Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 173 (2020) 254–263

Procedia
Computer Science

www.elsevier.com/locate/procedia

International Conference on *Smart Sustainable Intelligent Computing and Applications* under ICITETM2020

Real-Time Anomaly Recognition Through CCTV Using Neural Networks

Virender Singh[a,*], Swati Singh[a,**], Dr. Pooja Gupta[a]

[a]*Computer Science and Engineering (CSE), Maharaja Agrasen Institute of Technology, Rohini, Delhi - 110086, India*

Abstract

Nowadays, there has been a rise in the amount of disruptive and offensive activities that have been happening. Due to this, security has been given principal significance. Public places like shopping centres, avenues, banks, etc are increasingly being equipped with CCTVs to guarantee the security of individuals. Subsequently, this inconvenience is making a need to computerize this system with high accuracy. Since constant observation of these surveillance cameras by humans is a near-impossible task. It requires workforces and their constant attention to judge if the captured activities are anomalous or suspicious. Hence, this drawback is creating a need to automate this process with high accuracy. Moreover, there is a need to display which frame and which parts of the recording contain the uncommon activity which helps the quicker judgment of that unordinary action being unusual or suspicious. Therefore, to reduce the wastage of time and labour, we are utilizing deep learning algorithms for Automating Threat Recognition System. Its goal is to automatically identify signs of aggression and violence in real-time, which filters out irregularities from normal patterns. We intend to utilize different Deep Learning models (CNN and RNN) to identify and classify levels of high movement in the frame. From there, we can raise a detection alert for the situation of a threat, indicating the suspicious activities at an instance of time.

Keywords: Security; CCTV cameras; Real-time; Deep Learning Models; Suspicious Activities; Surveillance

1. Introduction

Presently, there has been an increase in the number of offensive or disruptive activities that have been taking place

* Corresponding author. Tel.: +919999487762
** Corresponding author. Tel.: +919968856977
E-mail address: virenjhijaria32@gmail.com (Virender Singh), swatisingh180@gmail.com (Swati Singh)

10.1016/j.procs.2020.06.030

these days. Due to this, security has been given uttermost importance lately. Installation of CCTVs for constant monitoring of people and their interactions is a very common practice in most of the organizations and fields. For a developed country with a population of millions, every person is captured by a camera many times a day. A lot of videos are generated and stored for a certain time duration. Since constant monitoring of these surveillance videos by the authorities to judge if the events are suspicious or not is nearly an impossible task as it requires a workforce and their constant attention. Hence, we are creating a need to automate this process with high accuracy. Moreover, there is a need to show in which frame and which parts of it contain the unusual activity which aids the faster judgment of that unusual activity being abnormal or suspicious. This will help the concerned authorities to identify the main cause of the anomalies occurred meanwhile saving time and labour required in searching the recordings manually.

Anomaly Recognition System is defined as a real-time surveillance program designed to automatically detect and account for the signs of offensive or disruptive activities immediately. This work plan to use different Deep Learning models to detect and classify levels of high movement in the frame. In this work, videos are categorized into segments. From there, a detection alert is raised in the case of a threat, indicating the suspicious activities at an instance of time. In this work, the videos are classified into two categories: Threat (anomalous activities) and Safe (normal activities). Further, we recognize each of the 12 anomalous activities - Abuse, Burglar, Explosion, Shooting, Fighting, Shoplifting, Road Accidents, Arson, Robbery, Stealing, Assault, and Vandalism. These anomalies would provide better security to the individuals.

To solve the above-mentioned problem, deep learning techniques are used which would create phenomenal results in the detection of the activities and their categorization. Here, two Different Neural Networks: CNN [3] and RNN [4] have been used. CNN is the basic neural network that is being used primarily for extracting advanced feature maps from the available recordings. This extraction of high-level feature maps alleviates the complexity of the input. To apply the technique of transfer learning, we use InceptionV3- a pre-trained model. The inceptionV3, pre-trained is selected by keeping in view that the modern models used for object recognition consider loads of parameters and thus take an enormous amount of time in to completely train it. However, the approach of transfer learning would enhance this task by considering initially the previously learned model for some set of classified inputs e.g. ImageNet; which further can be re-trained based on the new weights assigned to various new classes. The output of CNN is fed to the RNN as input. RNN has one additional capability of predicting the next item in a sequence. Therefore, it essentially acts as a forecasting engine. Providing the sense to the captured sequence of actions/movements in the recordings is the motivation behind using this neural network in this work. This network is having an LSTM cell in the primary layer, trailed by some hidden layers with appropriate activation functions, and the output layer will give the final classification of the video into the 13 groups (12 anomalies and 1 normal). The output of this system is used to perform real-time surveillance on the CCTV cameras of different organisations to avoid and detect any suspicious activity. Hence, the time complexity is reduced to a great extent.

2. Related Works

The automated video surveillance system has risen as a significant research topic in the field of public security. A lot of work has been reported, addressing the movement recognition and tracking of an object. Artificial intelligence has also contributed a lot to reduce workload and increase the efficiency of surveillance. There have been numerous attempts to partially or completely automate this task with applications like Event detection, human activity recognition, and behaviour analysis.

Anomaly recognition is quite a challenging and time-honoured concern of Computer Vision [27][28]. As to recognize the violent or aggressive pattern from the recordings in real-time surveillance applications; the system needs to perform multiple attempts. Gaurav Kumar Singh and Vipin Shukla [23] released a conference paper on “Automatic Alert of Security Threat through Video Surveillance System”. This paper proposes an approach to use the sensor systems which can alert on the occurrence of any suspicious activity. Although Sensors account for events, it does not provide any information about them. By analyzing the captured video, information about the threat and cause can be obtained very quickly and accurately to take mitigating actions. Thus, the use of CCTV camera and sensor systems, independently or jointly, may not be sufficient for real-time detection of undesired events. So, they designed a system that will detect a threat in time under different lighting conditions using a camera and sensor networks. The system that has been designed was assisted with an intelligent, measurable, nifty and unswerving algorithm. The author detects, tracks and analyze the motion. A background- based estimation and body-based detection were performed to analyze the human outline and capture the human motion and using various edge detection algorithms. Datta et al. [24] suggested a methodology for utilizing motion and limbs movements of people to detect aggression. Gao et al. [25] have distinguished violence and aggression in the crowd using violent flow description. Recently, Mohammadi et al. [2] proposed a way to deal with the classification of violent and peaceful recordings utilizing behaviour heuristic-based approach. Apart from violent and normal patterns identification, some authors proposed to utilize tracking to recognize the anomaly and characterizing peculiarity as a deviation from that normal movement. Kooij et al. [1] detected aggressive actions by employing video and audio data from surveillance videos. However, to avoid tracking several methodologies are proposed and implemented by various authors. A few among them are learn global motion patterns through motion patterns [13], Hidden Markov Model (HMM) on local Spatio-temporal volumes [9], mixtures of dynamic textures model [5], histogram-based techniques [6], social force models [12], topic modelling [7], and context-driven method [8] challenges in acquiring reliable tracks. Bharath Raj [10] published an article, for implementing a Deep Learning based surveillance framework using Object Detection. Jason Brownlee [26] has employed deep learning along with CNN to detect movement and identify objects in video.

From the so far literature review, it has been observed that the maximum number of researches have designed methodologies for learning distribution of ordinary movements from the training done using available recordings. They have tried to identify low partable patterns and consider them as irregularities. Some have proved that the use of sparse matrices for representation are more effective while solving the problems related to Computer Vision [14]. Further, while testing, some patterns which are resulting in enormous restoration error are considered as anomalous. Deep learning has resulted in being best for image classification and hence, is found suitable for video activity classification. However, it is troublesome and arduous to find annotations during tracing done for recording. Deep learning-based encoders have been utilized that automatically train the model for ordinary behaviour and engaged restoration loss to distinguish peculiarities.

3. Proposed Model

3.1. Architecture

Anomaly Recognition System consists of a design composed of convolutional and recurrent neural networks.

- The first neural network is convolutional, which has been utilized to obtain the high-level feature maps of the images. This will reduce the intricacy of the input for the second neural net. We are utilizing a pre-trained model called inceptionV3 created by Google. This model applies transfer learning [20] as a widely used object identification models. This has several parameters and can require a large period of time to train completely. Henceforth, Transfer learning utilizes a previously learned model that simplifies loads of this work. The model has learned for various classes like ImageNet which is then re-trained for the weights of new classes.
- The second neural network is utilized as a recurrent neural net to extract meaning from the chain of the actions portrayed in a fixed time duration. This model will be used to classify the segments of videos as a threat and safe.

3.2. Software Implementation

Fig. 1. presents the outline of the proposed approach for Anomaly Recognition System.

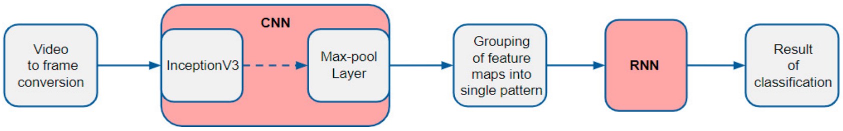
Video to frame conversion
CNN
InceptionV3
Max-pool Layer
Grouping of feature maps into single pattern
RNN
Result of classification

Fig. 1. Workflow of Anomaly Recognition System

The workflow of Anomaly Recognition System has been explained in the following steps:

- **Video to Frame Conversion:** Extracting frames from the captured CCTV recordings is the first step of this approach. The work extracts the frame after a fixed and small interval of time (say 1 sec). This extracted frame is then resized to the dimension 299x299 pixels which are the standard input dimensions for InceptionV3. The preprocess_input function is meant to adequate the resized image to the format that the model requires.
- **InceptionV3:** InceptionV3 is learned on the ImageNet dataset. It is a large dataset released in Visual Recognition competition. The model attempts to classify entire dataset into 1,000 categories, which is usually done in Computer Vision. The model concentrates general features from input pictures within the initial half. Later, it classifies those images based on the extracted features in the second half. The layers in the standard inception model are explained thoroughly in [22].
- **Convolution Neural Network:** The transfer Learning is used to train our CNN with the already trained InceptionV3 model. In transfer learning, we employ the feature extraction part to the new model and re-train the classification part with our original dataset. Since we don't need to learn the feature extraction part (which is a very complex part of the model), the overall learning process requires less computational resources and training time. The output of the Inception model is passed to the input of the CNN which isn't the final classification model. Rather, the outcome of the last pooling layer is extracted which is a vector containing 2,048 features to feed as an input to RNN. The vector is referred to as a high-level feature map.
- **Grouping of feature maps into a single pattern**: To give the framework a sense of the sequence, multiple prepossessed frames are considered. This chunk is then used to make the final classification. A chunk of these frames classifies a temporal segment of the video and can provide a sense of motion. For this, some feature maps are stored which are predicted by the inception model (CNN), generated in that fixed period of the video. Low-level features have been considered to generate a high-level feature map. These features are used for finding shapes and objects in computer images. This single combined feature map is then passed to the RNN. The reason to pass the feature map instead of the frame itself is to reduce the training complexity of the RNN.
- **Recurrent Neural Network:** The input of the second neural network is the concatenated collection of high-level feature maps generated in the previous step. This network has an LSTM cell with 5,727 neurons in the primary layer. This layer is followed by 2 Hidden layers. The first hidden layer contains 1,024 neurons with Relu as the activation function while the following layer has 50 neurons with sigmoid as the activation function. The actual probabilistic classification of the framework is produced the final layer having thirteen neurons with softmax as the activation function.

3.3. Hardware Implementation

Mostly, surveillance is carried with a view to monitoring a large portion of land. This brings forth the need to mull over some factors before computerizing surveillance. Moreover, this section explains the constraints for deep learning in surveillance and how we can overcome these constraints. We have 2 constraints for deep learning in surveillance: Video Feed and Processing Power.

Video Feed: Usually, to surveil or monitor a large area, multiple CCTVs are installed. These cameras require higher storage for recorded information; either locally, or at a faraway location. A good-quality recording can demand loads of memory than a low-quality recording. Memory being a limitation, we can't store a large quantity of information stream and thus quality is generally brought down to amplify storage capacity. Moreover, using a BW input stream instead of an RGB input stream can reduce the size by 3 times. Therefore, our deep learning surveillance system must be able to process low-quality videos as well. To tackle this issue, we have trained our model with videos captured at different durations of time with varying illuminations. The quality of our dataset is kept low to obtain better performance during real-time.

Processing Power: Where do we process the data collected from CCTV? This is a vital consideration in deciding the hardware cost of our system. This can be done in two ways:

- *Handling on a centralized server:* The extracted frames from the video streams recorded by the CCTV are processed by a GPU on a server operating at a remote location. This is a robust technique and allows us to achieve high accuracy even with a complex model. To overcome the problem of latency, fast Internet connectivity is required. Moreover, we have to use a commercial API to reduce the server setup and maintenance costs to a reasonable level. A lot of memory is consumed by most high-performance models.

- *Handling on the edge:* The transmission latency can be eliminated and abnormalities can be identified comparatively faster by appending a small microcontroller on the CCTV itself. Hence, we can perform real-time inference. Moreover, this is eliminating the dependency on the range of Wi-Fi/Bluetooth available and is an excellent add on for mobile bots (such as microdrones). However, the processing power of microcontrollers is comparatively less than GPUs. Thus, utilizing microcontrollers can bound our model to lower precision. This issue can be dodged by using onboard GPUs, however, that is a costly arrangement. Henceforth, we can install software packages like TensorRT, which can optimize our program for inference.

As examined previously, the frames in the CCTV feed may be of poor quality. Thus, the model must operate effectively in these conditions. An exceptionally exquisite way of doing that is by using data augmentation, which has been thoroughly explained in [19]. Introducing some noise to the frames can likewise reduce the quality of our dataset. Blurring the images and erosion effects are two effective ways for the same. In this manner, the ability to interpret low-quality recordings is a productive characteristic of the versatile real-time surveillance system. Hence, we have trained our model on such low-quality images as well. Also, we can process the data obtained from our camera sources by processing on a centralized server or processing on the edge. Handling on the edge is an excellent method - eliminating the transmission latency and reporting the variations from the norm faster than the previous strategies.

4. Dataset

4.1. Previously Used Datasets

Following is a concise review of the current video abnormality detection datasets. The UMN dataset [15] incorporates five distinctively arranged recordings where people walk around and begin running in various directions after some time. The suspicious activity in [15] is characterized by running action. In the following dataset, the author has used UCSD Ped1 dataset which contains 70 and UCSD Ped2 dataset having 28 CCTV recordings.[5] These videos have been recorded at a single place. The recorded suspicious activities are basic and do not add any significant value to CCTV surveillance. Avenue [11] is the third dataset that contains 37 recordings. Although it consists of comparatively more oddities, they are recorded at a single place. Similar to [5], the time span of the recordings in this dataset is not large with few anomalies such as throwing paper are unrealistic. Subway Exit and Subway Entrance datasets [16] forms the fourth dataset that contains single long surveillance recording for each. Basic irregularities like moving in the incorrect direction and payment skip are the part of this dataset. Lastly, BOSS dataset [17] is obtained from a surveillance CCTV installed on a train's roof. But these abnormalities are performed by actors containing peculiarities, for example, harassment, a person with a disease, panic circumstances, along with the normal recordings. All these previously utilized datasets are short in terms of quantity and time-span of the recording. Also, they cover limited anomalies and some abnormalities are not even realistic.

Table 1. Comparison of different dataset studied with our dataset.

	Video Count	Average Frames	Length
UMN	5	1290	5 min.
UCSD Ped1	70	201	5 min.
UCSD Ped2	28	163	5 min.
Avenue	37	839	30 min.
Subway Entrance	1	121,749	1.5 hours
Subway Exit	1	64,901	1.5 hours
BOSS	12	4052	27 min.
Ours	**1800**	**7247**	**128 hours**

4.2. Our Dataset

To overcome the incapabilities of previously used datasets, this work utilizes both normal and anomalous events captured by the surveillance CCTV cameras. It is a new and rarely explored large-scale dataset. These recordings are collectively 128 hours in length. It comprises various long and untrimmed real-world surveillance recordings captured at several locations, accounting for four realistic anomalies. Hence, these anomalies can be used to measure the degree of public safety.

Description of the dataset

- **UCF-Crime Dataset:** Text queries have been used to web scrap these videos from websites like LiveLeak and Youtube with minor alterations (for example “gunfire”, “public shootout”) for each anomaly, individually. The text queries in different languages are used to increase the dataset. Recordings that are manually altered, hoax recordings, news collected, the non-CCTV camera captured, or captured by a portable recording camera and containing aggregation are expelled from the dataset. The recordings in which the anomaly isn't clear have been removed too. To guarantee the quality of this dataset, 10 trained annotators are utilized with each having varying degrees of expertise in computer vision. Following the above constraints, we have gathered 950 unaltered real-world surveillance videos containing anomalies. Moreover, 940 normal scenarios are also collected, making a total of 1800 videos in this dataset. [18]
- **Annotation:** For detecting the variation from a norm, only labelled recordings are required to effectively train our model. However, to assess the performance of testing on recordings, the temporal annotations should be known. For this, the initial and last frames of the abnormal events for each recording has been identified. To mark the temporal extent for each anomalous behaviour, these videos are assigned to multiple annotators. The annotations from each annotator are averaged to obtain the temporal annotation. Intense efforts have been made for several weeks to create the final dataset.
- **Preprocessing Testing set:** Unnecessary footage like advertisements, inactivities and looped frames have been manually trimmed off from each video to reduce the size of the dataset and hence increasing the processing speed. The videos are also flipped horizontally to double the video counts to lower the overfitting in the training process. This is eventually increasing the effectiveness of the training process of the model. The entire dataset has been divided in the ratio 3:1 as training and validation set during the training of our model. The normal videos are shuffled with the videos containing the anomalies. Finally, these videos are passed through the pre-trained InceptionV3 model for dimension reduction.

5. Results and Discussions

In this work, we have trained 6 variations of our approach by altering different parameters and refining the dataset. The output layer of the RNN in model 1 has two neurons which are used to classify the entire dataset into 2 categories i.e. threat and safe. The anomalies considered for this model are Abuse, Arrest, Assault, and Arson along with a set of normal videos. The videos used are untrimmed and contain several unwanted footages. There are 940 chunks of un-shuffled frames with each chuck of 30 frames extracted at an interval of 1 second. The optimiser and loss function used for training this model are Adam and mean_squared_error resp. Fig. 2. (a) and (b) shows that the model trained is overfitted and producing fluctuating output with poor testing and training accuracy.

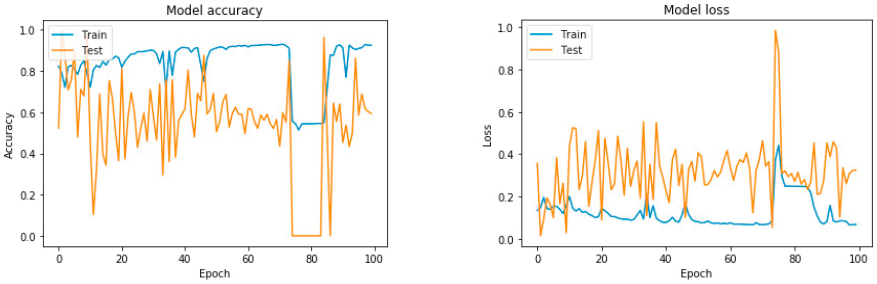

Fig. 2. (a) Accuracy of Model 1; (b) Loss of Model 1

To overcome the imperfections caused due to overfitting in model 1, the size of each chunk is reduced and the regularization [21] parameter is set to 0.01. The added term is considered to control the excessively fluctuating function. This prevents the coefficients to take extreme values. This forms the second model, Model 2. Fig. 3. (b) reflects that the loss function of model 2 is better tuned as compared to that of model 1. Moreover, overfitting is reduced to some extent as shown in Fig. 3. (a) producing a comparatively well-performing model.

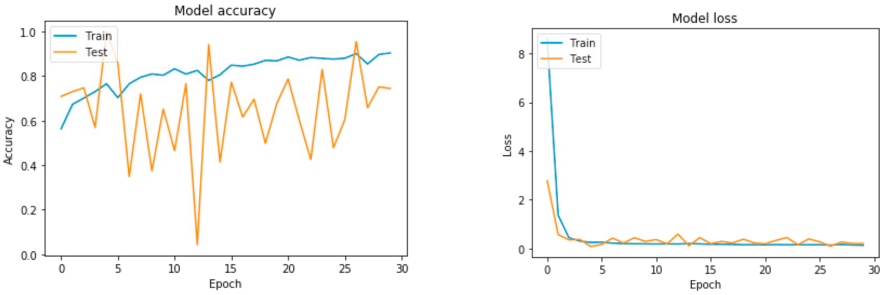

Fig. 3. (a) Accuracy of Model 2; (b) Loss of Model 2

Later, model 3 is designed to reduce overfitting to a considerable level. Along with regularization, the dataset is cleaned by manually trimming each video. Trimming is done to exclude the unwanted and useless footage from the videos that are causing improper training of the model. The dataset was shuffled this time which was not done in the case of model 1 and model 2. Moreover, the optimizer for model 3 is changed from Adam with the learning rate of 0.001 to SGD with the learning rate of 0.01.

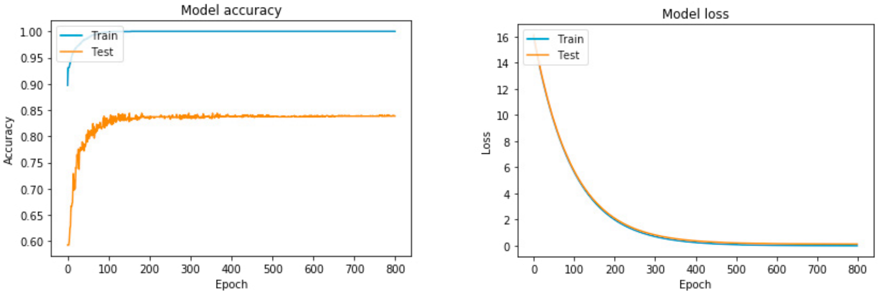

Fig. 4. (a) Accuracy of Model 3; (b) Loss of Model 3

Fig. 4. (a) shows that cleaning the dataset and changing the optimizer played an important role in reducing the overfitting in the trained model. But, the accuracy of the model is not as per expectations. Fig. 4. (b) presents a smooth curve for loss function with no fluctuation.

To enhance the accuracy of the model, eight more classes of anomalies; Road Accidents, Burglar, Explosion, Shooting, Fighting, Shoplifting, Robbery, Stealing, and Vandalism are added to the dataset. Also, the size of the chunks is decreased to 8. Model 4 is created to classify the segments of video into 13 different categories rather than 2 to give a better description of the anomaly detected by the algorithm.

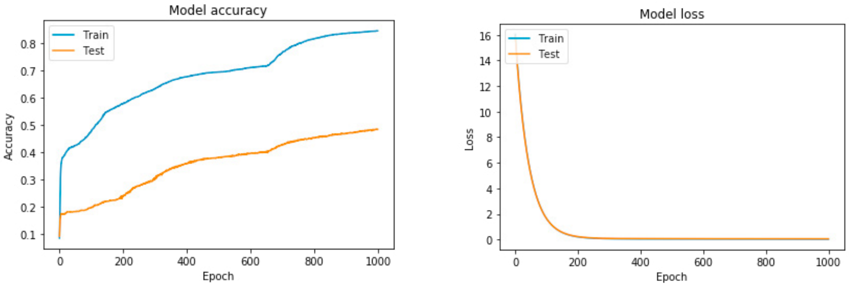

Fig. 5. (a) Accuracy of Model 4; (b) Loss of Model 4

Fig. 5. (a) and (b) shows that increasing the dataset alone, is not sufficient to improve the testing accuracy of the model.

In model 5, the loss function is changed from mean_square_error to categorical_crossentropy. The mean squared error (MSE) or mean squared deviation of an estimator measures the average of the squares of the errors while categorical_crossentropy (CCE) is a softmax activation plus a Cross-Entropy loss. It sets up a classification problem between classes more than 2 for every class in C.

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

where, y_i to the desired value and $\hat{y}_i$ to the actual value obtained.

$$CCE = \frac{1}{M}\sum_{p}^{M} -log\left(\frac{e^{s_p}}{\sum_j^c e^{s_j}}\right)$$

where, M: number of classes (Arson, Burglary, Shooting, etc), log: the natural log, s: is the CNN score for each positive class, 1/M: scaling factor to make the loss invariant and p: predicted probability observation o is of class c

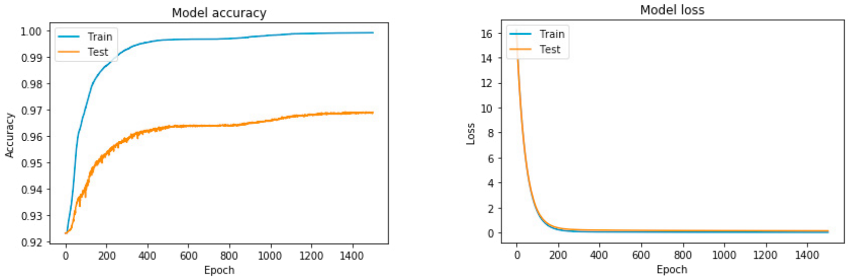

Fig. 6. (a) Accuracy of Model 5; (b) Loss of Model 5

Changing the error function has boosted the testing accuracy to a considerable amount and loss function is also converging faster as shown in Fig. 6. (a) and (b) resp. Thus, all the changes integrated so far have a positive impact on the overall performance of the model.

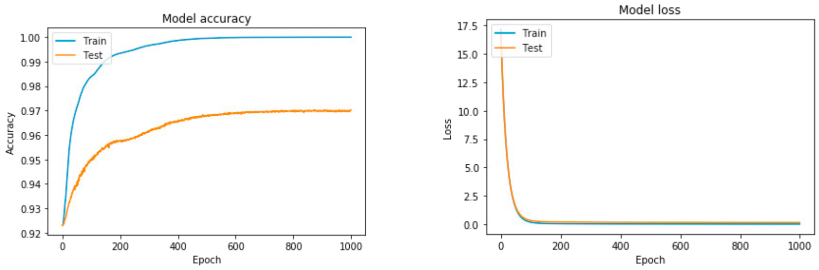

Fig. 7. (a) Accuracy of Model 6; (b) Loss of Model 6

Finally, the entire dataset is augmented by vertically flipping the dataset. Model 6 is trained on a dataset, double in size of the dataset used in model 5. This is done to optimise the validation accuracy of the model. Fig. 7. (a) shows that the training accuracy of the model increases logarithmically in accordance with the training curve. Fig. 7. (b) plots the loss function of the trained model.

Table 2. Performance Comparison of all models.

Models	train_loss	train_acc	val_loss	val_acc	Overfitting
Model 1	0.0652	0.9308	0.0142	0.9630	Maximum
Model 2	0.1819	0.9033	0.0793	0.9896	Considerable amount
Model 3	0.0062	1.0000	0.1333	0.8405	Reduced
Model 4	0.0248	0.8458	0.0569	0.4850	Reduced
Model 5	0.0148	0.9993	0.1341	0.9690	Reduced
Model 6	0.0098	0.9999	0.1548	0.9723	Least

Model 6 has comparatively the best overall performance amongst all the six models that have been implemented during the course of this work as shown in Table 2. Table 2 present the results obtained by six models that have been explained so far. Furthermore, we have tested model 6 on the self-collected dataset to examine its application in real-time scenarios.

6. Conclusion

This work suggests an approach to spot variation from the norm in real-world CCTV recordings. The normal data alone may not be effective to distinguish abnormalities in these recordings. Therefore, to handle the complexity of these realistic anomalies, both normal and anomalous videos have been considered and hence, maximised the accuracy of the model. Furthermore, to prevent the efforts-requiring temporal annotations of abnormal sections in training recordings, a general model of anomaly detection has been learned utilizing two distinct neural networks with a poorly labelled dataset. A rarely processed large-scale anomaly dataset consisting of 12 real-world anomalies has been utilized for learning with the aim of validating the suggested approach. The experimental results obtained during the work conclude that our suggested anomaly detection approach performs significantly better than the previously used methods.

Table 3. Details about the Optimized Model.

	Values
Categories Identified	Abuse, Burglar, Explosion, Shooting, Fighting, Shoplifting, Road Accidents, Arson, Robbery, Stealing, Assault, Vandalism, Normal
Chunk Size	8 frames
Optimizer	Stochastic Gradient
Error Function	Categorical Cross-Entropy
Regularization	Regularizers.l2 (0.01)
Activation Functions	Relu, Sigmoid, Softmax
Augmentation	Horizontal Flip

As specified in Table 3., this Threat Recognition Model classifies the anomalies into thirteen categories: Abuse, Burglar, Explosion, Shooting, Fighting, Shoplifting, Road Accidents, Arson, Robbery, Stealing, Assault, Vandalism, and Normal. The model concatenates 8 frames to form a chunk. The optimizer used in this model is Adam and the error function is categorical_crossentropy. The model uses three different types of activation function; Relu, Sigmoid, and Softmax. Moreover, to further increase the testing accuracy of the model the dataset has been doubled by flipping the videos horizontally. Hence, the overall accuracy of the model is 97.23% with reduced overfitting.

Eventually, to implement this model in real-time, it is necessary to consider all the hardware constraints explained in this work. Hence, a proper implementation plan will reduce the computation power, optimise the use of resources and eventually reduce the overall cost of the system.

7. Future Scope

This work offers the results of deep learning models for suspicious behaviour identification. The result of recognition obtained by these models reflects the significance of our dataset and provide the scope for further work.

The functionalities of this project can be scaled up in the future:

- In Anomaly Recognition System, the challenging part is the real-time execution of the model. A more effective and cost-efficient solution can be implemented in future to overcome this.
- The model can also be augmented to discover a potential threat and alert the authorities in advance for the incoming threat and hence, increasing the safety of people.
- Currently, the model has been trained on 12 anomalous activities. This can be broadened to integrate an enhanced variety of anomalies.

References

[1] J. Kooij, M. Liem, J. Krijnders, T. Andringa, and D. Gavrila. Multi-modal human aggression detection. Computer Vision and Image Understanding, 2016.

[2] S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio. Angry crowds: Detecting violent events in videos. In ECCV, 2016.

[3] Convolutional Neural Network (CNN) in Keras, https://towardsdatascience.com/building-a-convolutional-neural-network-cnn-in-keras-329fbbadc5f5

[4] Recurrent Neural Networks (RNN) in Keras, https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47

[5] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. TPAMI, 2014.

[6] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. In CVPR, 2011.

[7] T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behaviour in the video. In ICCV, 2009.

[8] Y. Zhu, I. M. Nayak, and A. K. Roy-Chowdhury. Context-aware activity recognition and anomaly detection in video. In IEEE Journal of Selected Topics in Signal Processing, 2013.

[9] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using Spatio-temporal motion pattern models. In CVPR, 2009.

[10] How to Automate Surveillance Easily with Deep Learning, https://medium.com/nanonets/how-to-automate-surveillance-easily-with-deep-learning-4eb4fa0cd68d, 2018.

[11] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in Matlab. In ICCV, 2013.

[12] IR. Mehran, A. Oyama, and M. Shah. Abnormal crowd behaviour detection using the social force model. In CVPR, 2009.

[13] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modelling of scene dynamics for applications in visual surveillance. TPAMI, 31(8):1472–1485, 2009.

[14] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In CVPR, 2011.

[15] Unusual crowd activity dataset of the University of Minnesota. In http://mha.cs.umn.edu/movies/crowdactivity-all.avi.

[16] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. TPAMI, 2008.

[17] Boss dataset, http://www.multitel.be/image/researchdevelopment/research-projects/boss.php.

[18] Waqas Sultani, Chen Chen, Mubarak Shah. Real-world anomaly detection in surveillance videos. In IEEE/CVF Conference, 2018.

[19] Data Augmentation, https://medium.com/nanonets/how-to-use-deep-learning-when-you-have-limited-data-part-2-data-augmentation-c26971dc8ced

[20] Transfer learning from pre-trained models, https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751, 2018.

[21] Regularization in Machine Learning, https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a, 2017.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In Google Research, 2015.

[23] Gaurav Kumar Singh, Vipin Shukla, Pratik Shah. Automatic Alert of Security Threat through Video Surveillance System. In 54th Institute of Nuclear Material and Management Annual Meeting, 2013.

[24] Biryukova EV, Roby-Brami A, Frolov AA, et al. Kinematics of human arm reconstructed from spatial tracking system recordings. J Biomech. 2000;33:985–995.

[25] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang. Violence detection using Oriented VIolent Flows. In ResearchGate, 2016.

[26] A Gentle Introduction to Object Recognition With Deep Learning, https://machinelearningmastery.com/object-recognition-with-deep-learning/, 2019.

[27] Waqas Sultani, Chen Chen, Mubarak Shah. "Real-World Anomaly Detection in Surveillance Videos", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[28] Junseok Kwon. "Rare-Event Detection by QuasiWang –Landau Monte Carlo Sampling with Approximate Bayesian Computation", Journal of Mathematical Imaging and Vision, 2019.