Datasets

The CUHK Avenue dataset [12] has 16 train video samples and 21 test video samples, with a resolution of 640 x 360 pixels. The normal behavior are pedestrians on the sidewalk and groups of pedestrians congregating on the sidewalk, whereas the anomalous events are people littering/discarding items, loitering, walking towards the camera, walking on the grass and abandoned objects. The UCSD pedestrian Dataset [13] has two datasets. UCSD Ped1 dataset has 34 training and 36 testing video clips, where each clip contains 200 frames, resolution of 238 x 158 pixels. Videos consist of groups of people walking towards and away from the camera. UCSD Ped2 dataset has 16 training and 12 testing video clips, where the number of frames of each clip varies and resolution of 360X240 pixels.The videos consist of walking pedestrians parallel to the camera plane. Anomalies of the two datasets include bikers, skaters, carts, wheelchairs and people walking in the grass area. The normal behaviors of the train video samples contain only scenarios of pedestrians walking on the pathway, whereas the test video samples contain anomalous pedestrian movement patterns such as walking across the sidewalk or walking on the grass, unexpected behavior such as skateboarding, cycling, and vehicular movement.

4.2 Experimental Setup

Here ISTL was implemented in Python with TensorFlow framework an optimizer Adam has been used with a learning rate set to 0.0001.Mean Squared Error (MSE) was taken as loss function during training the model. The training was performed with batch size of 4 with 100 number of epochs. At first step of experiment, preprocessing of the inputs has been performed. Here size of temporal sequences T=10 has been chosen using the sliding window technique. After the Spatiotemporal autoencoder model has been trained the model output goes for reconstruction cost calculation. Due to computational complexity while training the model layer normalization is used to reduce the training time by normalizing the activities of the neurons. Here anomaly threshold has been set to 0.1.

4.3 Qualitative Analysis: Visualizing Frame Regularity

By means of learned feature from spatiotemporal autoencoder model anomaly detection and localization has been performed by visualizing frames regularity.Each testing video has been tested individually. At first, test 001 of UCSD ped1 as shown in the regularity graph 5 has been performed. At the beginning of the curve video frames are highly regular but its regularity score drops from frame number 80 onwards till 153 frame. Right after the bicycle left regularity score starts to increase. By observing the graph and dataset, we can see a bicycle passing the walkway from frame 80 onwards.In this case model assume the bicycle as an anomaly.

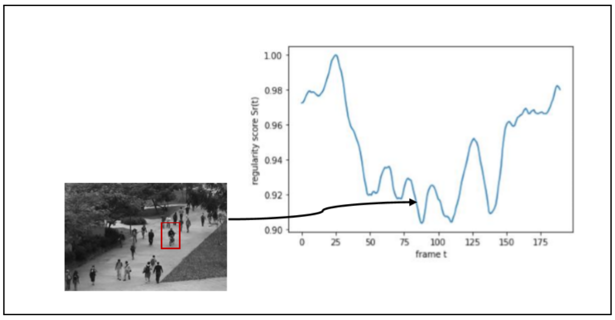

Figure 5: Regularity Score of video 001 from UCSD Ped1 dataset

During Test 004 of UCSDped1 dataset, as shown in figure 6 below. Which shows a skater entering the walkway at the beginning of the video, and someone walks on the grass at frame 140, which explains the two drops in the regularity score.

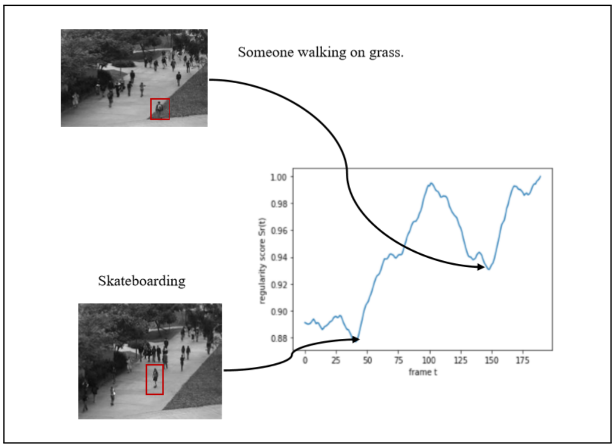

Figure 6: Regularity score of test 004 from UCSD Ped1

Now from the regularity score graph figure 7 of video 004 from UCSD Ped2.Where regularity score starts to decrease from high to low from frame 70 onwards. Where small vehicle and bicycle appears in the video and the model assumes these as anomaly.

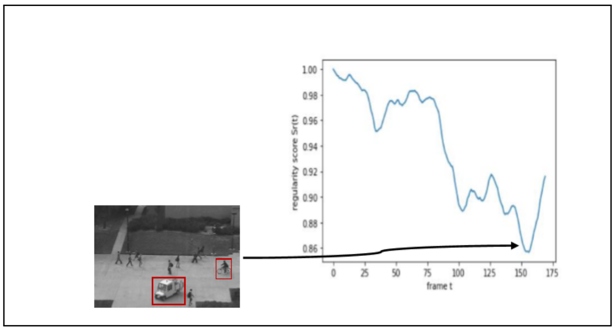

Figure 7: Regularity Score of video 004 from UCSD Ped2

Now figure 8 below shows the regularity curve of test 005 from avenue dataset. Where someone throwing a bag happens right after that regularity score drops from high to low and scores increases after he collects the bag and walk away from there.

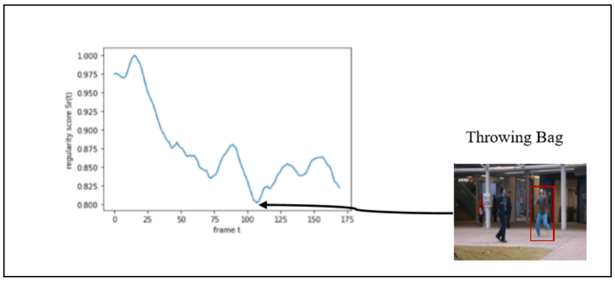

Figure 8: Regularity curve of test 005 of avenue dataset

4.4 Quantitative Analysis:

4.4.1 Anomalous Event Count by event type

Here we used the ground truth annotations for the performance evaluation of the model and we calculated the precision and recall comparing with ground truth for each dataset individually. The anomalous event count breakdown according to type of event is presented in Table 1-3 below.

Table 1: Anomalous event,false positives and total true positive count of UCSD ped1 dataset

	Biker	Skater	Cart	Trolly
Ground Truth	30	13	6	3
Ours	30	10	6	2
	Grass	Other	False Positive	True Positives
Ground Truth	4	4	0	60
Ours	3	2	10	53

In Table 1 grass means people walking on grass and other includes running and walking in a group. All bicycle and cart instances are well captured by the proposed system.These are strong abnormalities that are significantly different from what was captured in the normal scenes.These are the true positives detected by our model.However,it has difficulties in detecting certain types of event. A walking on grass, skater. Ten false positives are due to the pedestrian motion walking in an unusual direction, camera glitches and shake, camera angle whose elevation makes it difficult to differentiate between pedestrians and skateboarders by appearance. For this dataset our model gives the precision 0.841 and recall 0.88.

Table 2: Anomalous event,false positives and total true positives count of UCSD ped2 dataset

	Bicycle	Skater	Cart
Ground Truth	15	3	1
Ours	12	2	1
	False Positive	True Positive	
Ground Truth	0	19	
ours	3	15	

The main anomaly in the ped 2 test sample were cyclists. Most bicycle and cart instances are well captured by the proposed system. Missed cycles are located far away from the camera and due to occlusions in a crowded scene. Three false alarms are due to camera shake and occlusions of multiple pedestrians. For this dataset we obtained precision as 0.834 and recall as 0.789.

Table 3: Anomalous events, false positives and total true positives count in CHUK

	Run	Loiter	Throw
Ground Truth	12	8	19
Ours	11	7	19
	Opposite Direction	False Positive	True Positives
Ground Truth	8	0	47
ours	7	8	44

As in Table 3 above all throwing, loitering and irregular interaction events are well captured. These all are true positives.However, our system does have difficulties in detecting certain types of event. Missed detection of running events are due to the crowded activities where multiple and object of interest is far away from the camera. 5 out of 8 false alarms were due tocamera shake, whereas the rest of the false alarms are caused by obstruction to the camera, such as walking outside the shaded area of the station. Thus comparing with ground truth annotations, in this experiment we obtained precision 0.846 and recall 0.936.

4.4.2 Frame level AUC/EER

Quantitative evaluation in the form of frame level AUC and EER is shown in Table 4. Anomaly detection was evaluated with three state-of-the-art handcrafted feature representation-based approaches and five state-of-the-art deep learning-based approaches. For avenue dataset three state of the art hand crafted method did not publish their AUC/EER and our method exhibits the higher AUC of 81% than other five deep learning methods and EER of 27% which is comparable to other methods. Similarly for UCSD ped2 dataset our method give the better AUC of 92.5% than all other state of art methods i.e including hand crafted and deep learning methods and EER of 13% which is lower than all other methods except Chong which achieved the lower EER of 12%. Similarly for UCSD Ped1 dataset our method achieves the comparable AUC/EER (77%/28%) performance to other methods. Compared to the best result of the state-of-art approaches, proposed method shows an improvement of 2.5% in terms of frame-level AUC for UCSD Ped2 dataset and achieves the low EER than Hasan and Y.Zhao and exhibits the comparable performance to other datasets. Overall, the results prove that the proposed method is very effective on the Avenue and UCSD Ped2 dataset.

Table 4: Anomaly detection (AUC/EER) %. Higher AUC and lower EER indicates better performance

methods	ped1	ped2	Avenue
Adam [14]	77.1/38.0	NA/42.0	NA
Mehran [15]	67.5/31.0	55.6/42.0	NA
Mahadevan [12]	74.2/32.0	61.3/36.0	NA
Hasan [9]	81.0/27.9	90.0/21.7	70.2/25.1
Chong [2]	89.9/12.5	87.4/12.0	80.3/20.7
Luo [10]	75.5/NA	88.1/NA	77.0/NA
Y.Zhao [16]	87.1/18.3	88.6/20.9	80.9/24.4
H.Vu [11]	70.25/35.4	86.43/16.47	78.76/27.21
Ours	77/28	92.5/13	81/27

4.5 Active Learning

To demonstrate the active learning capability of this model, cycling scenarios on UCSD ped2 has been selected. Here we defined only bicycle movement on pathway as normal.We employed to test samples with human observer feedback that contains cycling for continuous learning. After training we evaluated the model excluding these two samples that were selected for continuous learning.The anomaly detection ratio is performed as detected anomalies/total test sample, ratio is 10/11 before active learning but it reduces to 3/11 after active learning. That means test video 4, 7 and 8 contains cycling and truck, cycling and skater on the video so the model assumes this as anomaly in active learning case.For easy we divided the total test samples in two categories where A)cycle only and B)cycle and vehicle or skater moving together in pedestrian pathway. This resulted in test video that A is detected as normal while B is detected as anomaly.

5. Conclusion

Here in this research we have proposed a new incremental spatiotemporal autoencoder model for anomaly detection in video surveillance system. Proposed end to end trainable model incorporates the spatial encoder decoder for spatial feature representation and temporal encoder decoder for temporal motion patterns. Qualitative analysis and quantitative comparison based on three benchmark dataset have been performed. As future work, we intend to use ISTL model for active learning of various datasets and also to implement for online video streaming cases.

References

[1] V. Chandola and V. Kumar A. Banerjee. Anomaly detection: A survey. *ACM Comput. Surv., vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009*, 2009.

[2] Y. S. Chong and Y. H. Tay. Abnormal event detection in videos using spatiotemporal autoencoder. *Advances in Neural Networks - ISNN 2017, 2017, pp. 189–196*, 2017.

[3] D. De Silva and D. Alahakoon. Incremental knowledge acquisition and self-learning from text. *he 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–8*, 2010.

[4] W. Hu, X. Xiao, Z. Fu, D Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 9, pp. 1450–1464, Sep. 2006.*, 2006.

[5] S. Xie and Y. Guan. Motion instability based unsupervised online abnormal behaviors detection. *Multimed Tools Appl, vol. 75, no. 12, pp. 7423–7444, Jun. 2016*, 2016.

[6] S. Wu and M. Shah B. E. Moore. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2054–2060*, 2010.

[7] B. Zhao and E. P. Xing L. Fei-Fei. Online detection of unusual events in videos via dynamic sparse coding. *CVPR 2011, 2011, pp. 3313–3320*, 2011.

[8] R. Nawaratne, T. Bandaragoda, A. Adikari, D. Alahakoon, D. De Silva, and X. Yu. Incremental knowledge acquisition and self-learning for autonomous video surveillance. *IECON 2017,2017, pp. 4790–4795.*, 2017.

[9] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. *IEEE Conference on Computer Vision and Pattern*, 2016.

[10] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. *2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 439–444.*, 2017.

[11] H. Vu. Deep abnormality detection in video data. *Electronic proceedings of IJCAI 2017, pp. 5217–5218*, 2017.

[12] V. Mahadevan and N. Vasconcelos W. Li, V. Bhalodia. Anomaly detection in crowded scenes. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[13] C. Lu and J. Jia J. Shi. Abnormal event detection at 150 fps in matlab. *2013 IEEE International Conference on Computer Vision, pp. 2720–2727*, 2013.

[14] A.Adam and D.Reinitz E.Rivlin, I.Shimshoni. Robust real-time unusual event detectin using multiple fixed-location monitors. *TPAMI,30(3):555-560*, 2008.

[15] R. Mehran and M. Shah A. Oyama. Abnormal crowd behavior detection using social force model. *IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942*, 2009.

[16] Yiru Zhao, Bing Deng, Chen Shen andYao liu, Hongtao lu, and Xian Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. *MM'17, October 23-27, 2017, Mountain View, CA, USA*, 2017.