International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206355

Automatic Detection of Violent Incidents from Video Footage of CCTV Cameras

Baswaraju Swathi, B L Deepika Chowdary, K Sai Sindhu, Ashika P
Department of Information Science, New Horizon College of Engineering, Bangalore, Karnataka, India

ABSTRACT

In the current era, the majority of public places such as supermarket, public garden, malls, university campus, etc. are under video surveillance. There is a need to provide essential security and monitor unusual anomaly activities at such places. The major drawback in the traditional approach, that there is a need to perform manual operation for 24 ∗ 7 and also there are possibilities of human errors. This paper focuses on anomaly detection and activity recognition of humans in the videos. Computer vision has evolved in the last decade as a key technology for numerous applications replacing human supervision. We present an efficient method for detecting anomalies in videos. Recent applications of convolutional neural networks have shown promises of convolutional layers for object detection and recognition, especially in images. Experimental results on challenging datasets show the superiority of the proposed method compared to the state of the art in both frame-level and pixel-level in anomaly detection task.

Keywords : Computer vision, Convolutional Neural Network(CNN), CCTV, Unusual Objects

I. INTRODUCTION

Abnormal event detection and localization is a challenging and exciting task in video monitoring. Indeed, the security context in recent years has led to the proliferation of surveillance[6] cameras, which generate large amounts of data. This flow of CCTV images creates a lack of efficiency of human operators. Moreover, studies show that they can miss up to 60% of the target events when they monitor nine or more displays.

In addition, after only 20 min of focus, the attention of most human operators decreases to well below acceptable levels. This can lead to potential security breaches, especially when monitoring crowded scene videos. A possible solution to this problem is the development of automated video surveillance systems that can learn the normal behaviour of a scene and detect any deviant event that may pose a security risk. Abnormal event detection, also known as anomaly detection, can be defined as a spatial temporal recognition problem, taking into account that the event to be recognized is not present in the training phase. In the context of video surveillance[7] systems, anomalies are formed by rare shapes, motions or their combinations. The main challenge in abnormal event detection is extracting robust descriptors and defining classification algorithms adapted to detect suspicious behaviors with the minimum values of false alarms, while ensuring a good rate of detection.

The initial studies in abnormal event detection[10] focused on trajectory analysis, where a moving object is considered as abnormal if its trajectory doesn't respect the fitted model during the training period. The main limits of such method are the sensitivity to occlusion and the effectiveness of detecting abnormal

shapes with normal trajectories. These methods can be used in detecting deviant trajectories in scenes containing few objects but not achieve satisfactory performance for other applications. Other methods such as low level local visual features [6–8] tried to overcome the limits of trajectory analysis and construct models by using handcrafted feature extractors. Among these methods, low local features such as histogram of oriented gradient (HOG), and histogram of optical flow (HOF) have been used to model the background and to construct the template behavior. However, these methods are usually specific to a given application and are not optimal for complex events. Besides, they don't link between local patterns, since local activity patterns of pixels is not efficient for behavior understanding. More complex methods used the concept of Bag of Video words (BOV) by extracting local video volumes obtained either by dense sampling or by selecting points of interest to construct the template behavior. However, the relationship between video volumes is often not taken into account. Derivatives of these method attempted to enhance the previous models by using not only the local region, but also the link between these regions for the overall understanding of the events. Usually the complexity of these methods makes the min efficient and time consuming for detection[12] of abnormalities in crowded scenes. Nowadays, deep learning has aroused the interest of the scientific community and works have been carried out in several fields including agriculture, biology and economics. More specifically, deep learning based methods have demonstrated a high capacity on image processing, which led to the use of supervised methods[13] for the anomaly detection.

These methods are generally based on the use of convolutional neural networks (CNNs)[9]. The main drawback of the supervised methods is the use of both normal and abnormal examples in the training phase which makes them not usable in real-world application for video surveillance, because it is very difficult to identify and reproduce all the abnormal events. Other deep learning based works have achieved good performance on anomaly detection datasets. These methods use unsupervised learning and their learning process do not require normal and abnormal training examples, which makes them suitable for abnormal event detection. In this article, we propose an unsupervised on-line adaptive method based on deep learning for the detection and localization of abnormal events.

The main contributions of this paper are as follows:
We adapt a trained 3D CNN to extract robust feature maps related to shapes and motions which allow us to detect and localize complex abnormal events in non-crowded and crowded scenes. This robust classifier is able to represent all normal events (redundant and rare ones) during the training phase, detects abnormalities and adapt to the appearance of new normal events during the testing phase.

II. RELATED WORK

1) In [1] the discussion is purely in the context of visual surveillance. Though most of the papers discussed in this survey address anomaly detection, authors have observed four key issues with these methods:
(i) Benchmark dataset-based comparisons are used to show the effectiveness against the state-of-the art Though benchmarks may be relevant for comparisons, they may not contain all real-life situations. For example, though anomaly detection works fine on Avenue dataset, it gives higher false alarms when applied on a real dataset QMUL using two of the proposed methods. Therefore, we believe, the methods need to be relevant for real-life scenarios and should be applicable to long duration videos.
(ii) Secondly, due to the aforementioned trend, very limited amount of research have been carried out for developing generic techniques applicable to a variety of datasets.

(iii) There has been hardly any illumination independent research except for accident type anomaly detection. The problem is not entirely due to the limitations of the learning models. It is equally dependent on the dataset types and lack of illumination independent feature extraction. Possibly with the emergence of DNN-based modeling, we hope to address these issues in future. An object oriented approach might yield better results than histogram based approaches as human do not think of pixels and their motion in detecting anomalies, but with mere object motion observations. Researchers can make datasets containing segments of the same scene at varying illumination conditions.
(iv) Some approaches remove the background and focus on foreground features for anomaly detections.

2) In [2] this research, authors have successfully applied deep learning to the challenging video anomaly detection problem. They formulated anomaly detection as a spatiotemporal sequence outlier detection problem and applied a combination of spatial feature extractor and temporal sequencer ConvLSTM to tackle the problem. The ConvLSTM layer not only preserves the advantages of FC-LSTM but is also suitable for spatiotemporal data due to its inherent convolutional structure. By incorporating convolutional feature extractor in both spatial and temporal space into the encoding-decoding structure, we build an end-to-end trainable model for video anomaly detection. The advantage of our model is that it is semi-supervised – the only ingredient required is a long video segment containing only normal events in a fixed view. Despite the models ability to detect abnormal events and its robustness to noise, depending on the activity complexity in the scene, it may produce more false alarms compared to other methods.

3) The paper [3] tells about a novel online adaptive method based on combination of pretrained 3D residual network and online classifier were developed and implemented. It is able to detect abnormal events, prevent the marginalization of normal behavior that rarely occurs during the training phase and adapt to the appearance of new normal events in the testing phase. In addition, our method does not require pretreatment methods such as tracking or background subtraction. This method is based on two main stages: Spatiotemporal feature extraction without any need of training, and the use of robust incremental classifier that prevents the redundancy of information in CCTV. It can also either be used online or offline. Authors have tested our proposed methodology on two main datasets using crowded (Ped2) and non-crowded scenes (CapSec). The results from the Ped2 dataset showed high performance in detection and localization for abnormal events based on The EERFL and EERPL. To the best of our knowledge, this method outperforms all existing techniques present in the literature and used for Ped2. Beside, the fastness and the simplicity of this method allow us to use it for real-world application. The results presented in this paper showed the effectiveness of using this framework in detecting abnormal events. This method is robust, takes into account rare normal events present in the training phase. Besides, it can be incorporated in online CCTV. Moreover, the method can be adapted so that human operators select false alarms to prevent its future appearance, which is suitable for dynamic environment. In this method, the localization of abnormal events is reflected as patches in the original image. In some cases, these patches may overflow on normal regions.

4) In [4] the growing demand for a secure and safe environment has enhanced the research for developing smart automated surveillance systems. These systems are expected to be adaptable, dynamic, reliable as well as affordable. The proposed anomaly and activities recognition system automatically detects anomaly events and notifies with a tag. The action recognition system implemented classifies the

actions of a person in the video with an accuracy of Top 1% accuracy of 71% and Top 5% accuracy of 94%. Also, the proposed algorithm is evaluated in terms of accuracy with three datasets: Avenue, UCSD and UMN, it is observed that proposed approach performs better in Avenue dataset as compare to other two.

5) In paper [5], Authors proposed a deep learning approach to detect real world anomalies in surveillance videos. Due to the complexity of these realistic anomalies, using only normal data alone may not be optimal for anomaly detection. Authors attempt to exploit both normal and anomalous videos. To avoid labor-intensive temporal annotations of anomalous segments in training videos, we learn a general model of anomaly detection using deep MIL framework with weakly labeled data. To validate the proposed approach, a new large-scale anomaly dataset consisting of a variety of real world anomalies is introduced. The experimental results on this dataset show that our proposed anomaly detection approach performs significantly better than baseline methods. Furthermore, we demonstrate the usefulness of our dataset for the task of anomalous activity recognition.

III. PROPOSED SYSTEM

In low resolution cameras, the equality of images will be less and hence tracking of objects in those images and further event detection becomes very difficult as there a loss of visual discriminative detail. To enhance the quality of these images, super resolution techniques can be used. But, enhancing the quality of images using super resolution methods require higher computaional cost and processing capabilities. There is also a need to examine the relevant physical information properties. The main objective of the proposed system is to provide a new technique for tracking of objects and event detection in the images of low resolution without the usage of any super resolution techniques or classifiers.

IV. METHODS AND IMPLEMENTATION

The system consists of some modules which are discussed in the following steps to interpret the Unusual Object Detection using deep learning method. The working method consists of main stages like:

- Loading the dataset
- Design of convolutional neural network
- Configuration of training
- Training of CNN unusual object detector
- Evaluation of trained detector

These stages and conventional methods are will be discussed in this section.

1. **Region Proposal:** Various recent studies have provided methods to produce categorical independent zone recommendations. These methods have examples such as the object ness of image windows, selective Search for Object Recognition, category independent object proposals, object segmentation using constrained parametric min-cuts, Multi scale combinatorial grouping and so on. These methods establish cells by implementing convolutional neural network with square cuts.

2. **CNN (Convolutional Neural Network) for Feature extraction:** In this study, a feature vector of size 4096 were extracted from each region proposal with Caffe deep learning framework. Features were calculated by forwarding the average output 227x227 red-green blue image with five convolution layers and two completely connected layers. In order to calculate an attribute in a region proposal, the image data is first converted to a form compatible with CNN

[9]. (In this study, fixed entrances of 227 * 227 pixels in size are used.). Then, the most simple of the possible transformations of the random-shaped regions was selected. Here, all the pixels in a tight bounding box around the candidate area are resolved unto the required size, regardless of the size or aspect ratio. Before dissolving, the tight bounding box was expanded to provide w pixels skewed picture content around the box at the skewed dimension (w = 16 was used). In addition, a simple bounding box regression was used to expand the localization performance within the application.

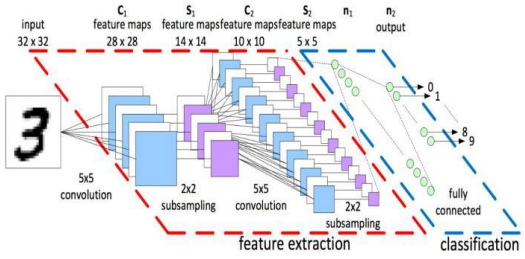

Figure 1. CNN Architecture

3. **CNN - Training (Pre-Processing):** In pre-processing the CNN was training on a large auxiliary data set (unusual objects classification) using only image-pixels level additional tags. CNN was trained on data set (keras layers to unusual objects) using only additional tags. This training was carried out using Deep Learning frame work (Image-net).
4. **Object Category Classifiers:** Here, binary classifier training was used to perceive cars. It is a positive example of an image area in which a car is tightly enclosed. In a similar way, a background region that is not interested in unusual object is a negative example. It is unclear how a partially overlapping region of the unusual object should be labeled. The unclear state is solved by specifying an overlap threshold value. Areas below this threshold value are identified as negative and those above the threshold value as positive. The overlap threshold "0.3" was chosen by conducting a grid search on the verification set. Once the features are removed and the training tags are applied, CNN is applied optimally to all classes.

A. CNN Algorithm Overview

Convolutional Neural Network (CNN) were used to achieve some breakthrough results and win well-known contests. The application of convolutional layers consists in convolving a signal or an image with kernels to obtain feature maps. So, a unit in a feature map is connected to the previous layer through the weights of the kernels. The weights of the kernels are adapted during the training phase by back propagation, in order to enhance certain characteristics of the input. Since the kernels are shared among all units of the same feature maps, convolutional layers have fewer weights to train than dense FC layers, making CNN easier to train and less prone to overfitting. Moreover, since the same kernel is convolved over all the image, the same feature is detected independently of the locating—translation invariance. By using kernels[11], information of the neighborhood is taken into account, which is an useful source of context information. Usually, a non-linear activation function is applied on the output of each neural unit. If we stack several convolutional layers, the extracted features become more abstract with the increasing depth. The first layers enhance features such as edges, which are aggregated in the following layers as modifies, parts, or objects.

The following concepts are important in the context of CNN:

i. **Initialization:** It is important to achieve convergence. We use the Xavier initialization. With this, the activations and the gradients are maintained in controlled levels, otherwise back-propagated gradients could vanish or explode.

ii. **Activation Function:** It is responsible for non-

linearly transforming the data. Rectifier linear units (ReLU), defined as

$$f(x)= \max(0, x),$$

were found to achieve better results than the more classical sigmoid, or hyperbolic tangent functions, and speed up training. However, imposing a constant 0 can impair the gradient flowing and consequent adjustment of the weights. We cope with these limitations using a variant called leaky rectifier linear unit (LReLU) that introduces a small slope on the negative part of the function. This function is defined as

$$f(x)= \max(0, x)+\alpha\min(0,x)$$

where is the leakyness parameter. In the last FC layer, we use softmax.

iii. **Pooling:** It combines spatially nearby features in the feature maps. This combination of possibly redundant features makes the representation more compact and invariant to small image changes, such as insignificant details; it also decreases the computational load of the next stages. To join features it is more common to use max-pooling or average-pooling.

iv. **Regularization:** It is used to reduce overfitting. We use Dropout in the FC layers. In each training step, it removes nodes from the network with probability. In this way, it forces all nodes of the FC layers to learn better representations of the data, preventing nodes from co-adapting to each other. At test time, all nodes are used. Dropout can be seen as an ensemble of different networks and a form of bagging, since each network is trained with a portion of the training data.

v. **Data Augmentation:** It can be used to increase the size of training sets and reduce overfitting. Since the class of the patch is obtained by the central voxel, we restricted the data augmentation to rotating operations.

B. Network Architecture

i. **ImageInput Layer:** An imageInput Layer is the place you initialize the size of input image, here, 128-by-128-by-1 is used. These numbers represent height, width, and the number of channels. In this case, input data is a grayscale image, hence the number of channel is 1.

ii. **Convolutional Layer:** Input arguments for this layer are filtering size, the number of filters, and padding. Here, the filter of size 10 is used, which determines 10×10 filter. The number of channels used is 10, means 10 neurons are connected. Padding of 1 specifies that the size of the output image is same as that of an input image.

iii. **ReLU Layer:** ReLU (rectified linear unit) layer is a batch normalization layer, which is placed after initializing a nonlinear activation function. Importance of this layer is to decrease the sensitivity and increase the pace of the training.

iv. **Max Pooling Layer:** Max pooling layer is one of the down sampling technique which is used for convolutional layers. In this architecture, poolSize is set to 3 and training function's step size is 3.

v. **Fully Connected Layer:** Fully connected layers follow max pooling layer. In this layer, all the neurons of all layers are interconnected to the previous layer. The given input argument for this layer is 10, which indicate 10 classes.

vi. **Softmax Layer:** Fully connected layers are followed by softmax layer, which is normalization technique. This layer generates positive numbers as output such that the sum of numbers is one. Classification layer uses these numbers for classification.

vii. **Classification Layer:** Classification layer is the final layer of the architecture. This layer classifies the classes based on probabilities obtained from softmax layer and also calculate cost function.

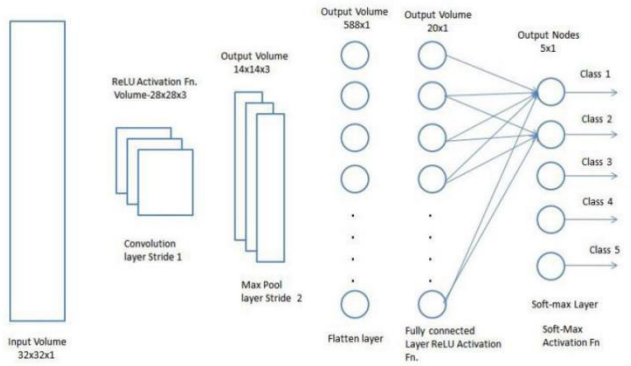

Figure 2. Fully Connected Model

Now that we have converted our input image into a suitable form, we shall flatten the image into a column vector. The flattened output is fed to a feed-forward neural network and backpropagation applied to every iteration of training. Over a series of epochs, the model is able to distinguish between dominating and certain low-level features in images and classify them using the Softmax Classification technique.

So now we have all the pieces required to build a CNN. Convolution, ReLU and Pooling. The output of max pooling is fed into the classifier which is usually a multi-layer perceptron layer. Usually in CNNs these layers are used more than once i.e. Convolution -> ReLU -> Max-Pool -> Convolution -> ReLU -> Max-Pool and so on.

Training Options: The maximum number of epochs set to 50 and initial learning rate is 0.001.

V. SYSTEM DESIGN

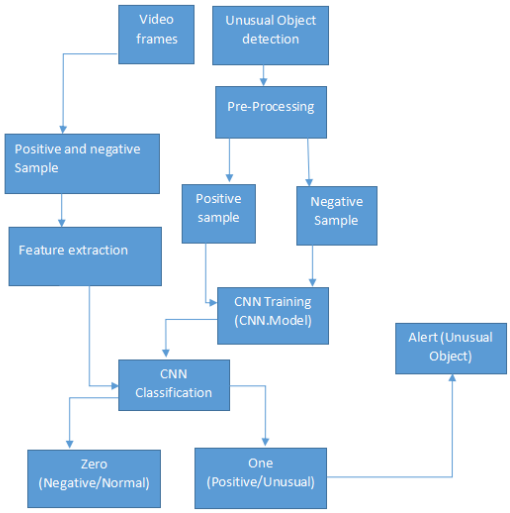

Figure 3. System design of Automatic Detection of Violent Incidents from Video Footage of CCTV Cameras

Unusual Object Detection

The proposed unusual object detector has been successfully trained by using CNN deep learning methods on the sample unusual object datasets and the unusual object detection process has been successfully performed by the trained unusual object detector being tested on the test data set. Different images and videos were tested and found that the new technique of classification was found to show above 90% accuracy. Some images, videos tested with other database images and videos are given in the results analysis. In result analysis are real time detect in types of unusual object with functions like Gun, Knife, Normal image and unusual object recognize when live camera is start then capture the test videos (unusual objects)that time compare the training model **'activity.model** ' and **'lb.pickle** ' if it is match the dataset after the process in display the result.

Figure 4. Detection of unusual object Knife

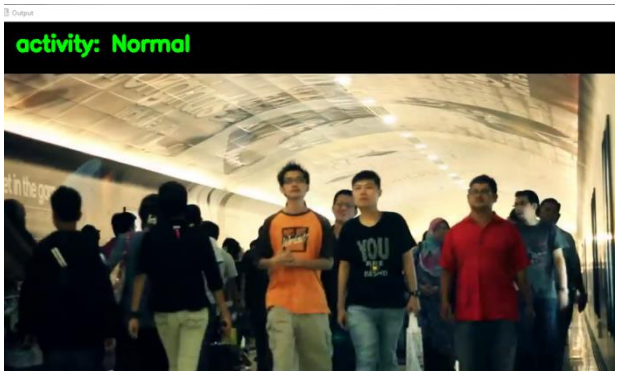

Figure 5. Detection of Normal

Figure 6. Detection of unusual object stick

VI. PERFORMANCE COMPARISON & RESULTS

Table 1. Performance Comparison

Algorithms	Accuracy(%)	Time(ms)
SVM	40.00	2325
LBP	46.00	3210
HOG	38.00	3300
CNN	99.00	1500

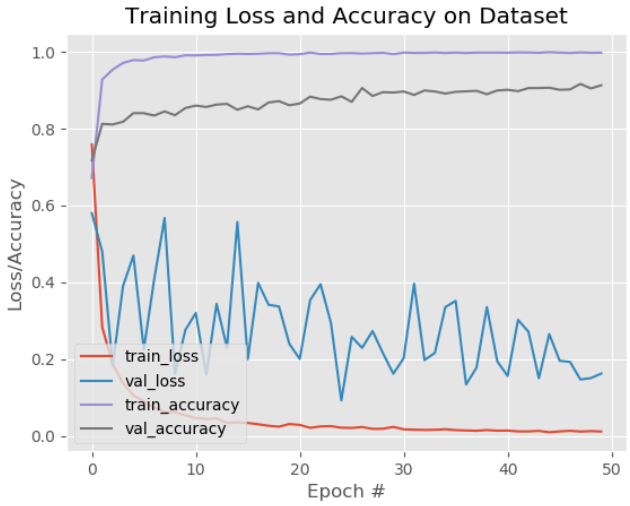

Figure 7. Accuracy graph

From the above performance table and the accuracy graph, compare to the SVM, LBP, HOG, CNN gives the best accurate results for the unsual object detection.

VII. CONCLUSION

The growing demand for a secure and safe environment has enhanced the research for developing smart automated surveillance systems. These systems are expected to be adaptable, dynamic, reliable as well as affordable. The proposed anomaly and activities recognition system automatically detects anomaly events and notifies with a tag. The action recognition system implemented classifies the actions of a person in the video with an accuracy of more than 90%. Also, the proposed algorithm is evaluated in terms of accuracy with trained model.

VIII. REFERENCES

1. Kelathodi Kumaran Santhosh, Debi Prosad Dogra, Partha Pratim Roy "Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey", Jan 2019.
2. Yong Shean Chong, Yong Haur Tay, "Abnormal

Event Detection in videos using Spatiotemporal Autoencoder" Jan 2017.

3. Samir Bouindour, Hichem Snoussi, Mohamad Mazen Hittawe, Nacef Tazi and Tian Wang, "An On-Line and Adaptive Method for Detecting Abnormal Events in Videos Using Spatio-Temporal ConvNet" Feb 2019.
4. Aiswarya Mohan, Meghavi Choksi, Mukesh A Zaveri "Anomaly and Activity Recognition Using Machine Learning Approach for Video Based Surveillance", July 2019.
5. Waqas Sultani, Chen Chen, Mubarak Shah, "Real-world Anomaly Detection in Surveillance Videos" Conference on Computer Vision and Pattern Recognition (2018).
6. Somesh balani, Baswaraju Swathi, Niza Barun Shretha, "Survey on home security surveillance system based on wifi connectivity using Raspberry Pi and IOT module", International Journal of Advanced Research in Computer Science 9 (2), 452, 2018.
7. H Kaushik, B Mounica, B Swathi, "A survey on monitoring systems", International Journal of Computer Science and Mobile Computing, 2015.
8. V. Bastani, L. Marcenaro, and C. Regazzoni. A particle filter based sequential trajectory classifier for behavior analysis in video surveillance. In ICIP, 2015.
9. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W., Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 802–810. NIPS 2015. MIT Press, Cambridge, MA, USA (2015).
10. Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, "Angry Crowds: Detecting Violent Events in Videos", Conference: European Conference on Computer Vision.(2016).
11. www.tensorflow.org
12. Darpan Majumder & Dr. S. Mohan Kumar , Review of Security Strategies used in Vehicular Adhoc Networks, International Conference on Innovative Research in Engineering, Management and Sciences ISBN : 978-93-5391-778-4.
13. Dr. S. Mohan Kumar, Ashika.A, A Survey on Big Data Analysis, Approaches and its Applications in the real World, Journal of Emerging Technologies and Innovative Research, ISSN: 2349-5162, May 2018 , Volume 5, Issue 5, pp. no.: 93-100

Cite this article as :

Baswaraju Swathi, B L Deepika Chowdary, K Sai Sindhu, Ashika P, "Automatic Detection of Violent Incidents from Video Footage of CCTV Cameras", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.464-472, May-June-2020. Available at doi : https://doi.org/10.32628/CSEIT206355
Journal URL : http://ijsrcseit.com/CSEIT206355