JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Intelligent Video Surveillance Through Deep Learning

Mohammed Shoeb,M.tech Scholar
Department of Computer Science and Engineering
Mahatma Gandhi Institute of Technology,
Hyderabad,India-500075 mshoeb_pg19cse2502@mgit.ac.in

C Ramesh kumar Reddy, Professor
Department of Computer Science and Engineering
Mahatma Gandhi Institute of Technology
Hyderabad,India-500075 crkreddy_cse@mgit.ac.in

Abstract : — Detecting suspicious activity in public places has become an important task as the number of shootings, knife attacks, terrorist attacks, and other incidents in public places around the world has increased. This paper focuses on using ***deep learning technique***s for identifying suspicious activities in videos using ***3D Convolution Neural Networks***. We present the architecture of our system, which can process video footage from cameras in real time and predict whether or not the activity is suspicious, in an educational environment. Deep learning techniques are used to identify suspicious or normal activity and send an emergency message signals to the corresponding authority, if a doubtful act is identified. Frequent Monitoring is happening through consecutive frames extricated from surveillance clips. Initially the work involves, extricating the video clips into frames and preprocess them; secondly the data is trained using 3D CNN model, to predict the suspicious activity in a video footage. We also propose future developments in the area of suspicious act detection in a video footage.

IndexTerms - 3DCNN Algorithm, video surveillance, Human Recognition technique, Deep Learning.

I. INTRODUCTION

The main objective of video surveillance is to develop smart video surveillance to oppose the traditional passive surveillance video such that unusual human activities can be recorded and after analysis, a notification can be generated through alarms, massages, or other methods to prevent unusual behavior. Abnormal activities such as abandoned component detection, robbery recognition, health identification of patients or elder home care (i.e. fall recognition), accidents or traffic rule breaking such as unauthorized U-turns, parking issues, and dangerous driving recognition on the road, violence recognition such as hitting, kicking, punching, slapping, shooting in public areas, and fire detection necessarily require the use of a smart monitoring system capable of automating [1]. Terrorists' use of explosives in public places has become a more dangerous activity in recent years. Terrorists target highly sensitive crowded public places like airports, bus stops, railway stations, government offices, and shopping centers. Terrorist goes to these locations and put their packages that contain bombs to be used in nuclear or bomb attacks [1]. It is extremely difficult for guards to keep a watch on crowded public locations and identify suspicious components. These types of attacks in public locations, which are recorded and traditionally investigated with cameras, are not fully evaluated by modern technologies. A real time smart video surveillance method can protect public areas by sensing and controlling abandoned belongings and making an alarm to alert security personnel to eliminate the components [2]. As a result, creating a fully automated, effective and efficient smart surveillance system is very helpful.

Video surveillance is a new area in which Artificial Intelligence, Machine Learning, and Deep Learning are being used. Artificial intelligence enables a machine to think in the manner of a human. Critical elements of machine learning are discovering from

training information and making assumptions on upcoming data. Due to GPU (Graphics Processing Unit) processors and large data-sets are now obtainable; the deep learning concept is preferred [3].

II. OVERVIEW .

The use of computer vision is associated with video surveillance will guarantee security and stability. The following steps are involved in computer vision techniques: modeling of environments, gesture recognition, and categorization of moving components, monitoring, behavior understanding, summary, and merging of data from various CCTV footages [5]. This technique necessitates extensive pre-processing in order to collect attributes from various video segments. There are two types of identification techniques: Supervised learning and Unsupervised learning. Supervised classification makes use of manually labeled training data, whereas unsupervised classification is entirely machine-driven and requires no human involvement [6].

The proposed system will use CCTV footage to monitor human behavior on a university and softly notify to the guards whenever a suspicious incident happens. Activity recognition and human nature identification are the two important elements of smart surveillance monitoring [3]. The video clip footage derived from university was utilized for test purposes.

The proposed system will utilize footages collected from CCTV camera for monitoring the human action behavior in a college campus and provide warnings when any suspicious action occurs. The significant components in smart video monitoring are action detection and human behavior identification. Understanding human activities automatically is a complex job [7]. Various areas on a university are under CCTV and different activities are to be surveilled. The recorded video achieved from university is being used for experimental purposes. Convolution Neural Network and Recurrent Neural Networks are two types of Neural Networks [3]. Convolutional Neural Network is employed to extricate high-level features from pictures to minimize the complexity of the input. Recurrent Neural Networks can be used for classification and is well suitable for video stream computation. The model used in the paper is 3D Convolution Neural Network that takes extricated frames to train the machine, so that the model can understand the difference between normal and suspicious activity in the video frames. The system will make use of a “VGG-16 (Visual Geometry Group)" pre-trained method that's been learned on the ImageNet data-set [4]. Presently, the framework has been trained to predict behavioral patterns on the footage. The method estimates dubious or normal human activity in recordings that is used to aid in the evaluation.

The majority of the present system relies on footage obtained through Surveillance cameras. If a crime or act of violence occurs, this clip would be used to conduct investigations. However, a system that automatically identifies any unordinary or abnormal circumstances in advance and has a process to alert the authorized person is more fascinating and can be implemented to both inside and outside locations. The recommended approach is to create such a technique in a campus area.

Numerous investigators use the particular procedure listed below to create an intelligent surveillance system capable of detecting the aforementioned abnormal human activities.

- *Background component Detection:* Foreground subtraction is an effective tool for detecting various changes in frame sequences and extracting foreground components [6].
- *Component identification:* in video sequence could be achieved either using non monitoring or monitoring approaches. To make the trajectory of an object over time, a tracking-based method is used to locate its position throughout every image frame [10].
- *Feature extracting*: A lot of algorithms extract structure and kinetic features of the object for component identification, and its extracted features is often delivered as input to the system.
- *Component classification:* Component classification is a process for separating the components in a video. The above process aids in distinguishing among various items such as individuals, automobiles, and so on. “Support Vector Machine”, “Haar-classifier”, “Bayesian, K-Nearest Neighbor, Skin color detection, and Face recognition” are some of the techniques used to classify objects.
- *Component analyzing:* After classifying the components in the video, behavior analysis is conducted for evaluating the different threshold values to ensure abnormal actions.

III. DESIGN AND IMPLEMENTATION:

In this section, we have represented a technique for identifying uninhabited components, identifying theft, identifying falls, and identifying violence (shown in below figure 1). The following significant step ladders are used to identify malicious human behavior: Detection of foreground components, tracking or non-tracking based component classification, extrication of features, classification, behavior analysis, and identification typically, research teams will update on these measure with distinguish methods or approaches in enhancing identification performance.

A. System Architecture

Figure1 describe the system architecture of the video surveillance process as described below:

- **Foreground component Identification:**

The removal of foreground component from video is the first and most essential step in detecting suspicious human activity. Background subtraction is really a strong instrument for observing the changes in frame sequences and extracting foreground components. In a video, foreground components are moving components as well as recently arrived components that have become stationary after a limited period of time, such as left luggage. Moving components, on the other hand, are considered the foreground components in background subtraction methods, whereas static components are recognized the background of the video. This theory helps to simplify the detection of motion components from a fixed camera video, but it is harder to identify recently arrived stationary components.

Background patterning and change identification based techniques are two techniques that can be used to detect motion components. To recover motion, change identification methods locate the distinguish among two adjacent frames and apply post-processing techniques to restore the whole components. These techniques are more expedient in terms of execution, but they are less accurate. Modeling-based technique attempt to produce the background model using temporal and spatial cues. In comparison to earlier techniques, a fairly accurate background prototype for the background can effectively remove the foreground components even more efficiently. In terms of implementation and execution, these techniques can differentiate from very easy to extremely complex.

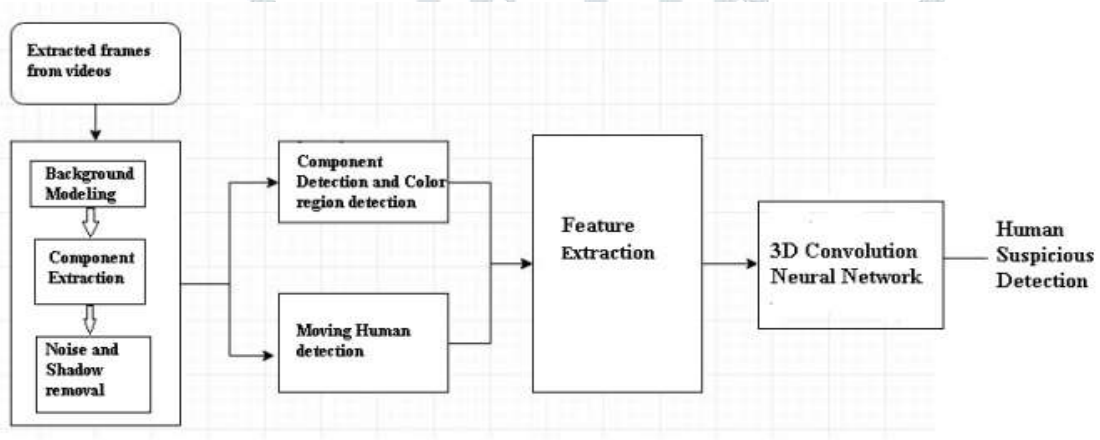

Figure 1 Block diagram for intelligent video Surveillance.

Figure 1 Techniques determining the distinction between different adjacent frames and then applying post-processing techniques to retrieve the entire components such techniques are quicker in terms of execution but cheaply accurate. Modeling-based technique attempt to produce the components with a little temporal and spatial cue in comparison to earlier methodology, a specific reference background method can extricate foreground components even more effectively. The contexts of formulation and execution, these techniques could range from easy to extremely complex.

Recently emerged stationary components in a clip could be harmful to the people as well as the general public. It's difficult to separate such stationary foreground components from surveillance footage using background subtraction techniques. Researchers used a variety of methods to retrieve and recognize stationary components.

- **Moving foreground components detection:**

Several researchers have worked over the last decade to detect moving foreground components in surveillance tape. These techniques aid in the retrieval of human activities like burglary, crowd running, vandalism, fights and threats, illegally crossing the border, punching, kicking, slapping, hitting, jumping, chain snatching, and falling from the background of surveillance tape applies general background technique based on a combination of Gaussian distributions. This technique uses a combination of different Gaussian distributions to handle a multimodal distribution. The proposed method does not accurately model a rapidly changing background due to its small Gaussian distribution. Background subtraction is a very common technique for segmenting foreground objects in a video sequence captured by a static camera, basically detecting moving objects from the difference between the current frame and the background model. To get good segmentation results, the background model should be updated regularly to adapt to the constant changes and changing lighting conditions of the scene Therefore, background subtraction is often inadequate to detect stationary objects and is therefore complemented using compliment method.

B. Stationary component identification:

Unusual activity detection involves the investigation of neglected components in order to avoid terrorists attacks. Foreground techniques in surveillance cameras comprises moving components as foreground components and stationary components as backgrounds. As a result, when a recently returned component will become static, is therefore consumed into the background. Several papers have different noise removal techniques with dual background initiatives and various understanding levels to retrieve the two foreground components for sensing stationary components in video content.

- **Shadow and Noise separating and luminosity manipulation methods**

Detecting foreground components without noise, lighting, or shadow is a critical task in the portion of computer vision. Noise enhance issues with object recognition, illumination ended up causing false identification, and shadow changes the looks of the components, making components tracking complicated.

C. Component Monitoring:

Component monitoring is a critical and difficult task in the discipline of computer vision. It aids in the generation of a trajectory of a component over time by identifying its place in successive frames of surveillance tape in order to analyze human activity. Points, component contour, component silhouette, elementary geometric patterns, articulated patterns, and skeletal frameworks are the component shape representations being used monitoring. Monitoring a component can be challenging at times due to visual noise, temporary occlusion of things, difficult component shapes, illumination changes, difficult component motion, and deformable component.

The majority of the proposed technique for anomalous activity monitoring relies on tracking data. These techniques do not perform in dynamic situations, such as scenes with a large number of people and a lot of occlusion. Several authors have prevented which use tracking-based abnormal behavior identification because of occlusion, complex component patterns, deformable things, and a stationary camera position, which all produce incorrect monitoring.

D. Feature Filtration:

Choosing suitable features is critical in fully automated detection of various activities from surveillance cameras. The primary goal of feature extrication is to discover the more hopeful knowledge in a video feed.

Feature filtration for abandoned component identification/theft identification: Detecting stationary components in clip is a difficult task. As a result, some components features are extracted from clip to distinguish among movable and static components.

A double foreground with varying learning methods: The use of a double foreground method with two separate long and the short learning methods. With these two distinguish learning operations; two foreground covers FL and FS are developed. If (FL; FS) = (1, 0), then components is stationary.

Centroid, height and width of a component Centroid are defined as an average of the pixels in x and y coordinates belonging to the component that can be measured through the given below formula:

$$R_x = \frac{\sum_{i=1}^{p} X_i}{N}$$

$$R_y = \frac{\sum_{i=1}^{p} Y_i}{N}$$

R*x* and R*y* =Centriod of x and y

N= component

i = 1,2,3…

Length and breadth are the Y and X axis of the distance. If the component Centroid, length and breadth in each state are similar then component is stationary.

E. Identification and unusual activity Detection:

Following the detection of shifting in addition to static foreground additives in a video frame, the element reputation method is used to decide whether the conduct is appropriate or anomalous. After locating shifting or desk bound foreground gadgets around video frame, the item category spot is implemented to the popularity of regular or bizarre conduct. For instance, a static character and an unused object in a crowded place could be handled as suspect objects if the capabilities of the objects are

uncertain. Items reputation certainly distinguishes among desk bound residing characters and desk bound deserted object, combating in addition to boxing, face and pores and skin color objects, hearth place and mild sensor, solar rays, or any synthetic mild, falling human pose and laying human pose, and so on.

Overall, there are 3 varieties of category techniques: feature primarily based totally, movement primarily based totally and pattern primarily based totally. The preliminary step was to decide which suspicious spots must be examined. We classified the following suspicious spots: Abuse, Arrest, Arson, Stalking, Robbery, Explosion, Fight, Accident, Robbery, Shooting, Shoplifting, Theft, and Vandalism. These 5 moves led to 5 lessons on categorical algorithms. Suspicious movements were up to exactly 14 categories. The next segment started getting records in each category.

A JavaScript code snippet is used to extract photos from Google Photos. After getting enough number of different pictures, the excess pictures were removed. This method began to be performed for each of the 14 categories. A version selection method is used as soon as a record is available. 3D CNN and i3d inception technique is adapted after looking at a complete neural community-based architecture for a real-time project.

F. Algorithmic Description and Model Analyzing:

Python language is used for deep learning programming with the help of available techniques such as Tensorflow, keras, PyTorch, Flux and others.

Convolutional neural network training follows the same conceptual framework as fully - connected network training. The gradients are generated by using a forward pass and back propagation, and that they are upgraded with first order technique like Adam in Keras. The major aspect is that trained variables in convolution layer serve as filters and biases.

3D CNN classification: After feature extraction from the dataset, the training of the data in the 3d CNN and i3d models to be initiated. 3D convolutions implement a three-dimensional filtration system towards the input data, filter follows a three directions (a,b,c) to measure low-level feature representations. Their output shape is a three-dimensional volume space similar to cuboidal shaped. They are useful for detecting activities in video content, 3D images, and so on. Following three - dimensional space it can also be used to measure the two - dimensional space inputs like images.

A filter containing size I x I applied to an input that includes RR channels is a I×I×*R* volume that conducts convolution on input has size *I*×*I*×*R* to give the output of a feature map (is also known as activation map) of size I×*R*×I.

The testing input (CCTV footages) performs three dimensional convolutions followed by intermediate convolutions features, rectification and max pooling layer and in 3D CNN the convolution kernel is also in 3D Kernel. The input volume of data in 3d CNN uses three dimensional weights to extract each pixel resolution which scans the entire clip to produce a feature map.

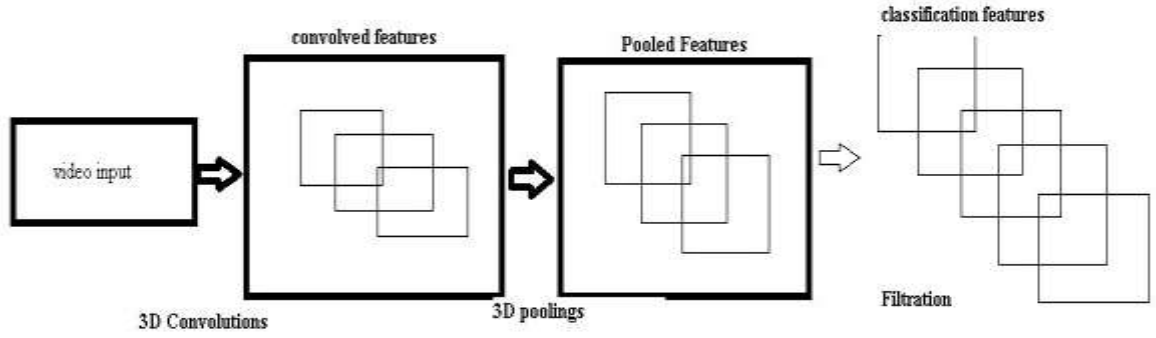

Figure 2: 3D Convolution Neural Network Pipeline

Figure 2 shows the classification of input data in 3D Convolution Neural Network:

Video input: The testing and Training data is given to the 3D CNN as a video clips after feature extraction process.

3D Pooling and Filtration: It reduces the size and divides the pixels to perform some convolution measurement to produce the pool layer by selection the maximum value from each pixel.

Here in the Architecture 3D CNN model we are giving the video frames containing size *E X W X L X H* where E represent channels, W represents width, H represents height and L represents length of video frames kernel size and 3D convolution are based on kernel depth and spatial size

Each video clip contains almost 28 frames, so we are providing each 3D input is having 224 X224 pixels among 3 different cannel colours with 28 consecutive frames.

There are three convolution modems are prepared where initial modem contains a 3D layers comprises 32 filters with kernel size around 3x3x3 in max pooling surface.1x2x2 is the size of a pooling surfaces with no temporal pooling across 2x2 pixels is

performed and make sure that temporary measurable data must not be thrown by any operations around pooling surface. In between we have converted string data into numerical data. Similarly second convolution modem takes place with filters 64 and third convolution modem contains 2 convolution layers with each filter around 128.Thereby the classification of 3D CNN process is completed.

The i3D Inception model: This model employs inception neural layers and intends to consider a various convolutions to improve performance using feature diversification. In general it employs a 3x3 convolution technique to reduce computational load.

As we know I3D inception contains two 3D convolution neural network surfaces .The classified 3D convolutions frames are focus on 3(LxHxB) in the inception since the inception is i3D, we build 3 dimensions to the frames .The i3D classified frames are then takes place in max pooling similar to the 3D CNN but the data is unsymmetrical to the nature in this model. They contain 16 and more units in the inception max pool with a limited group of frames and overlap each other with different groups. In every individual frame of 16 modems all groups are extricated and modem contain 128 x 128 sizes. Thus the given data is trained and ready for testing the video data clips.

The data is then trained which took around 4 hours to load in the machine after compiling the model. The sample is given below.

```
Epoch 00008: val_loss did not improve from 1.47325
Epoch 9/10
13/13 [=========] - 0s 20ms/step - loss: 0.6740 - accuracy: 0.8627 - val_loss: 1.8800 - val_accuracy: 0.3464
```

The final stage is testing stage in which the 10 videos are taken that contains robbery, fire accidents, car crash, fighting and others in the model. For sample we have took the moving individuals collections that gave the result as shown below.

```
CLASS NAME: HorseRace  AVERAGED PROBABILITY: 7e+01
CLASS NAME: Swing  AVERAGED PROBABILITY: 2.9e+01
CLASS NAME: TaiChi  AVERAGED PROBABILITY: 0.21
CLASS NAME: WalkingWithDog  AVERAGED PROBABILITY: 0.012
100%|██████████| 1213/1213 [00:00<00:00, 1281.24it/s]
100%|██████████| 1651/1651 [00:21<00:00, 76.05it/s]
```

The model gives the accuracy around 78% for the taken dataset video collections using 3D CNN and i3D inception model.

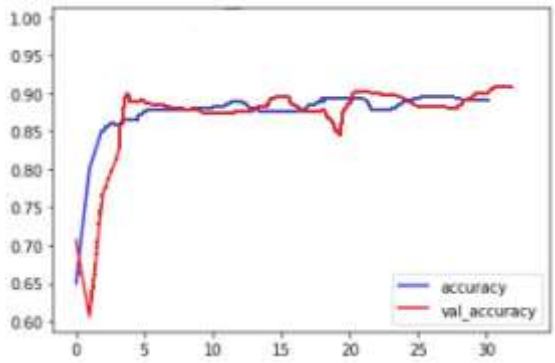

Figure 3 Accuracy Graph

IV. RESULTS

The training data that are used for training the machine to analyze the given surveillance video as suspicious or normal.

Figure 4: Dataset samples

- **Motions capturing**:

To capture the movement, two successive frames are necessary. If the component take actions, the pixel intensity in the consecutive frames will differ periodically when comparison to the prior frame. The below figure shows the normal and outcome pixel of the video content comprises two frames of the clip.

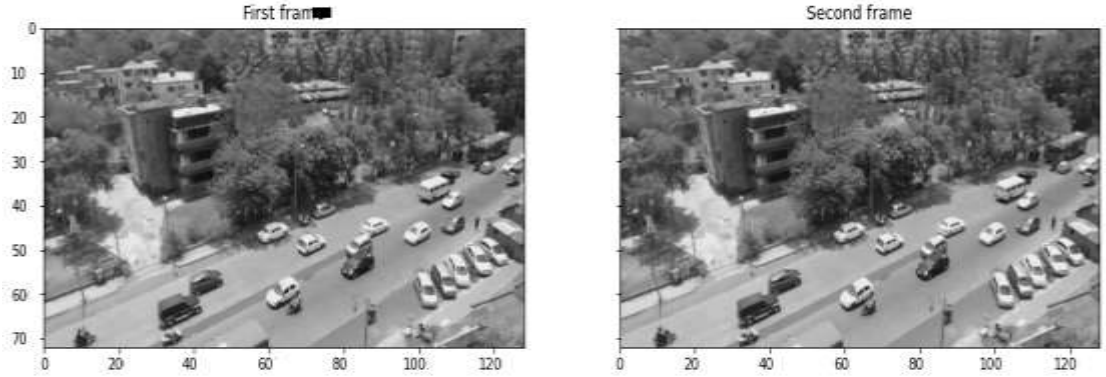

Figure 5 Movements and actions capturing

- **Background deduction**:

It acts a very normal method for separate prominence parts in any video sequence photograph by a changeless camcorder, essentially detecting motion components to equate the frame and the background technique. The Gaussian technique is used to calculate the background of the given frames as center pixel in the buffer of the clips as shown in figure below.

Figure 6: Background subtraction in the frames.

- **Component identification Outcome**:

To determine the components and objects in the frames, the pixels frames are compared with the previously available frames and then threshold the frames .The sample is shown below in figure 7.

Figure 7 Component Detection in the frames.

The prediction and result of the given input video shows that the given input test data Is suspicious or non suspicious are shown below Figure 8 and Figure 9.

Figure 8 Suspicious Recognition Activities in Test Data

In the figure 8, given the arson as video input to test, then the outcome of the prediction is given as Arson.

Figure 9: The Normal Activity of Test Input

Here the figure 9 show the normal output of the given video input while testing the data. The normal output of the video does not give any alarm signal because the input is not suspicious as shown in figure 9, while the abnormal event provides a warning to alert signals to the operator.

V. Conclusion

In the present generation, almost everyone understands the significance of CCTV footage, but in many other cases, these short clips are used for purposes of investigation after the violence has occurred. The suggested model is capable of preventing crime before it took place. CCTV footage is being monitored and analyzed accordingly. The outcome of the study is an instruction to the suggested authority for taking action, if the results suggest that an unfavorable incident is about to occur. As a result, this can be prevented.

Detection of abandoned component and detection of robbery of the majority of the research has been completed for the identification of abandoned components from surveillance tapes recorded by stationary cameras. Only a limited number of works recognized the stationary human as an unusual component. To address such issues, human identification techniques should be highly efficient, and the framework should verify for the availability of the owner in the incident; if the owner is invisible in the incident for an extended period of time, an emergency signal should be raised. To restructure the problems of thievery or component removing, the structure of the individual picking up the stationary component must match the owner's face; otherwise, a warning should be raised to alerting the guards. Future enhancements could include the creation of complexity cues inside the portion of 3D evidence aggregation and detailed occlusion analyzation. Spatial-temporal features could be stretched to three dimensions to improve unused component detection methods in a variety of complex surroundings. Future work on thresholding can increase the efficiency of the monitoring model by employing acceptable or hysteresis threshold strategies. A few works were also suggested for unidentified component identification using numerous camera angles. There is a wide range of possibilities for detecting abandoned components in video sequences captured by moving cameras.

Reference

[1] Wangli Hao, Ruixian Zhang,Shancang Li, Junyu Li,Fuzhong Li,Shanshan Zhao, and Wuping Zhang"Anomaly 2020. Event Detection in Security Surveillance Using Two-Stream Based Model" Security and Communication Networks(IF1.791) Volume 2020, Article ID 8876056.

[2] G. Sreenu, M. A. Saleem Durai 2019 "Intelligent video surveillance: a review through deep learning techniques for crowd analysis". *Journal of Big Data* volume 6, Article number: 48.

[3] Amrutha C.V, C. Jyotsna, Amudha J 2020. "Deep Learning Approach for Suspicious Activity Detection from Surveillance Video"Conference: 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).

[4] P.Bhagya Divya, S.Shalini, R.Deepa, Baddeli Sravya Reddy, December 2017. "Inspection of suspicious human activity in the crowdsourced areas captured in surveillance cameras",International Research Journal of Engineering and Technology (IRJET).

[5] Zahraa Kain, Abir Youness, Ismail El Sayad, Samih Abdul-Nabi, Hussein Kassem,2018. " Detecting Abnormal Events in University Areas ", International conference on Computer and Application

[6] Jitendra Musale,Akshata Gavhane, Liyakat Shaikh, Pournima Hagwane, Snehalata Tadge, December 2017. "Suspicious Movement Detection and Tracking of Human Behavior and Object with Fire Detection using A Closed Circuit TV (CCTV) cameras ", International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue XII.

[7] U.M.Kamthe,C.G.Patil 16-18 Aug, 2018. "Suspicious Activity Recognition in Video Surveillance System", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India. (**DOI:** 10.1109/ICCUBEA.2018.8697408).

[8] Tian Wanga, Meina Qia, Yingjun Deng, Yi Zhouc, Huan Wangd, Qi Lyua, Hichem Snoussie, January-2018. "Abnormal event detection based on analysis of movement information of video sequence",Article-Optik,vol- 152.

[9]Xu, H.; Yao,W.; Cheng, L.; Li, B, 25 February 2021 . "Multiple Spectral Resolution 3D Convolutional Neural Network for Hyperspectral Image Classification". Remote Sens. 2021, 13, 1248.

[10]LucaSmaira , JoãoCarreira (DeepMind), EricNoland (DeepMind), EllenClancy (DeepMind), Amy Wu , Andrew Zisserman (DeepMind) Wed, 21 Oct 2020 "A Short Note on the Kinetics-700-2020 Human Action Dataset" Remote sense Journal,Article *13*(7) .

[11] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola Jan 2021 "Dive into Deep Learning" MC University Canada ,Release version 0.16.1.

[12] P.Bhagya Divya, S.Shalini, R.Deepa, Baddeli Sravya Reddy, December 2017 "Inspection of suspicious human activity in the crowdsourced areas captured in surveillance cameras",International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 12.

[13] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang and Shanshan Hao,2019 "I3D-LSTM: A New Model for Human Action Recognition" School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China ,AMIMAIOP Conference Series Materials Science and Engineering 569(3):032035.