International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Anomaly Detection through Video Surveillance using Machine Learning

Sarika Late, Mrunmay Pathe, Riya Makhija, Jagruti Tatiya, Prof. Mrunal Pathak
Department of Information Technology, AISSMS Institute of Information Technology, Maharashtra, India

Article Info
Volume 8, Issue 4
Page Number : 137-145

Publication Issue
July-August-2021

Article History
Accepted : 02 July 2021
Published : 08 July 2021

ABSTRACT

Anomaly Detection is identification of suspicious human behavior using real-time CCTV video. Human Anomaly Behavior has been studied as one of the main problems of computer vision for more than 15 years. It is important because of the sheer number of applications that can benefit from activity detection. For applications such as image monitoring, object tracking and formed to oversee, sign language identification, advanced human contact, and less motion capture markers, for example, human pose estimates are used. Low-cost depth sensors have disadvantages, such as restricted indoor use, and with low resolution and noisy depth information, it is difficult to estimate human poses from deep images.

The proposed system therefore plans to use neural networks to solve these problems. Suspicious identification of human activity through video surveillance is an active research area in the field of image recognition and computer vision. Human activities can be monitored by video surveillance in critical and public places, such as bus stations, train stations, airports, banks, shopping malls, schools and colleges, parking lots, highways, etc. to detect terrorism, robbery, chain snatching crimes, and other suspicious activities. It is very difficult to monitor public places continuously, so it is important to have intelligent video surveillance that can track human behavior in real time and categorize it as common and unusual, and that can generate an alarm. The experimental results show that the proposed algorithm could reliably detect the unusual events in the video.

Keywords : Video Surveillance, Anomaly detection, Image Processing, CNN, Machine Learning.

I. INTRODUCTION

With the increasing demand for security, surveillance cameras have been widely deployed as the infrastructure for video analysis [1]. Surveillance cameras are being used in public places e.g., streets, intersections, banks, shopping malls, etc. to increase public safety and it only store records of CCTV footages. One major challenge faced by surveillance video analysis [1] is detecting abnormal, which

requires exhausting human efforts. However, the monitoring capability of law enforcement agencies has not kept pace [2]. The result is that there is a glaring deficiency in the utilization of surveillance cameras and an unworkable ratio of cameras to human monitors [3]. One critical task in video surveillance is detecting anomalous events such as traffic accidents, crimes, or illegal activities. Anomalies in videos are broadly defined as events that are unusual and signify irregular behavior [3]. Consequently,anomaly detection has broad applications in many different areas, including surveillance, intrusion detection,health monitoring, and event detection [4]. Unusual events of interest in long video sequences, e.g. surveillance footage,often have an extremely low probability of occurring [5]. Assuch, manually detecting these rare events, or anomalies, is a very meticulous task that often requires moremanpower than is generally available. This has prompted the need for automated detection and segmentation of sequences of interest [6].

The proposed system tends to create an application for the identification of anomaly in public places in real time [6]. The proposed system can be used for surveillance in areas such as malls, airports, train stations, etc. where there is a chance of theft or shooting [6]. The proposed system will use deep learning and neural networks [1] to train our system. This model will then be implemented as a desktop app that will take real-time [6,7] CCTV footage as input and send a warning to the administrator when a suspicious activity is detected. Real life implementations range from gaming to AR/VR, to health care and gesture recognition [7]. Compared to the image data domain, very little work is being done to apply CNNs [8] to the video classification. This is because a video is more complex than a picture because it has a different dimension-temporal [8].Unsupervised learning [8] exploits the temporal dependency between frames and has proved to be effective in video analysis [8]. Some anomaly activity approaches use CPU instead of GPU so that anomaly activity can run on low-cost hardware like embedded systems and cell phones [9]. Low-cost depth sensors are another new technology for computer vision [9]. They are present in game consoles including the Kinect for Xbox 360. They are motion sensors that allow the user to communicate with a console without a game controller, only by hand gestures [9]. These are RGB-D sensors that provide depth information through standardized lighting technology [9]. However, these sensors are limited to indoor use and their low resolution and noisy depth information make it difficult to estimate human anomaly in depth images [9].

Real-world anomalous events are complicated and diverse [9]. It is hard to list all of the possible malicious events at the same time [9]. Therefore, there is a need that the anomaly detection algorithm does not depend on any former information about the events [10]. In other words, anomaly detection should be done with minimal supervision. Sparse-coding based approaches [10] are considered as representative methods that achieve modern anomaly detection outcomes [10]. These methods assume that only a small initial portion of a video contains normal events, and therefore the initial portion is used to build the normal event glossary [10]. Then, the main idea for anomaly detection is that anomalous events are not accurately reassemble from the normal event glossary [10]. However, since the environment captured by 6479 surveillance cameras [10] can change drastically over the time, these approaches produce high false alarm rates for different normal behaviors [11].

It is desirable to learn an anomaly detection model which works well under multiple scenes with multiple view angles [11]. However, almost every current dataset is not perfect fit for such kind of assessment because of the need of scene variation [11]. In fact, almost all existing dataset contains videos

captured by one placed camera [9]. Our framework identifies an anomaly in a real-time manner. Our anomaly detection method has high true-positive and low false-positive rates which make it quite dependent [11]. We evaluate our anomaly detection and localization framework on popular datasets and provide a report of result on running time of our framework [11]. The comparison with modern methods shows the advantage of our method, in terms of performance and running time [11].

With the rapid development of deep learning methods, Convolutional Neural Networks (CNN) [11] are emerging as eminent ways of extracting representations (features) from imaging data [11]. However, like other machine learning methods of CNNs are prone to capturing any bias present in the task or dataset when not properly controlled [12].Inthis paper, we propose a representation learningscheme that learns features predictive of class labels with minimal bias to any generic type of protected variables [12].

II. LITERATURE SURVEY

The definition of an anomaly is always depending on what context is of interest. A video event is considered as an anomaly if it is not very likely to occur in video [6]. In 2015, For describing such unusual events i.e., anomaly in complex scenes Real-time anomaly detection and localization method [6] was developed, where each video is defined as a set of non-overlapping cubic patches and is described using two local and global descriptors. These descriptors capture video properties in a variety of ways. By incorporating simple and cost-effective Gaussian classifiers [6], the system can distinguish normal activities and anomalies in videos. The local and global features are based on structure similarity between adjacent patches and the features learned in an unsupervised way, using a sparse autoencoder [3,7,8]. The experiments confirm that the system can reliably detect and localize anomalies as soon as they happen in a video.

Another research presents, an unsupervised dynamic sparse coding approach for detecting unusual events in videos where, system accepts video sequence as an input then the proposed method uses a sliding [5] window along both the spatial and temporal axes to define an event. The task of unusual event detection is therefore formulated as detecting unusual group of cuboids residing in the same sliding window. A dictionary is first learnt from the video using sparse coding and later updated in an online fashion as more data become available. Given the learned dictionary, a reconstruction weight vector is learned for each query event and a normality measure [5] is computed from the reconstruction vectors. The proposed algorithm only needs to scan through the video once, and online updating of the learned dictionary makes the algorithm capable of handling concept drift in the video sequence. Finally, using sparse coding enables the algorithm to robustly discriminate between truly unusual events and noisy usual events.

In 2017, Weixin Luo, Wen Liu, recommended a Temporally coherent Sparse Coding (TSC) [7] in one of his paper, inspired by the ability to detect sparse coding-based anomaly, where the proposed system implement identical neighbouring frames with similar coefficients of reconstruction. Then the proposed system maps the TSC with a special kind of stacked Recurrent Neural Network (sRNN). The nontrivial selection of hyper-parameters to TSC can be avoided 1by taking advantage of sRNN in learning all parameters simultaneously, while the reconstruction coefficients [7] can be inferred within a forward passage with a shallow sRNN, which decreases the computational cost of learning sparse coefficients.

Yong Sheen Chong and Yong Haur Tay presented an effective method for detecting anomalies in videos. Recent applications of convolutional neural networks

[1,8], especially in photos, have shown promises of convolutional layers for object detection and recognition. Coevolutionary neural networks, however 1are supervised and involve labels as signals of learning. In videos that involve crowded scenes, the proposed system proposes a spatiotemporal architecture [8] for anomaly detection. Our architecture consists of two main components, one for the representation of spatial features and one for learning about the temporal evolution of spatial features. Experimental findings on benchmarks for Avenue, Subway and UCSD indicate that our method's detection accuracy is comparable at a substantial speed 2of up to 140 fps to state-of-the-art methods.

Authors	Algorithms	Drawbacks	Result
M Sabokrou, M Fathy, M Hoseini, R Klette	Gaussian Classifier Algorithm	Time consuming	The system representsavideo using both global and local descriptors. Two classifiers are proposed based on these two forms of representation. The fusion strategy on the outputs of these two classifiers achieves accurate and reliable anomaly detection and localization.
Bin Zhao, Li Fei-Fei, Eric P. Xing	Sparse coding technique	Less Accuracy	Experimental results on hours of real-world surveillance video and several YouTube videos show that the proposed algorithm could reliably locate the unusual events in the video sequence, outperforming the current state-of-the-art methods.
weixin luo,wenliu,sheng huagao	TCS and sRNN	Less Accuracy	Extensive experiments on both toy and real datasets demonstrate thar out TSC and sRNN based method consistently outperform existing method which validate effectiveness of our method.
yougsheanchong , yonghaurtay	Spatiotemporal Autoencoder	Time consuming	Experimental results on avenue, subway and UCSD benchmarks confirm that, the detection accuracy of method is comparable to state-of-the-art methods at considerable speed.

III. PROPOSED SYSTEM

In the proposed system first the user have register themselves and then their information i.e. name, address, phone number, email id,gender, username,password all this data will be stored in the database and then using username and password the user can log into the system. After that user have to select a video from the device and then after selecting the video the machine learning algorithm starts its working.

In this system we are using CNN algorithm, first the system takes video as an input and then the video is divided into different frames called segmentation and then the work of CNN begins the first two layers of CNN that is convolutional and pooling are used for feature extraction, after that fully connected layer is used for the final outcome that is classification.

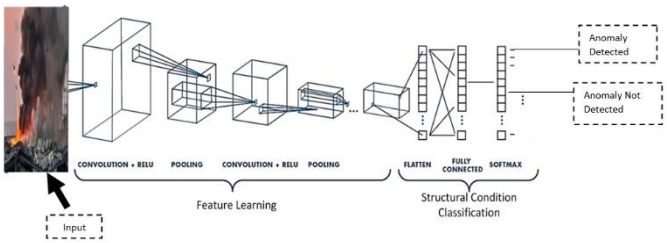

Fig.1. Above figure illustrates the working of CNN which contains layers that is convolutional, ReLU, pooling and fully connected

Working of CNN:

CNN consist of 3 layers i.e. convolutional, pooling, ReLU and fully connected these are further discussed below:

Convolutional layer:

The Convolutional layer is related to feature extraction.

It is a special operation applied on a particular matrix (, usually the image matrix) using another matrix (, usually the filter-matrix). The operation involves multiplying the values of a cell corresponding to a particular row and column, of the image matrix, with the value of the corresponding cell in the filter matrix. We do this for the values of all the cells within the span of the filter matrix and add them together to form an output.

Lets see the steps for convolutional layer:

Step 1: Line up the feature and the frame.

Step 2: multiply each pixel of frame with each pixel of feature.

For eg.

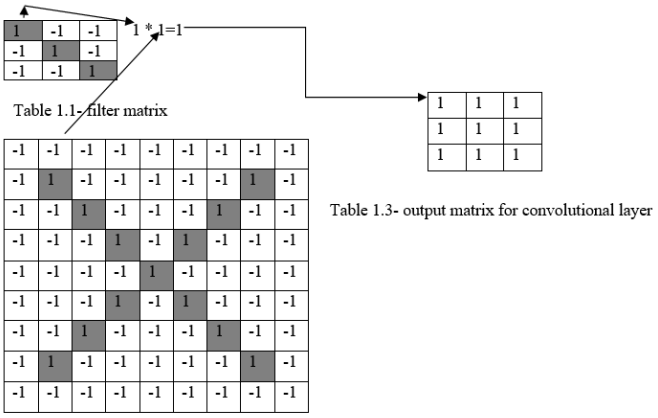

Table 1.1- filter matrix

1	-1	-1
-1	1	-1
-1	-1	1

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

Table 1.2- input matrix of image for convolutional layer.

1	1	1
1	1	1
1	1	1

Table 1.3- output matrix for convolutional layer

After multiplying the values add all the values and divide by the total no. of pixels.
(1+1+1+1+1+1+1+1+1)/9 = 1

Now to keep the track of where that feature was , we create a map and put the value of that filter at that place.

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

Table 2.1 – Input matrix for convolutional layer convolutional layer

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.0	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.0	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.0	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.0	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

Table 2.2 – final output matrix of

Similarly, we will perform same activity using different features and we will get matrix for that other filters.
Convolution is often represented mathematically with an asterisk * sign. If we have an input image represented as X and a filter represented with *f*, then the expression would be:

$$Z = X * f$$

ReLU Layer:

In this layer we will remove all the negative values from the matrix and replace those values with zero(0).it is an activation function.this is done to avoid the values from summing upto zero. ReLU transform function only activates if the value is above zero ,if the input is below zero than the output is

zero.so to remove that problem ReLU layer is used. *RELU or Rectified Linear Unit* is applied on all the cells of all the output-matrixThe basic intuition to derive from here is that, after convolution, if a particular convolution function results in '0' or a negative value, it implies that the feature is not present there and we denote it by '0', and for all the other cases we keep the value as it is.Together with all the operations and the functions applied on the input image, we form the first part of the Convolutional Block.

x	f(x)=x	F(x)
-3	f(-3)=0	0
-1	f(-1)=0	0
-0.51	f(-0.51)=0	0
-2	f(-2)=0	0

Table 3- working of ReLU layer function

Formula for ReLU layer is as follows:

f(x)=max(0,x)

For e.g.Lets apply ReLU layer to the matrix we created above then the matrix after applying ReLU will be as follows:

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	0.1	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.0	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.0	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.0	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

Table 4.1- input for ReLU layer

0.77	0	0.11	0.33	0.55	0	0.33
0	0.1	0	0.33	0	0.11	0
0.11	0	1.0	0	0.11	0	0.55
0.33	0.33	0	0.55	0	0.33	0.33
0.55	0	0.11	0	1.0	0	0.11
0	0.11	0	0.33	0	1.0	0
0.33	0	0.55	0.33	0.11	0	0.77

Table 4.2- output of ReLU layer

Pooling Layer:

The Pooling layer consist of performing the process of extracting a particular value from a set of values, usually the max value or the average value of all the values. This reduces the size of the output matrix. For example, for MAX-POOLING, we take in the max value among all the values of say a 2 X 2 part of the matrix. Thus, we are actually taking in the values denoting the presence of a feature in that section of the image**.** In this way we are getting rid of unwanted information regarding the presence of a feature in a particular portion of the image and considering only what is required to know. It is common to periodically insert a Pooling layer in-between successive convolutional blocks in a CNN architecture. Its function is to progressively reduce the spatial size of the representation to reduce the number of parameters and computation in the network.

Together with the **CONVOLUTIONAL LAYER** and the **POOLING LAYER,** we form the **CONVOLUTIONAL BLOCK** of the CNN architecture. Generally, a simple CNN architecture constitutes of a minimum of three of these Convolutional Block, that performs feature extraction at various levels. In our system we are using max-pooling.

In this layer we shrink the image stack:

1.Pick the window size as 2 or 3.
2.Pick a stride.
3.Move the window across the filtered images.
4.From each window, take the maximum value.

0.77	0	0.11	0.33	0.55	0	0.33
0	1.0	0	0.33	0	0.11	0
0.11	0	1.0	0	0.11	0	0.55
0.33	0.33	0	0.55	0	0.33	0.33
0.55	0	0.11	0	1.0	0	0.11
0	0.11	0	0.33	0	1.0	0
0.33	0	0.55	0.33	0.11	0	0.77

Table 5.1- output of ReLU layer taken as input for pooling layer.

1.00	0.33	0.55	0.33
0.33	1.00	0.33	0.55
0.55	0.33	1.00	0.11
0.33	0.55	0.11	0.77

Table 5.2- shrinked image matrix (output of pooling layer)

So after applying this from 8*8 matrix we will shrink that to 4*4 matrix. this will be done to all the other

outputs of the ReLU layer and all the output matrix will be shrinked and converted to 4*4 matrix.

After the images are shrinked we will again pass that oupt matrix through convolutional ,ReLU and pooling layer once again we will do this till the images are shrinked and converted to 2*2 matrix.

$$f_{x,y(S)=\max_0 1\, S_{2x}} + a, 2Y + b$$

Fully Connected Layer:

After the passing image from 2 different sets of convolutional, ReLU and pooling output is 2*2 matrix Now we will merge these 2*2 matrices into a vector .

$$Z = W.X + b$$

Here, X is the input, W is the weight and b is the constant, W in this case will be a randomly initialized matrix.

For example

1	0.55
0.55	1.00

1	0.55
0.55	0.55

0.55	1.00
1.00	0.55

Table 6.1, 6.2, 6.3 - shrinked images matrix takes as input for fully-connected layer

1
0.55
0.55
1.00
1
0.55
0.55
0.55
0.55
1.00
1.00
0.55

table 6.4- vector(Final output of fully connected layer)

After that the the predictions are based on that particular vector obtained.

Fig 2- some example images of anomalous and normal events

Implementation Details

OpenCV comes with many powerful video editing functions. In current scenario, techniques such as image scanning, face recognition can be accomplished using OpenCV.

Image Analysis is a very common field in the area of Computer Vision. It is the extraction of meaningful information from videos or images. OpenCv library can be used to perform multiple operations on videos.

Steps:

1. Open the Video file using cv2.VideoCapture()
2. Read frame by frame using read().
3. Release the VideoCapture and destroy all windows

In our system we are creating 30 frames per second. And the size of input video should be less than or equal to 5 mins and in mb it should be less then or equal to 50mb.

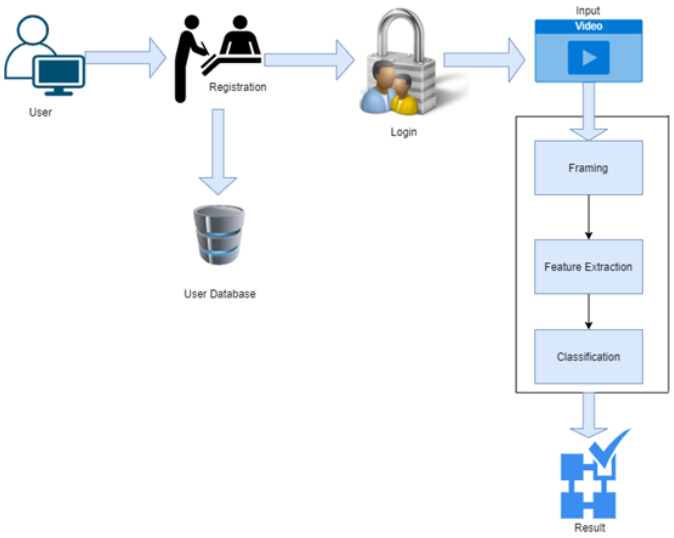

Fig 3- System diagram for Anomaly detection

EXPERIMENTAL RESULTS

We present a system to find suspicious activities using convolutional neural network, our system first registers the user and stores that details in the database and after that using username and password user can log in and then system takes video as an input then processing is done and the result or the output is generated in a text format if the suspicious activity is done then the message is displayed as "Anomaly detected" otherwise "No anomaly detected" is displayed.

For image classification there are many different machine learning algorithms , but CNN has proved to be efficient and reliable in many different ways. For example if we take deep neural networks in that the no. of parameters increases and more the parameters then more is the complexity which leads to over fitting problems. Thus our system uses CNN algorithm. And this system can be solution to many problems.

In this paper, we first empirically evaluate our proposed method under a controlled setting on a synthesized dataset, then we compare our method with Spatiotemporal Autoencoder, sRNN,Gaussian Classifier and Sparse coding technique.

Table 1 : AUC and EER of different methods

Method	AUC(%)	EER(%)
Spatiotemporal Autoencoder	80.3	20.7
TCS and sRNN	68	NA
Gaussian Classifier	99.6	2.5
Ours		

CONCLUSION

The proposed system analyzed and discussed about the Anomaly Activity detection and the advantages of implementing Machine learning-based Anomaly Activity detection using case study of manual CCTV Camera. The proposed system has mentioned and outlined the architecture of our system, design, and basis structure along with the brief introduction to the system. On that basis the proposed system has concluded that is one of the best technologies to be used for the Anomaly detection. Its more reliable more accurate and more efficient to be used as the part of Anomaly Activity Detection. A system for processing real-time CCTV footage will help to create better protection and less human intervention. There have been tremendous strides in the field of human anomaly operation, allowing us to better serve the countless applications that are possible with it. Moreover, research in related areas, such as Activity Tracking, can greatly improve its productive usage in many areas.

FUTURE WORK

The proposed system takes input from device videos and generate detection message on that particular frame, so in future we can upgrade this system to take live CCTV footages as inputs or connect this system to

CCTV cameras and detect the anomalies or any suspicious activities. And we can also generate alerts to the authorities. Other than that this system can also be upgraded so that it can detect activities even in darkness or in night time.

IV. REFERENCES

[1]. Steven Diamond, Vincent Sitzzmann, Felix Heide,Gordon Wetzstein, "Unrolled optimization with deep priors," 2018.

[2]. M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in Crowded Scenes," IEEE Computer Vision and Pattern Recognition, 2015.

[3]. B. Zhao, L. Fei-Fei, and E. P. Xing, "Online Detection of unusual events in videos via dynamic sparse Coding," IEEE Computer Vision and Pattern Recognition , Vols. 3313-3320, 2011.

[4]. 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 Recognition, vol. pp. 733–742, 2016.

[5]. Trong-Nguyen Nguyen, Jean Meunier, "Hybrid Deep Network for Anomaly Detection," vol. 1908.06347v1, 2019.

[6]. Z. Zhan, J.-F. Cai, D. Guo, Y. Liu, Z. Chen, and X. Qu, "Fast multiclass dictionaries learning with geometrical directions in MRI reconstruction," IEEE.

[7]. Jefferson Ryan Medel, Andreas Savakis, "Anomaly Detection in Video Using Predictive Convolutional Long Short-Term," 2016.

[8]. Trong-Nguyen Nguyen, Jean Meunie, "Anomaly Detection in Video Sequence with Appearance Motion Correspondence," vol. 1908.06351v, 2019.

[9]. Youg Shean Chong, Yong Haur Tay, "Abnormal event detection in videos using spatiotemporal Autoencoder," vol. 43000, 2017.

[10]. C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in MATLAB," IEEE International Conference on Computer Vision, vol. 2720–2727, 2013.

[11]. W. Luo, W. Liu, and S. Gao, "A Revist of Sparse coding-based anomaly detection in stacked RNN Framework," The IEEE International Conference on Computer Vision, 2017.

[12]. Naimat Ullah Khan, Wanggen Wan, "A Review of Human Pose Estimation from Single Image," IEEE, Vols. 978-1-53865195-7, 2018.

Cite this article as :

Sarika Late, Mrunmay Pathe, Riya Makhija, Jagruti Tatiya, Prof. Mrunal Pathak, " Anomaly Detection through Video Surveillance using Machine Learning ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 4, pp. 137-145, July-August 2021.

Journal URL : https://ijsrst.com/IJSRST21846