IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANOMALY DETECTION FOR ABNORMAL ACTIVITY WITH VIDEO SURVEILLANCE USING DEEP LEARNING

[1]Shruti Trivedi, [2]Mayuresh Kulkarni

[1]M.E. Student, [2]Assistant Professor

[1]Department of Computer Engineering,

[1]Silver Oak College of Engineering and Technology, Ahmedabad, India

Abstract: Detection of abnormal activities is essential but demanding research problem in computer vision. Anomaly detection in video surveillance involves monitoring of the designated area of interest to detect anomalies. Video surveillance is extensively used for a variety of fields, for instance, traffic monitoring, medical monitoring, security guarding, etc. Amid these various research fields, the detection of anomalous events plays a significant role. The goal is to identify abnormal behavior through the use of reliable features by utilizing deep learning methods. To address that, a fusion approach is proposed in videos based on three-dimensional convolutional neural networks (C3D) for extracting spatiotemporal features and convolutional long-short-term-memory (ConvLSTM) to aggregate the features that enhances the accuracy and efficiency of the model. The popular and open source UCSD dataset consisting of two sub datasets UCSD Ped1 and UCSD Ped2 has been used to evaluate the performance of the proposed method. The result demonstrates that we get high accuracy as compared to other approaches which is 94.3% and 92.6% for the first and second datasets respectively.

Index Terms **- Anomaly Detection, Video Surveillance, 3D convolutional neural network, model fusion, Convolutional long-short-term-memory.**

I. INTRODUCTION

Anomaly Detection is a process for the identification of events, unusual patterns, or observation in Datasets that give rise to suspicions as they differ from normal. Anomalies shall be referred to as Exceptions, outsiders, noise, and deviations as well.

Detection of human behavior is always seen as a challenge to safety and security in crowded locations. Throughout all activities, anomaly detection is an event that has been monitored continuously and frequently over the years. It is considered to be very important for a number of security reasons. Surveillance cameras are often used in public places such as banks, shopping malls, convenience stores, streets to increase people's safety. However, the task of detecting anomalous events, such as crimes, accidents or illegal activities, is critical. Any real-world irregularities are considered to be complicated and show a great deal of variety as the camera-captured environment can change significantly over time. Defining a region as normal is difficult, as the boundaries between regular and anomalous data are not accurate. As a result, normal behavior can be detected as anomalies and abnormalities can be identified as normal observations. Sometimes the definition of normal or abnormal may often change, and the behavior that is considered normal today may be abnormal in the future. In general, the task of detecting anomaly is not easy to solve due to these challenges. Many of the techniques used to detect anomalies solve a specific formulation of the problem, which is followed by different factors, such as the availability of labeled data, the nature of the data, etc.

Over the years, several traditional approaches have been introduced to detect abnormal behavior. Many of them use handcrafted features, but the task depends on human assumptions. These approaches are based on the assumption that any behavior or pattern that is different from the normal scenario learned would be considered an abnormal activity that might not be true due to the notion of anomaly that varies from domain to domain. Many solutions currently available seem promising, but there are drawbacks that affect the efficiency of the model and prevent the system from achieving higher accuracy.

This research paper proposes a fusion model to detect anomalies in video surveillance. There are measures to identify the anomaly in the model. At first, the data is preprocessed to convert them into valid input form. Instead, to extract the information, we use a three-dimensional convolutional neural network as a feature extractor. The benefit of using a deep learning technique is that the three-dimensional filters used in the task provide both spatial and temporal information, whereas the two-dimensional filters used in handcrafted techniques provide only spatial information. In order to enhance model accuracy and performance, the features are aggregated using a convolutional LSTM. Finally, by implementing and evaluating the model, we can obtain improved outcomes relative to other state-of-the-art approaches currently in use.

The structure of this research paper is described as follows. Section II represents some important related work done for abnormality detection task. Section III represents the architecture and the methodology of the proposed fusion model with the steps implemented. Section IV represents the evaluation of the model performance and comparison with other methods. Ultimately, we sum up the conclusions and suggestions for future investigation in Section V.

II. RELATED WORK

Most of such rare events are uncertain beforehand, because that would require anticipating all the ways that anything could happen outside the standard. Hence, learning a model which is abnormal or uncommon is almost impossible. The task of getting a normal video data in which the scene is more regular in comparison of what is anomalous; we should concentrate on the setting where the training clips or data includes only regular patterns. A common approach used by researchers in this field is to learn the usual patterns from the training videos first, and then the irregularities are observed as events deviated from usual patterns. Most anomaly detection work relies on extracting local features from videos, which are then used to train a model. In video analysis and detection of anomalies, trajectories have been used extensively. Approaches like trajectory based or pixel based are used for the task of detection. In trajectory-based approaches, irregularity such as directions or speed can be identified, but there are some events or action like body movements which might not be identified through observation of spatiotemporal trajectory points. In the same way, pixel-based methods might not identify the activity such as loitering of a person or any criminal activity as these activities are related to global movement instead of body movement. The researcher of the paper [1] proposed a fused approach of trajectory-based and pixel-based in a unified framework combining object trajectory analysis output with the pixel-based analysis to identify irregular behavior related to object trajectories direction and speed, and also complex behaviors related to finer movement of the object. An approach that focuses on spatiotemporal abnormalities in the video data exists as well. These depend mainly upon extraction and analysis of local level features, such as histogram of oriented gradients, histogram of oriented flows as well as optical flow [2] that focuses on focuses on the optical flow information. Their success is due to their low computational cost and their ability to concentrate on abnormal behavior, even in highly crowded scenes. Another technique is sparse reconstruction. The researchers of [3] propose a methodology which focuses on sparse components. Given the input it is calculated in matrix representing entire video. Matrix decomposition is used to acquire sparse components. The purpose is to capture the usual activities by the low-ranking variable and the sparsely occurring unusual activities by sparse components. Further filter is utilized to refine the obtained sparse components that help to decrease deviations and noise. These are all traditional methods that mostly focus on local features, handcrafted feature within limited regions. Consequently, the effectiveness of deep learning approaches through various applications has caused the growth of methods in detections of abnormalities. A concept Deep Learning refers to the learning of hierarchical set features in an artificial neural network (ANN) from multiple layers of hidden nodes. Unlike conventional approaches, using a limited amount of pre-processing, deep learning techniques learn important features directly from the data. Specifically, CNN has demonstrated effectiveness in a wide variety of applications such as person detection, object and behavior recognition, and much more. It consists of a stack of convolutional layers with a fully-connected layer and a softmax function or classifier. Convolutional Autoencoder on the other hand is eventually a ConvNet with mirrored stack of convolutional layer replacing classifier. Researchers of [4] proposed a convolutional spatiotemporal autoencoder for learning usual patterns in the training video data clips. The framework includes two parts. First, spatial autoencoder to learn the spatial structure of single frames of video where spatial encoder and decoder has two convolutional as well as deconvolutional layers. Second, temporal encoder-decoder to learn temporal patterns of encoder spatial structure where temporal encoder is three-layered convolutional LSTM. The LSTM model is widely used for sequence learning and time series modeling, and has demonstrated its success in applications such as speech translation and recognition of handwriting. Once the model is trained, videos with normal events are expected to have small number of reconstruction error, and video volumes with unusual activities are expected to have high reconstruction error. The system applies Thresholding on the error generated by testing input volumes. The authors of [5] introduced a hybrid model autoencoder model. It is based on two parts where one is to use convolutional autoencoder(ConvAE) for the learning of each frame's spatial structure; other part employs the improvised LSTM encoder-decoder with Convolutional-LSTM to learn temporal video motion feature for learnt spatial feature map of each and every time step. These methods have been introduced over the time with fusion of variety of architectures. Some of them are traditional and some are deep learning techniques. The idea is to extract spatiotemporal features in more enhanced way with higher accuracy and efficiency of the model.

III. OUR APPROACH

The phase of anomaly detection with our proposed approach starts with the task of preprocessing data followed by learning features, including the extraction and aggregation of information. Finally, the anomaly score is checked to evaluate the efficiency of the model. The detailed proposed flow is shown in Figure 1.

3.1 Preprocessing

The job at this point is to transform the original raw input data into the form that is suitable to our proposed model. All frames are taken from the video and resized to 112 x 112. The pixel values are scaled from 0 to 1 to ensure that the size of all input images is the same and, for normalization, each frame is subtracted from the global mean image. The mean image is computed by averaging the pixel values of each frame at each position in the training dataset. After that, the model is considered ready for the model training task. Here, the frames of the video clips are divided into chunks of 8 or 16, in our case 16 frames and provided as input to the function extractor.

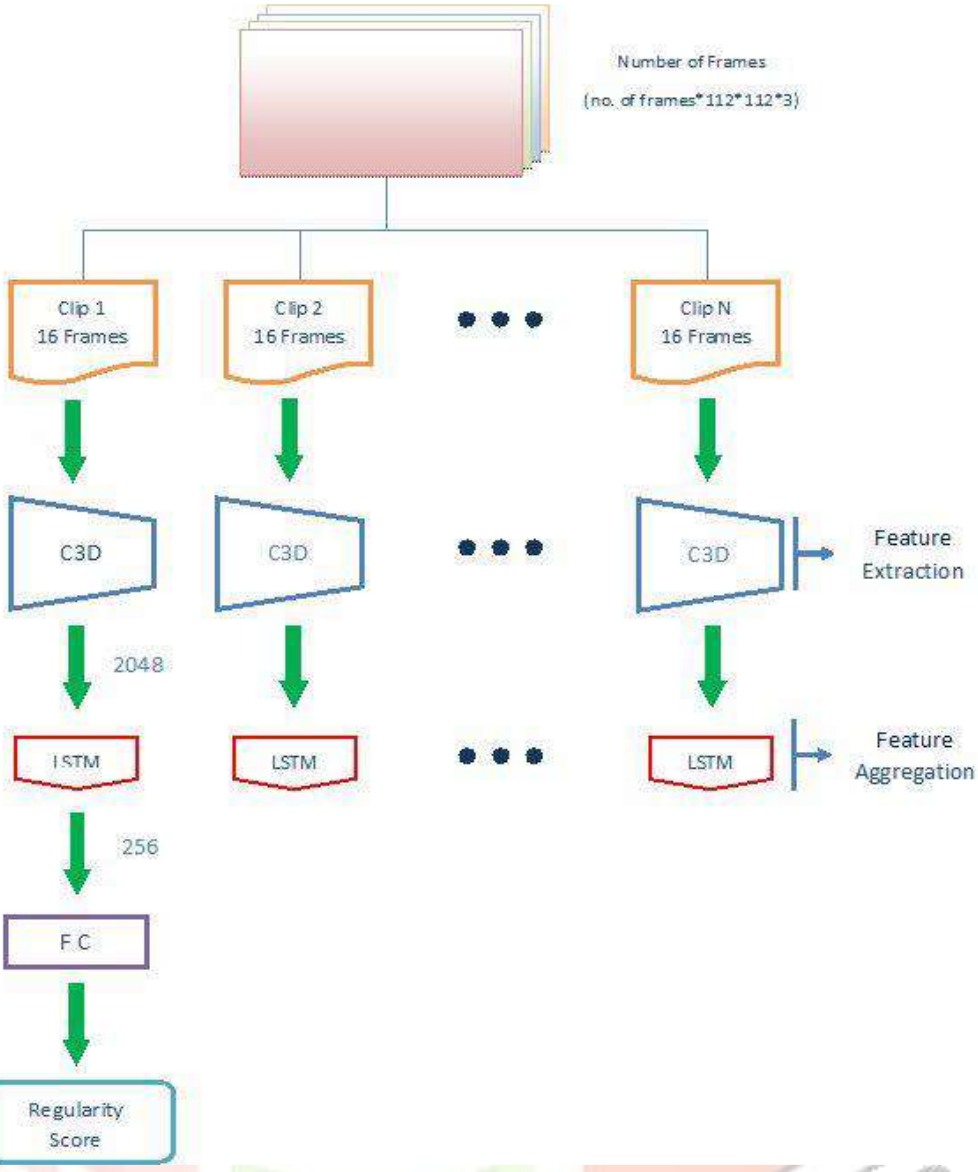

Figure 1 Proposed Model Flow

3.2 Feature Extraction

In this stage, the features are extracted from the preprocessed data. The feature learning stage may use a variety of architectures or approaches as a feature extractor.

A 2D convolutional neural network is a type of model that automates the task of extracting features by acting directly on the original raw data. This model is ideally suited for storing input spatial information. But when it comes to time information for the same input image, it collapses. Therefore, we have chosen a 3D convolutional network as a feature extractor that preserves the spatial and temporal information of the input images. It has been observed in the research that 3D CNN is very suitable for a spatiotemporal feature extractor as compared with 2D CNN because of its ability to model temporal information and model motion and appearance at the same time. When the feature extraction process is completed, the features are sent to the next stage of the feature aggregation.

3.3 Feature Aggregation

This stage is considered to be part of the process of feature learning. At this point, the Convolutional LSTM is used to aggregate the features extracted by C3D [6]. Simple LSTM ensures that backpropogated error does not burst or vanish, and can function on sequences that are long and can be stacked with each other to obtain high-level information.

Convolutional LSTM is a variant of LSTM that requires less weight and provides better spatial features. It has the capability to temporarily disseminate spatial features through convolutional LSTM state; it also has an extra peephole connection that helps it to obtain better knowledge of the past. This additional relation is an advantage in the process of detecting anomaly as it can look up to past knowledge with long sequences and recall the data collected from the training to differentiate abnormality from normal.

3.4 Regularity Score

After the model training is done, to evaluate performance of our model, the testing data is fed and checked if it is capable to detect abnormality among normal behavior while retaining false alarm rate to low. The equation to compute regularity score for every frame is same as [4] to better compare. Here, the I is reconstruction error of pixel values within frame t of a video sequence is chosen as Euclidean distance among frames which is input and reconstructed one:

$$e(t) = ||x(t) \square fw(x(t))||2 \quad (3.1)$$

Where fw = learned weights from model. Then the abnormality score sa(t) is calculated by scaling within 0 and 1. There after regularity score which sr(t) can be calculated by deducting anomaly score from the equation (3.1).

$$sa(t)=(e(t) - e(t)min)/(e(t)max) \quad (3.2)$$

$$sr(t) = 1 - sa(t) \quad (3.3)$$

IV. EXPERIMENT

The performance of our model is evaluated with state-of-art methods. We have used Keras, Tensorflow and many other libraries to implement proposed model on the dataset and compare with the previously used algorithms or techniques.

4.1 Dataset Description

In the proposed method, For the purpose of identifying abnormalities, there are several datasets available publicly, such as UCSD dataset [9], Avenue dataset [10], UMN dataset [11], and Subway dataset [12]. Amid all, we have selected UCSD dataset for the experiments of accuracy and efficiency.. UCSD datasets is divided into two individual datasets which are UCSD Ped1 and UCSD Ped2 respectively.

- UCSD Ped1 consists of 34 training and 36 testing video clips containing 200 frames in video sequences with resolution of 238 x 158.
- UCSD Ped2 consists of 16 training and 12 testing video samples containing variety number of frames which ranges between 120, 150 and 180 in training video clips whereas 180 frame for the testing video clips with the resolution of 360 x 240.

Ped1 test clips contain and detect anomalies such as skaters, cycling, wheelchairs, carts etc in the scene. People's movements on the pedestrian changes after a while from rarely crowded to heavily crowded. Ped2 test clips detects the anomalies like skaters, people riding bicycles, trucks in the middle of pedestrian, etc.

4.2 Evaluation

The efficiency of our model is evaluated using state-of-the-art methods. Reconstruction error decides how the frame is classified as abnormal. Thresholding is a rather straight-line approach to decide how reactive we want our system to be. Further warnings are activated when the threshold is set to low as the system becomes more sensitive to its surroundings. In the table, frame lever comparison is shown of area under curve (AUC) and equal error rate (EER) of previously introduced methods with ours. AUC is considered to be the higher the better and EER is considered to be the lower the better.

Table 1 Overall ratio of AUC and EER of different methods

Method	AUC/EER(%)	
	UCSD Ped1	UCSD Ped2
HOFME[7]	72.7/33.1	87.5/20.0
ConvAE[8]	81.0/27.9	90.0/21.7
ST-AE[4]	89.9/12.5	87.4/12.0
HST-AE[5]	90.5/13.5	88.9/11.5
Ours	94.3/10.7	92.6/12.9

Figure 2 illustrates the output of proposed model on the video samples of UCSD Ped1; the method we have proposed detects abnormality properly even in slightly crowded scenes.

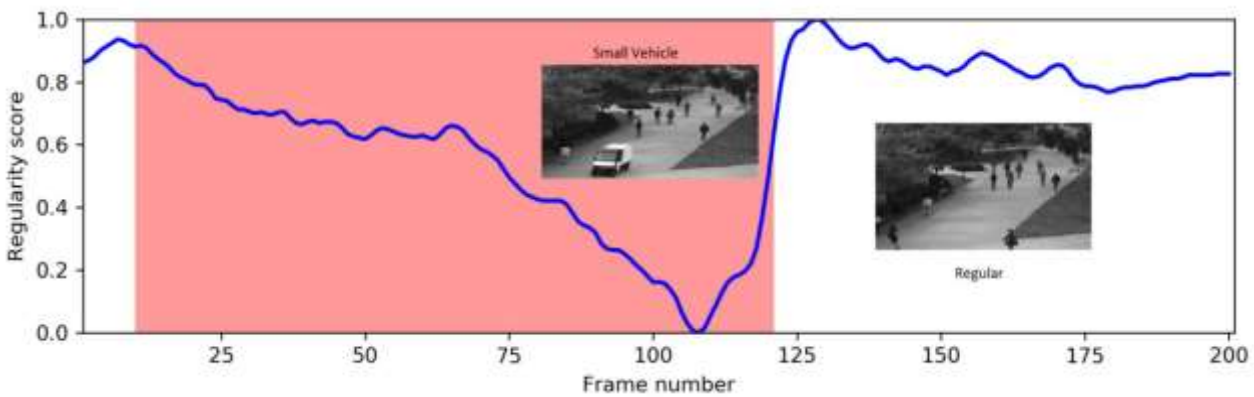

Figure 2 Regularity Score of video #27 from UCSD Ped1

In the figure, the abnormality is identified as the small vehicle on the walkway in this video. When it is closely seen in the video, it is observed, and then the graph goes regular when the situation is normal.

Figure 3 and 4 compares the output of proposed model with standalone 3D CNN and appearance.

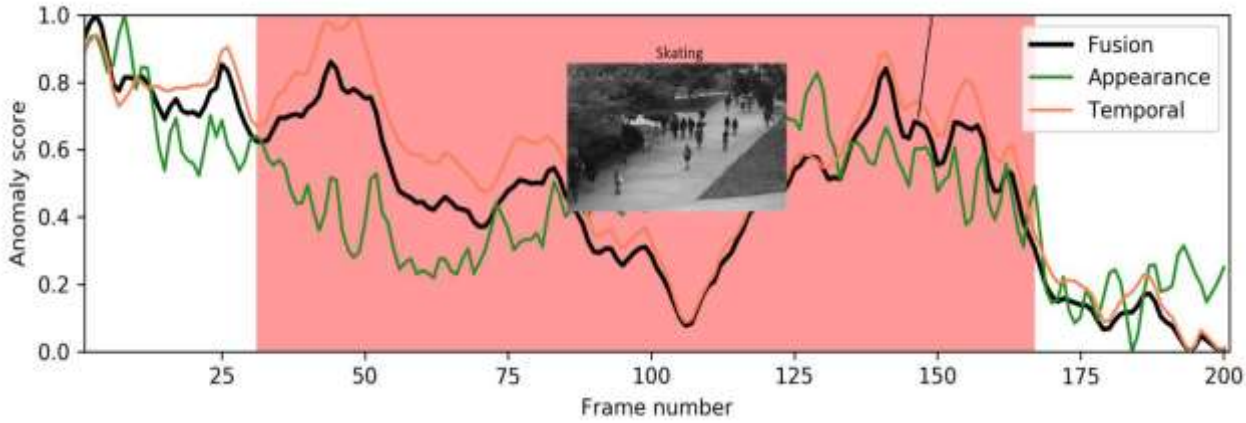

Figure 3 Comparison with method of video #4 of UCSD Ped1

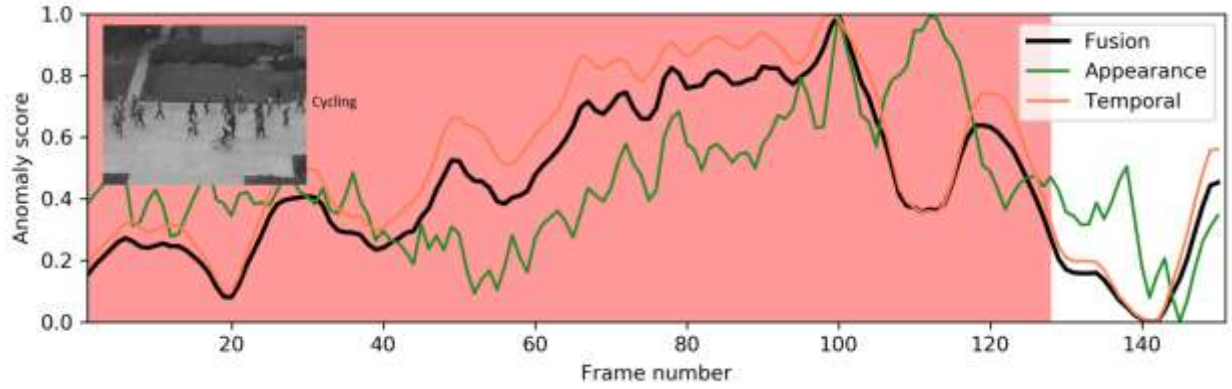

Figure 4 Comparison with method of video #5 of UCSD Ped2

V. CONCLUSION

Detection of suspicious behavior in video surveillance has become a very demanding field of study in computer vision. Among the strategies available for the task, the critical task is to reliably identify irregular activity in the test data and to classify them into individual groups according to their behavior. Past traditional approaches operate on local and handcrafted features. In the identification of anomalies, the efficiency of conventional techniques on sequence data and images is not adequate due to their inability to capture complex data structures. In order to solve a problem such as lack of training data and difficulties in obtaining spatial and temporal information, we propose a fusion model that combines C3D and ConvLSTM. The primary step of the model is to preprocess the data to convert them into more acceptable input form for the model. Feature extraction is carried out on the input from preprocessing through the use of three-dimensional convolutional neural network. To make a model more efficient, we aggregate the features using convolutional long-short-term-memory. It enhances the performance of model by increasing accuracy. Previous models generate accuracy less than 90% for both dataset pedestrians. Our model's accuracy is above 92%.

There are several problems with deep anomaly detection techniques, the shortage of training data or lack of availability is one of them. Model cannot be able to identify the anomaly more precisely if the size of the data increases or the scene becomes more crowded. The boundary between normal and abnormal behavior might not be very accurate and the definition can differ depending on the situation and passing time. These are some of the issues of present approach that gives a direction for future work in the field.

REFERENCES

[1] Coşar, S., Donatiello, G., Bogorny, V., Garate, C., Alvares, L. O., & Brémond, F. (2016). Toward abnormal trajectory and event detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology, 27(3), 683-695.

[2] Colque, R. V. H. M., Caetano, C., de Andrade, M. T. L., & Schwartz, W. R. (2016). Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Transactions on Circuits and Systems for Video Technology, 27(3), 673-682.

[3] Biswas, S., & Gupta, V. (2017). Abnormality detection in crowd videos by tracking sparse components. Machine Vision and Applications, 28(1-2), 35-48.

[4] Chong, Y. S., & Tay, Y. H. (2017, June). Abnormal event detection in videos using spatiotemporal autoencoder. In International Symposium on Neural Networks (pp. 189-196). Springer, Cham.

[5] Wang, L., Zhou, F., Li, Z., Zuo, W., & Tan, H. (2018, October). Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 2276-2280). IEEE.

[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[7] Wang, T., & Snoussi, H. (2012, September). Histograms of optical flow orientation for visual abnormal events detection. In 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (pp. 13-18). IEEE.

[8] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 733-742).

[9] Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010, June). Anomaly detection in crowded scenes. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 1975-1981). IEEE.

[10] Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE international conference on computer vision (pp. 2720-2727).

[11] "Abnormal Crowd Behavior Detection using Social Force Model", accessed on 11th of June,2020, https://www.crcv.ucf.edu/projects/Abnormal_Crowd/

[12] 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, 30(3), 555-560.