Smart Surveillance System for Anomaly Recognition

Kunal Kamble[1,*], *Pranit* Jadhav[1,**], *Atharva* Shanware[1,***], and *Pallavi* Chitte[1,****]

[1]Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Navi Mumbai

Abstract. Situation awareness is the key to security. Surveillance systems are installed in all places where security is very important. Manually observing all the surveillance footage captured is a monotonous and time consuming task. Security can be defined in different terms in different conditions like violence detection, theft identification, detecting harmful activities etc. In crowded public places the term security covers almost all type of unusual events. To eliminate the tedious manual surveillance we have developed a smart surveillance which will detect an anomaly and alert the user and authority without any human interference. It is a very critical issue in a smart surveillance system to instantly detect an anomalous behaviour in video surveillance system. In this project, a unified framework based on deep neural network framework is proposed to detect anomalous activities. This neural network framework consists of (a) an object detection module, (b) an object discriminator and tracking module, (c) an anomalous activity detection module based on recurrent neural network. The system is a web application where user can apply for three different security services namely motion detection, fall detection and anomaly detection which is applicable for monitoring different environment like homes, roads, offices, schools, shops, etc. On detection of anomalous activity the system will notify the user and responsible authority regarding the anomaly through mail with an anomaly detected frame attachment.

1 Introduction

The term security covers various type of abnormal events like explosion, arson, burglary, assault etc. But some sorts of anomalies like violence detection is very difficult task to recognize in crowded place because it involves group activities. The anomaly analysis of a input stream involving crowd is very difficult because of large number of people are involved. Smart video surveillance systems are spatio-temporal and are capable of improving situational awareness. The goal is to explore the concepts of multi-scale spatio-temporal tracking through the use of real time video analysis and long term pattern analysis to encourage situational awareness. The aim is to presents a unique approach to recognize anomalies in intelligent video surveillance systems.We aim to solve the issues that arises in manual surveillance systems. We aim to recognize human activity and classify them as normal or one of the anomalies. We also aim to recognize anomalies that won't require human action interference. If a anomalous activity is detected the user and the responsible authority will get notified through mail along with an attachment of image in which the anomalous activity is detected. The application areas of Surveillance cameras are public places e.g. streets, traffic lights, banks, malls, schools, etc. in order to maintain public safety. However, the surveillance systems are not that efficient.

*e-mail: kunalkamble20k@gmail.com
**e-mail: pranit812jadhav@gmail.com
***e-mail: atharvamshanware@gmail.com
****e-mail: pallavi.chitte@rait.ac.in

- This system aims on providing the user with 3 distinct types on anomaly namely motion detection, fall detection and anomaly detection.
- The system detect the anomaly and alerts the authorities but cam also be enhanced to discover a potential threat in advance for the incoming threat and hence, increasing the safety of people.
- The deep learning model is developed for only anomaly detection as motion detection and fall detection does not require an convolutional neural network to be trained on a dataset.
- Currently, the anomaly recognition model has been trained on different types of anomalous activities and normal activities. This can be broadened to integrate an enhanced variety of anomalies.
- The motion detection module doesn't require CNN model as it can be detected using only image preprocessing. The fall detection module only requires an object recognition framework to work hence, it saves a lot of time and results to quick response time.
- Video Surveillance systems, are being integrated with other industrial automation systems for intelligent application in fields like fire, access control, police services, manufacturing, transportation and many other fields so that surveillance cameras become a part of the Internet of Things.

Existing work focuses primarily on providing context-specific solutions on:

- Highways, traffic signals and main junctions

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

- Apartment buildings, houses or other residential areas
- Crowd pull gatherings
- Different religious festival gatherings
- Inside office, school buildings, malls and shops

Resulting, insufficiency in the implementation of cctv cameras and an unfeasible ratio of human monitors to surveillance cameras. Maintaining a constant watch using surveillance cams is very tiresome for the operators. In real word situations it is necessary for intelligence to be visible. In real world scenarios response time taken for generation is highly important. In emergency situations like stampede prediction of certain behaviour, actions and movements is highly useful. The primary contribution of this paper is to offer a deep learning framework for identification of anomalous activities. A framework that utilizes YOLO network and Kalman filter as an object detector and discriminator model for the video frames, from which VGG-16 based CNN model extracts multiple features and finally a LSTM model trained on these features classifies whether the activity being performed or the general environment is anomalous or normal. The result obtained by this model reflects the significance of the dataset and provide scope for further work.

2 Literature Survey

Anomaly recognition involves identifying human activities that are carried out in an anomalous event. This involve identifying the objects of interests for feature extraction. The region of interests part of the video with some human interaction or objects that effect the environment and leads to the anomaly. Since the final results depends on the identification, the particular objects of interest are crucially important. There have been multiple attempts in developing a human activity or anomaly recognition model. In 2020, Shreyas D, Raksha S and Prasad B implemented an "anomalous human activity recognition system" [2], their methodology uses adaptive video compression method which compresses parts of the video that are irrelevant, retaining only the objects of importance. This method combines adaptive video compression and 3D CNN with contextual multiple scales based on the temporal features, to provide an accurate anomalous activity recognition system that works in real time. This method can recognize anomalous events with an accuracy of 80 percent. In 2019, Dinesh Jackson Samuel, Fenil E and Gunasekaran Manogaran implemented a "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM" [4], which uses a spark framework to process real time videos from different sources,the video frames are separated and the features of each frames are extracted by using Histogram of Oriented Gradients function. Then the frames are labeled as violent and normal, which are used to train the Bidirectional LSTM network for recognition of anomalous activities. 94.5 percent accuracy was achieved using this method. In 2018, Waqas Sultani, Chen Chen and Mubarak Shah implemented a "Real-world Anomaly Detection System in Surveillance Videos" [6] which followed a method based on videos which are considered normal and violent as bags, and segments of video as instances, an anomaly ranking model is automatically learned that predicts high anomaly scores for anomalous segments of videos. In 2017, Kwang-EunKo and Kwee-BoSim implemented a "deep convolutional framework for abnormal behavior detection in a smart surveillance system" [9], the proposed method discriminates and tracks objects of same type by using object recognition algorithm and Kalman filter based object entity discriminator. Developed a deep convolutional framework for extracting dynamic feature of human behavior from a video frame. The VGG-16 feature extractor used in this method is pose based CNN, which identifies different poses that are involved in anomalous activities. This method managed to achieve 89 percent of accuracy. In 2016, D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe in "Learning deep representations of appearance and motion for anomalous event detection" [23], utilised deep learning-based auto encoders to develop a model of normal behaviours and reconstruction loss to detect abnormalities recently and proposed an unsupervised deep learning framework for building discriminative representations for video anomaly detection automatically. This method also known as double fusion scheme is based on learning the correlations between appearance and motion features which are combined to discover abnormal activities. In 2017, Cheng-Bin Jin, Trung Dung Do, Mingjie Liu and Hakil Kim implemented a "multilevel action descriptor" [5], which consists of three levels: posture, locomotion, and gesture level; each of which corresponds to a different group of sub-actions describing a single human action, for example, eating while sitting using appearance based temporal features with multiple CNN. Considering all the mentioned case studies for anomalous action recognition system, the viable approach to this problem is in implementing a anomalous recognition system using a CNN+RNN model. The regions of interest for anomaly detection involves different objects that affect change in the state of environment. For identifying region of interest, an object detection and recognition framework can be used. In 2020, A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao developed "YOLOv4: Optimal Speed and Accuracy of Object Detection" [1], which is an object detection framework. This network is trained on coco dataset for recognizing 72 different objects from images. YOLOv4 is a state-of-the-art detector which is faster in terms of FPS and more accurate than all available alternative detectors. YOLOv4 detects and recognized multiple object visible within the frame. It has a lot of features and out of them the features that improves the accuracy of both the classifier and and detector are selected. Training a machine learning model on dataset that includes data in form of videos requires implementing spatio-temporal tracking algorithm. This means the features extracted on single frame should be tracked throughout the sequence of frames. Hence it is necessary to implement a algorithm to track the object detected by YOLO network through entire sequence of frames. In 2012, Z. Fu and Y. Han implemented a algorithm, “Centroid weighted kalman filter for visual object tracking” [20]. It is a state estimation method which pre-

dicts the target state using the state model and combines the observation model to estimate the posterior probability density function of the state. Videos are sequences of images which are trained on a CNN network. Multiple examples of algorithm based on Convolutional Neural Network with impressive capabilities in computer vision problems have been developed and has managed to achieve good accuracy in tasks of object detection and recognition in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) 2012, ZF-Net, GoogLeNet, and ResNet were the best-performing entries to the ILSVRC in 2013, 2014, and 2015. Among which VGGNet was one of the top-performing models. In 2014, Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University proposed "Very deep convolutional networks for large-scale image recognition" [18]. Even for video recognition, VGGNet is simple to use and has a high recognition rate.

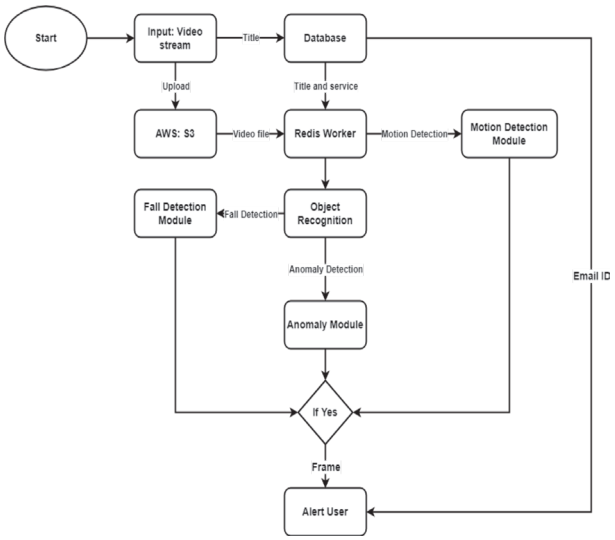

Figure 1. Proposed System activity flow

3 Proposed Methodologies

The goal of this problem statement is to develop a application that takes input from surveillance footage from the user's surveillance system, acknowledge any anomaly occurred in the footage and notifies the user. Mistreat, arrest, arson, assault, explosion, brawling, gunshot, and hooliganism are among the abnormalities. There are two broad categories of activities that human beings engage in: normal activities and anomalous activities. An anomalous behavior is when an individual behaves differently from normal in a way that causes harm to him or to others. Mental discomfort usually causes such behaviors. Intensive research into the recognition of human activity and its applications has shed light on detection of anomalies. The contexts identified are listed as application areas. The proposed system for anomaly detection takes stream of data i.e. frames from the video stream from different sources and analyze them for anomalous activities. Anomalies can be categorized in different types ranging from a simple motion in an unrestricted areas to a fight or an explosion. The sequence of frames of videos are handled according to the different category of the anomaly. Here the anomaly detection has been divided into three main categories:

- Motion Detection
- Fall Detection
- Anomaly Detection

The work flow of the proposed system is illustrated in Fig 1. The proposed system is a web application which takes video stream as an input from multiple users. These users selected category for anomaly detection is stored into the database. The system takes stream of data from different sources as input, breaks it down into small video clips and are stored on a cloud storage. The video clips on the cloud are fetched sequentially. The video clips are converted into overlapped frames and are passed to the respective detection to be processed further. If any find of anomaly is detected the user is alerted.

3.1 Motion Detection

Motion detection module detects any kind of motion captured and reports it. The motion detection module involve following steps:

1. Capture the video as a sequence of frames.
2. Read two frames from the sequence.
3. Get the difference between the current frame and the previous frame.
4. Apply Image manipulations like grayscaling and thresholding.
5. Finding Area of Contours to detect Motion.

3.2 Fall Detection

The fall detection module will recognize the human subject and the action of falling performed by the human. To recognize the human subject from the video an object detection framework is You only look one(YOLO) is used. The version of YOLO used in the proposed system is YOLOv4. YOLOv4 is a state-of-the-art detector which is faster in terms of FPS and more accurate than all available alternative detectors. YOLOv4 detects and recognized multiple object visible within the frame. It has a lot of features and out of them the features that improves the accuracy of both the classifier and and detector are selected. YOLOv4 recognize object within the frame and returns it's co-ordinates. The co-ordinates are used to determine the position of a person. Using these co-ordinates, height and width of the objects can also be retrieved. These co-ordinates are used for the fall detection of a human subject. Fall detection module involves following steps:

1. Capture the video as a sequence of frames.
2. Predict the person object using YOLOv4.

3. Retrieve the height and width of the object using the co-ordinates returned by YOLO.
4. Find the difference between height and width of the object between current and the previous frame.
5. If the height calculated in the current frame is smaller and the width is bigger than the person is considered as fallen.

3.3 Anomaly Detection

Anomaly detection is a machine learning framework Fig. 2 shows the flow of data through different modules for anomaly detection in a video.

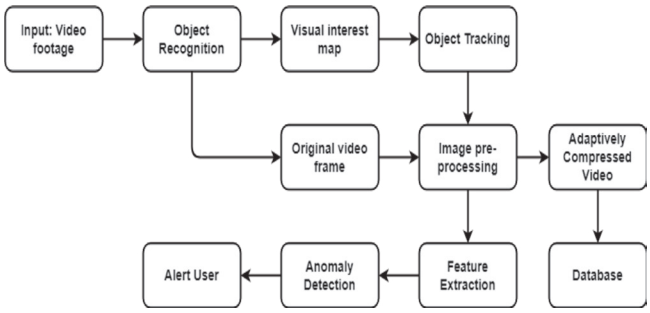

Figure 2. Anomalous Activity Detection

Input Module

This module takes video files as input. It converts video file into series of images and passes it to the next module.

Object Detection module

YOLOv4 object detection model is used for detecting different objects within each frame. It draws a bounding box around each object and mentions the object class. Following calculation shows the way each feature map is calculated:

$$W^1 = \frac{(W - F + 2P)}{S} + 1 \quad (1)$$

where W denotes the size of input image, F denotes size of receptive field, P denotes the number of pads, S denotes step size of stride, and W^1denotes the size of resulted feature map. The input image is equally divided into a cell grid with S x S cells by the YOLO network. The bounding boxes B and $Pr(obj)$, the probability that each box will contain the object are calculated through the forward propagation of the network in each grid cell. The confidence score is computed using the probability of the object in the box and this reflects both the expectation of the object in the box and the extent of accuracy of the box prediction. The confidence score is depicted as $Pr(obj) * IOU^{truth}_{pred}$. Thus $Pr(obj) * IOU^{truth}_{pred}$ denotes that a value is computed by the intersection over union between two B, namely a box for ground truth (GT) box and a box for the predicted bounding box. If an object is not present in the predicted B the confidence score corresponds to 0. If the B and the GT box are completely matched the confidence score corresponds to 1.

Object tracking module

A YOLO object detector is able to detect an object extremely fast and accurately. However, the model is not able to discriminate between object of same type. The YOLO network detects all the specified objects within the frame discounting of their number. YOLO network unifies the object localizer and classifier into single CNN network to speed up the process of detecting objects. To develop a practical applications such as a smart surveillance system only object detection without discrimination is not enough. Hence, it is important to discriminate and track object entities among a group of the detected objects of same type.The centroid weighted Kalman filter based object entity discriminator not only discriminates the objects with same label but also tracks them through multiple frames is integrated in the system. For implementing an activity recognition it is necessary to discriminate each object and track them which is done by the object tracking module. The tracking region of the dynamic target is selected using background subtraction and the target position at the beginning of tracking is predicted by Kalman filter, and then centroid weighted method is utilized to optimize the target position for further enhancing tracking accuracy.

Image pre-processing module

In image pre-processing frames with bounding box are resized to (244, 244).

Adaptive Video Compression module

: The resized frames are combined together to form a compressed version of the video file and save it into the database.

Feature Extraction Module

Feature extractor is an important step in machine learning. Feature extraction decreases dimensionality by condensing a large number of input variables into a smaller number of feature groups. Even though the dimentionality is reduced this does not disturb the description of the original dataset. The feature extractor module use VGG. The Visual Geometry Group at the University of Oxford devised the VGG convolutional neural network model for image recognition, where VGG16 refers to a VGG model with 16 weight layers. The input layer accepts an image with dimensions of (224 x 224 x 3), while the output layer is a 1000-class softmax prediction. The feature extraction component of the model runs from the input layer to the last max pooling layer (labelled by 7 x 7 x 512), whereas the classification part of the model runs from the input layer to the last max pooling layer (labelled by 7 x 7 x 512). The weigths used in the feature extractor are trained on 14 million

Anomaly Detection Module

The next step is the analysis of the dynamics of anomalous behaviour based on RNN model. Feeding the output

of a particular layer to the input in order to predict the output of the layer is the basis of the principle on which a RNN works. RNN accepts the current input data and previously received inputs in order to handle the sequential data. Due to the internal memory present in a RNN it is able to memorize the previous inputs. In RNN the data is cycled through a loop to the middle hidden layer. The input layer in RNN takes x as an input, processes it and passed it to the middle layer. Multiple hidden layers are present in the middle layer h and each has its own activation functions, weights and biases. Each layer generates out y. Fig 3 shows the working of RNN.

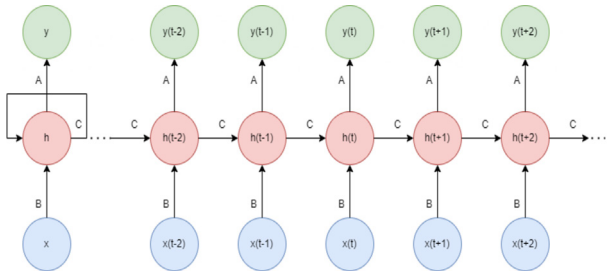

Figure 3. Working of RNN

The internal state of the network created by the architecture of RNN allows it to exhibit dynamic temporal behaviour. RNN uses backpropagation through time (BPTT) to perform weight update along with the time axis. However, the BPTT algorithm causes the RNN suffer through the Vanishing gradient problem. For an anomaly detection, it is assumed that it is necessary to identify the action sequence from the input which is also is assumed to be a sequence representing a series of actions. In an anomalous series of action the current always tend to be dependent on the previous or vice versa. The weight gradient gradually disappears in the learning process of a RNN model due to the increase in time interval of dependency. Long short-term memory networks (LSTMs) is a special type of RNN that are designed to avoid long term dependency problem. The LSTMs are suitable for usage in action recognition applications. By preventing gradient vanishing during the learning period, the structure overcomes the problem of long-term reliance. At every time step an input data sequence is received by the LSTM and the cell state and cell output vector is the result derived by each cell. Utilization of these two vectors is done at the next time step in the propagation of LSTM. The cell gives an output which is estimated as a class estimation which interprets the behaviour as normal or anomalous corresponding to the input data sequence as shown in the Fig 4.

The operation of a LSTM cell at time t is described. The output vector from the VGG-16 feature extractor is received as input X_t by each cell, along with the previous time's cell $t-1$, C_{t-1}, and the cell output h_1. The cell state updating procedure is developed in the following order based on these inputs: forget-input-output gate:

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}X_t + b_f) \tag{2}$$

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}X_t + b_i) \tag{3}$$

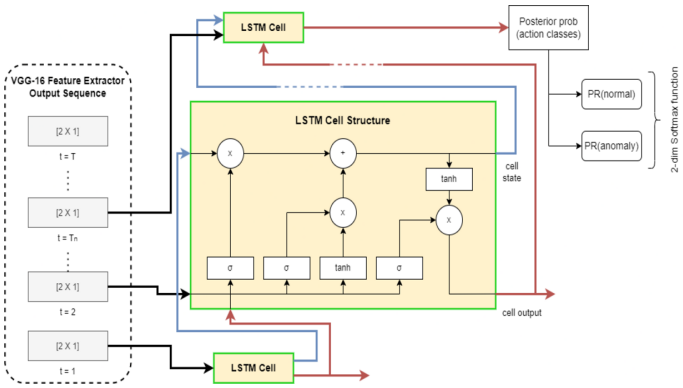

Figure 4. Anomaly detection process using LSTM with vector sequence recieved from VGG-16 feature extractor

$$o_t = \sigma(W_{oh}h_{t-1} + W_{ox}X_t + b_o) \tag{4}$$

h_t the final output and C_t the cell state of the corresponding cell at time t are obtained as follows:

$$C_t = f_t * C_{t-1} + i_t * tanh(W_{ch}h_{t-1} + W_{cx}X_t + b_c) \tag{5}$$

$$C_t = f_t * tanh(C_t) \tag{6}$$

The final output h_t from cell at time$t$ is the object class that is obtained from the LSTM. Based on the input given to the LSTM network at that time point along with output generated by that cell at previous time point the LSTM network outputs the result of behaviour class estimation i.e. normal or anomalous. The proposed method identifies ongoing activity from continuous streaming videos and enables in identifying whether the ongoing activity is normal or anomalous without completely perceiving the entire image sequence.

4 Results and Analysis

The primary objective of this proposed system is to develop a framework that is applicable in smart surveillance system for monitoring the well being of physically disabled or elderly people and also in infrastructure which require constant monitoring for commotion. A system capable of detecting normal and abnormal behaviour from the ongoing feature input generated from a video clip that of standard RBG image frame sequence has been developed. The hardware and software specifications of the environment in which the proposed system is developed and tested are as follows:

- Hardware specifications:
  - CPU: AMD Ryzen 5 3550H (8 CPUs, 2.1GHz)
  - GPU: NVIDIA GeForce GTX 1050 (3GB) CUDA enabled
  - Memory: 12GB
- Software specifications:

– Windows: 10
– CUDA: 10.1
– CUDNN
– OpenCV: 4.5
– Python: 3.7

4.1 Fall Detection

This module only utilizes YOLO object detection framework. The way object detection works is consider Fig. 5, which is a normal image, after giving this image as an input to YOLO network it gives Fig. 6 as an output.

Figure 5. Normal Image

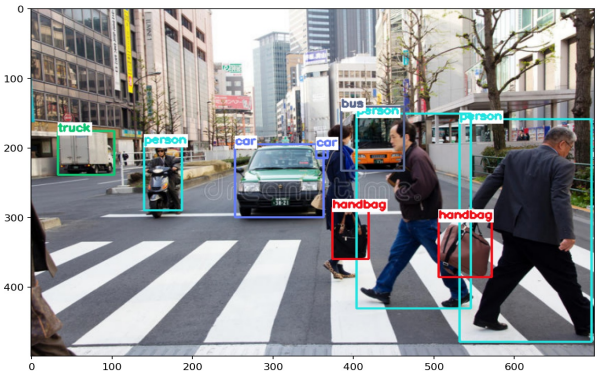

Figure 6. Object detection output

As it is seen in Fig. 6 that objects such as people, car motorcycle present in the Fig. 5 is displayed in bounding boxes. YOLO network returns the co-ordinates for the bounding box of the object along with the probability and class labels which is used to display the frames along with boxes and labels highlighting the objects within the image. The fall detection algorithm will takes sequence of images as output and will detect the person visible in the frames and whether the person is standing up or has fallen down. Fig. 7 shows a person standing upright, the YOLO network detects the person and determines the person is standing and is alright.

In Fig. 8 it is visible that the person is lying on the floor hence the algorithm detects that the detected person has fallen down and does the required action.

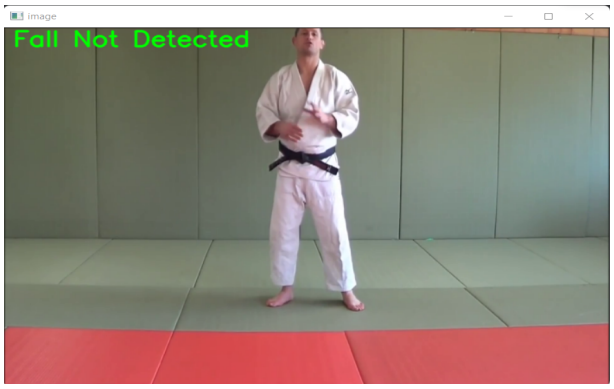

Figure 7. Person standing upright

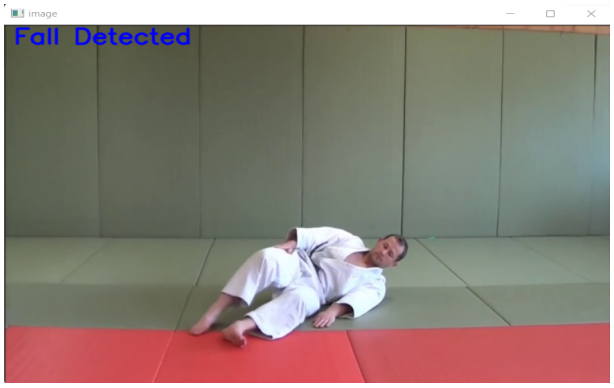

Figure 8. Person fallen down

4.2 Anomaly Detection

The proposed method for anomaly detection is trained and tested on UCF Crime dataset. UCF-Crime dataset is a new large-scale dataset which consists of 128 hours of 1900 videos which are untrimmed real world surveillance footage captured on cctv cameras, with 13 realistic anomalies including arrest, shoplifting, arson, burglary, assault, explosion, vandalism, fighting, road accidents, robbery, abuse, shooting, stealing. These anomalies are selected because they have a significant impact on public safety. This categories present in the dataset is reduced to involving only arrest, arson, assault, explosion, vandalism, fighting, abuse, shooting, stealing which consists of 550 video clips. This is done to improve the computation with respect to the hardware specification. The above mentioned activities are considered as anomalous activities and the performance the methodology is evaluated. YOLO object detection and discriminator is implemented through Darknet library which is a open source deep learning library. Fig. 9 is the image in which the objects detected in Fig. 6 are discriminated fraom one another. Features are extracted from these frames which are used for training the VGG-16 feature extractor. VGG-16, a pre-trained CNN model based on large scale image dataset is used due to the dataset being small which is compatible for using a transfer learning approach. Caffe library, an open source deep learning framework is used in utilizing the VGG-16 feature extractor. Finally the feature vectors extracted from VGG-16 feature extractor is used for training the LSTM

Figure 9. Object Discriminator Output

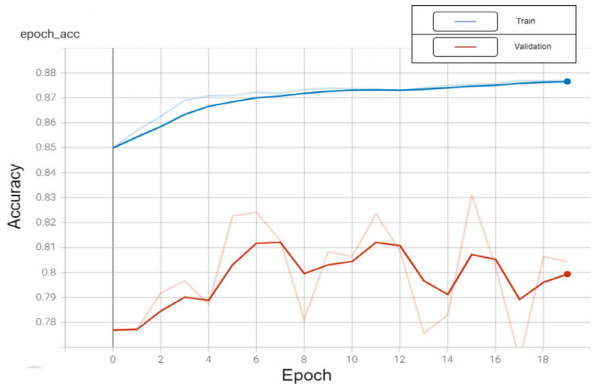

Figure 10. Train vs Validation Accuracy

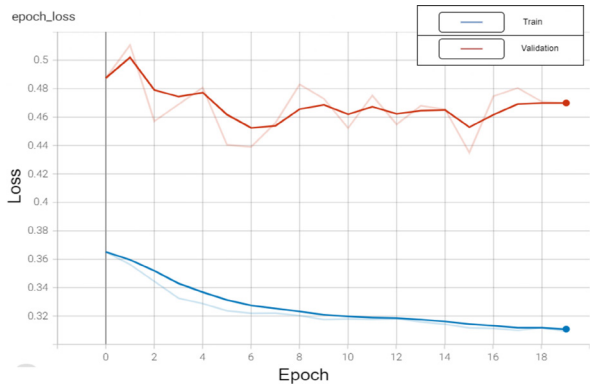

Figure 11. Train vs Validation Loss

network. For validation on the anomaly detection system the, precision and recall values are calculated.

The precision is measured as the ratio of the number of correctly identified Positive samples to the total number of Positive samples (either correctly or incorrectly). The precision of the model in categorising a sample as positive is measured. The precision is calculated in the following manner.

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}} \tag{7}$$

The recall is determined by dividing the total number of Positive samples by the number of Positive samples accurately categorised as Positive. The model's ability to recognise Positive samples is measured by the recall. The higher the recall, the greater the number of positive samples found. The following is how recall is computed.

$$Recall = \frac{True_{positive}}{True_{positive} + False_{negative}} \tag{8}$$

The performance of the model across all classes is described by a metric called accuracy. It is useful when each class has equal importance. The ratio between the number of right guesses and the total number of forecasts is used to compute it. The accuracy is calculated as follows.

$$Accuracy = \frac{T}{T + F} \tag{9}$$

Where,

$$T = True_{positive} + True_{negative} \tag{10}$$

and

$$F = False_{positive} + False_{negative} \tag{11}$$

Accuracy measured during each epoch of training and validating the LSTM network in given in Fig. 10. Loss is also an important factor in training a LSTM model. The increase in loss causes decrease in the accuracy and vice versa. Hence, it is advised to keep loss to the minimum while training a machine learning model. Loss measured during each epoch of training and validating the LSTM network in given in Fig. 11.

The classification report generated for validation of the LSTM models is given in Table 1. The classification report consists of precision, recall, f1-score and support valued of both the classes along with the accuracy

Table 1. Classification report of LSTM model

	Precision	Recall	F1-Score	Support
Normal	0.81	0.75	0.78	147595
Anomaly	0.80	0.85	0.82	172345
Accuracy	-	-	0.80	319940
Macro avg	0.80	0.80	0.80	319940
Weighted avg	0.80	0.80	0.80	319940

Another metric for evaluating the performance of machine learning model is a confusion matrix. It is is N x N matrix where N is the number of target classes. It is a tabular summary of the number of correct and incorrect predictions made by a classifier. Since the LSTM model can predict between two target classes, hence $N = 2$. Table. 2 is the confusion matrix generated from the validation of the LSTM model.

Table 2. Confusion matrix of validation

	Predicted values	
Actual values	Normal	Anomaly
Normal	111124	36471
Anomaly	26142	146203

The way this model can be implemented is, consider Fig. 12, the frame is from the video clip of a protect taken place on 27 February 2022 in St. Petersburg, Russia. The video clip depict police brutality shown while arresting a protester, which is considered as a crime, hence it is recognized as an anomalous activity by the anomaly detection system.

Figure 12. Anomalous activity detection

On the other hand Fig. 13 is a frame from video clip taken from a cctv camera located on street-side. The video clip depicts people walking on the street and doing their job. There isn't any crime taking place in that clip, it is just people following their daily routine, hence it is recognized as an normal event by the anomaly detection system

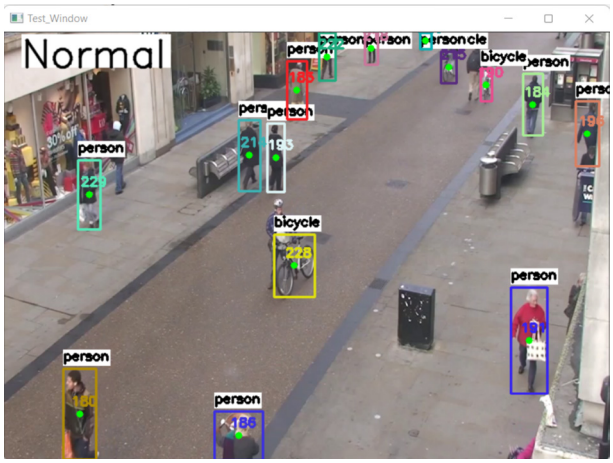

Figure 13. Normal Acivity Detection

After training and analysis of the model, it has been observed that the proposed system can recognize normal and anomalous events with an accuracy of 80.43% with precision of 80% and recall of also 80%

5 Conclusion and Future scope

Due to the increase in crime rates, there is a great need for surveillance systems to be installed in all public spaces, such as residences, schools, and streets. However, simply installing surveillance cameras will not prevent criminal activity; there must also be a structure in place to give prompt assistance to victims of crime, as well as swift action against offenders. This is possible through continuous and careful monitoring of the surveillance footage which requires a lot of manpower. Smart surveillance system for anomaly recognition is a automated system which reads the surveillance footage from the cameras and alerts the user if any anomaly is found. As this system is automated there is no need of manpower to monitor the videos as it is done by the system itself. This system is helpful in many environments such as hospitals, schools, offices, highways, etc. The current system will detect whether the event is anomaly or not and will notify the user and authority that an anomaly has been detected. This system can be further updated to recognize the type of anomaly occurred and notify the specific authority to handle the anomaly. As this system is developed is developed in a low end environment, the dataset used to train the model is limited and also more complex algorithm require high computational power for training the model, hence training the model in high end environment will give better result than the current environment. In the current environment, actions performed in the video are processed and detected in real time. The YOLO and Kalman based combined network labels each object within the frame, vgg-16 CNN network extracts features from each frames, these features are labelled as normal and anomalous by the LSTM network. The system classifies the sequence of frames as normal and anomalous at an accuracy of 80%,

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 23 2020.

[2] D. G. Shreyas, S. Raksha, and B. G. Prasad, "Implementation of an anomalous human activity recognition system," SN Computer Science, vol. 1, no. 3, p. 168, May 2020.

[3] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," Journal of Big Data, vol. 6, no. 1, p. 48, Jun 2019.

[4] D. J. Samuel R., F. E, G. Manogaran, V. G.N, T. T, J. S, and A. A, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," Computer Networks, vol. 151, pp. 191–200, 2019.

[5] C.-B. Jin, T. D. Do, M. Liu, and H. Kim, "Real-time action recognition using multi-level action descriptor and dnn," in Intelligent Video Surveillance, A. J. R. Neves, Ed. Rijeka: IntechOpen, 2019, ch. 4.

[6] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[7] Abdel, M., M, G., Rashad, H. and Zaied, A. N. H. (2018). A comprehensive review of quadratic assign-

ment problem: variants, hybrids and applications. Journal of Ambient Intelligence and Humanized Computing, 1-24.

[8] Huang W, Ding H, Chen G. A novel deep multi-channel residual networks-based metric learning method for moving human localization in video surveillance. Signal Process. 2018;142:104–13 (ISSN 0165-1684).

[9] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," Engineering Applications of Artificial Intelligence, vol. 67, pp. 226–234, 2018.

[10] Wang C, Yang H, Bartz C, Meinel C. Image captioning with deep bidirectional LSTMs and multi-task learning. ACM Trans Multimedia Comput Commun Appl. 2018;14:40.

[11] Zhang C, Tian Y, Guo X, Liu J. DAAL: deep activation-based attribute learning for action recognition in depth videos. Comput Vis Image Underst. 2018;167:37–49.

[12] Lee WK, Leong CF, Lai WK, Leow LK, Yap TH. ArchCam: real time expert system for suspicious behaviour detection in ATM site. Expert Syst Appl. 2018;109:12–24

[13] Dan X, Yan Y, Ricci E, Sebe N. Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst. 2017;156:117–27

[14] Tsakanikas V, Dagiuklas T. Video surveillance systems-current status and future trends. Comput Electr Eng.

[15] Feng Y, Yuan Y, Lu X. Learning deep event models for crowd anomaly detection. Neurocomputing. 2017;219:548–56.

[16] Pang S, del Coz JJ, Yu Z, Luaces O, Díez J. Deep learning to frame objects for visual target tracking. Eng Appl Artif Intell. 2017;65:406–20 (ISSN 0952-1976).

[17] Zhou S, Shen W, Zeng D, Fang M, Zhang Z. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process Image Commun. 2016;47:358–68

[18] Simonyan, K., Zisserman, A., 2014b. Very deep convolutional networks for large-scale image recognition, 2014b, aArXiv preprint arXiv:1409.1556.

[19] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks, IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1944–1957, Aug. 2013.

[20] Z. Fu and Y. Han, "Centroid weighted kalman filter for visual object tracking," Measurement, vol. 45, no. 4, pp. 650–655, 2012.

[21] Şaykol E, Güdükbay U, Ulusoy Ö. Scenario-based query processing for video-surveillance archives. Eng Appl Artif Intell. 2010;23(3):331–45.

[22] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, June 2016.

[23] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. In BMVC, 2015.