International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 27–34

p-ISSN 2655-8564, e-ISSN 2685-9432

Human Detection in Video Surveillance

Sushama Khanvilkar[1], Santosh Gupta[1], Hinal Rane[1],
Calvin Galbaw[1,*]

[1]*Department of Computer Engineering,*
Xavier Institute of Engineering, Mahim, Mumbai, Maharashtra, India
[*]*Corresponding Author: calving2012@gmail.com*

(Received 17-07-2020; Revised 31-07-2020; Accepted 01-08-2020)

Abstract

Recognition of the human activities in videos has gathered numerous demands in various applications of computer vision like Ambient Assisted Living, intelligent surveillance, Human-Computer interaction. One of the most pioneering techniques for Human Detection in Video Surveillance based on deep learning and this project mainly focuses on various approaches based on that. This paper provides an idea of solution to use video surveillance more effectively, by detecting any humans present and notifying the concerned people. The deep learning model, preferred for fast computation, Convolution Neural Network is used by stacking 3 blocks of layers on fully connected layers. This provided an identification of humans and naïve approach to eliminate inanimate human like objects such as mannequins.

Keywords: deep learning, CNN, human detection

1 Introduction

Human Activity detection is a major problem in smart videos surveillance. It is an elementary drawback in computer vision, i.e. to notice the activity of human in

Volume 3, Issue 1, pages 27–34
p-ISSN 2655-8564, e-ISSN 2685-9432

surveillance videos. These applicants need real time detection performance, but it is generally very time consuming to detect the actual activity. Since the use of CCTV, the cases of forced entries and robberies have decreased drastically. But the delay in response to such cases can cause problems. If the owner can get the notification of such events, the culprit can be caught red handed. It becomes important to alert the user by detecting what activity is been performed by the subjects [1-3].

2 Research Methodology

This prospective implementation was carried out using simple programming tools and cloud resources. The Convolutional Neural Network (CNN) is the most promising network to work with images and videos. Hence, developing an architecture using CNN was an optimal and efficient choice.

Implementation Design. In order to implement the system Modified AlexNet design which is trained on frames of video has been used.

Dataset size. 8 videos have been used as a dataset.

Sample size calculation. The sample size was chosen from multiple videos which satisfies the needs of the required datasets. Each video chosen have average of 8000 frames from which about 10% are taken into consideration. This is to reduce redundancy of the data.

Subjects and selection method. The dataset is formed of videos which are taken using CCTV cameras. These videos all include people trying to break into the shops and houses. Some videos also include mannequins and are taken mostly at night. The dataset are labeled according to visibility of humanity.

International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 27–34

p-ISSN 2655-8564, e-ISSN 2685-9432

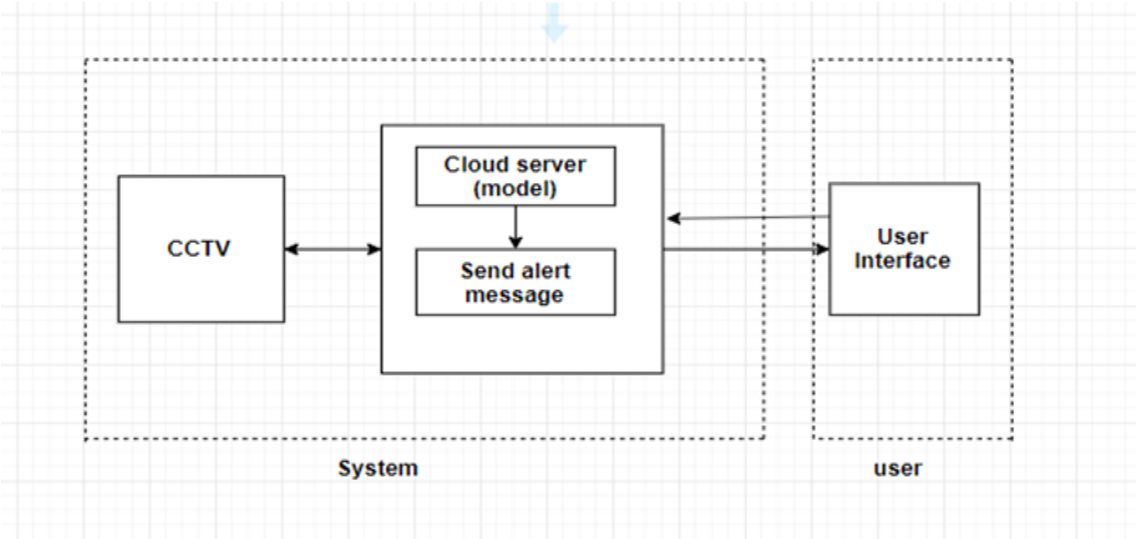

Figure 1. Block diagram

Preprocessing methodology. The main source is a raw video recorded by the CCTV, as in Figure 1. Such videos have very high fps and using such videos requires a lot of processing power by the system. To reduce the processing power, we reduce the fps i.e. dividing current fps by 10. These frames are further sent to be processed [4].

The frames extracted from the videos are RGB images. Processing of images begins with resizing the image into 227×227 and then converting RGB images into Grayscale images. This method converts or compresses the three channels of RGB to a single channel. This single channel contains the values of luminance. Luminance can also be described as brightness or intensity, which can be measured on a scale from black (zero intensity) to white (full intensity). Therefore, the output will have the monochromatic range of black and white.

Most of the theft and break-ins occur at night; hence the images will be dark and will be not clear. To brighten up the image techniques like histogram equalization, alpha and beta transformation can be used. We choose histogram equalization to brighten up the images. Histogram equalization improves the contrast of the image by spreading out the most frequent intensity values.

To remove the noise from the images, blurs are used. This reduces the sharpness of the image and smoothen it. Gaussian blur is the most popular blur and is used for processing. Blur also helps in detection of edges and for thresholding. Thresholding

Volume 3, Issue 1, pages 27–34

p-ISSN 2655-8564, e-ISSN 2685-9432

converts the image to have only two intensities or values. Thresholding using OTSU is used in the project. Finding edges using Canny edge detection is the last pre-processing of the image. Edges help reduce the processing done by the neural model. It reveals the important parts of the image discarding others and helps in extraction of features by CNN.

Neural Network Model. Human detection module will make use of a convolutional neural network to detect and recognize human in the video surveillance. For creating the network, CNN are regularized versions of multilayer perceptron.

Multilayer perceptron's usually means fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. CNN uses relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand engineered. The structure of network consists of different components: Input layer, Hidden layer and Output layer. Input Layer is reflecting the potential descriptive factors that may help in prediction. Hidden Layer is defined number of layers with a specified number of neurons in each layer. Output Layer is reflecting the thing is a human present or not.

The CNN architecture used is a modified AlexNet. The input is a series of 3 continuous frames to help whether the entity is a human or a human like mannequin. Due to this, each frame in the input stack is has its own CNN layers. The features extracted or output of the CNN layers are concatenated and given to the fully connected network. The classification of the images is done by using the softmax activation layer. Figure 2 depicts the CNN block for each frame [5], [6]. The fully connected network is illustrated in Figure 3.

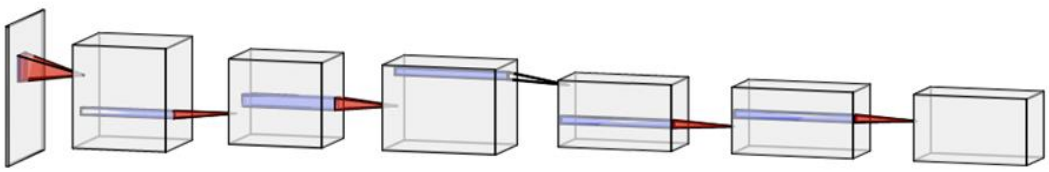

Figure 2. CNN block

International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 27–34

p-ISSN 2655-8564, e-ISSN 2685-9432

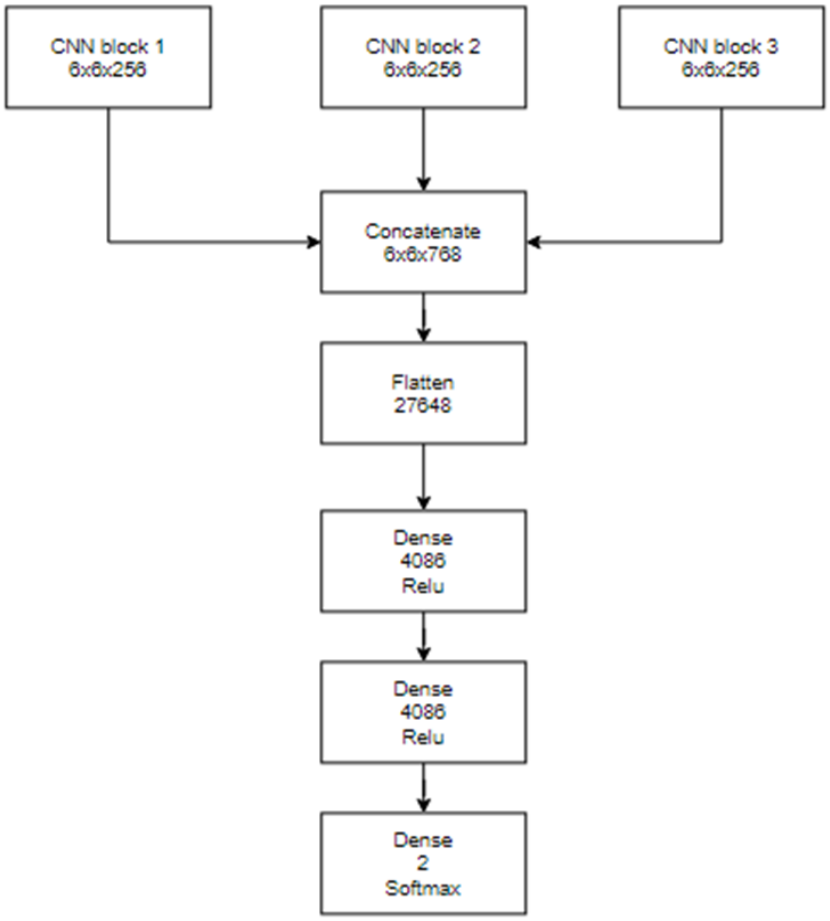

Figure 3. Fully connected network

3 Results and Discussion

The model classifies the data properly at the accuracy rate of 87%. This accuracy is measured by feeding the test data containing both positive and negative labelled images. From the predicted labels, the number of correctly labelled data, positive and negative both, is divided by the total number of data gives the accuracy of the model. The model is trained also in the way that it does not detect mannequins. The model implementation uses android GUI to alert the user of the CCTV and system. This will help the damage done due to the robbery or catch the intruder.

The model is fast and efficient but the delay due to cloud and pre-processing hamper the performance a little bit. This can be neglected by using faster network speed and faster hardware.

Volume 3, Issue 1, pages 27–34

p-ISSN 2655-8564, e-ISSN 2685-9432

Example. Figure 4 shows the correct prediction on the GUI of the system. This depicts the notification and alert used in the system.

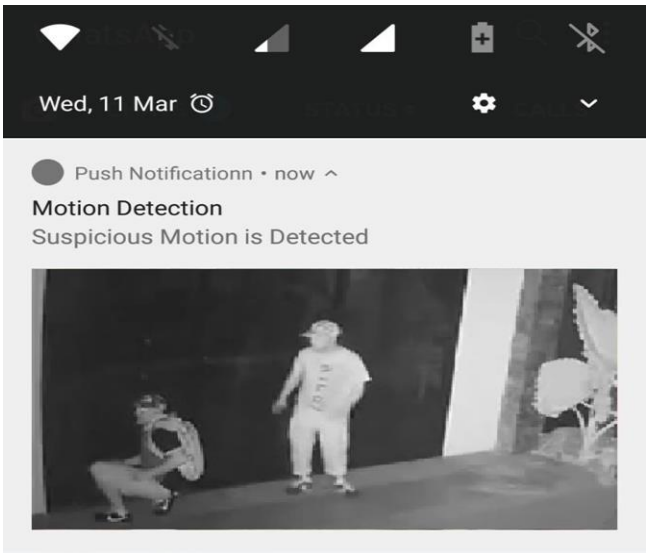

Figure 4. Prediction on Android GUI

Figure 5 shows the incorrect prediction which is also called False Positive. This depicts that the model used is not perfect having accuracy of 87%.

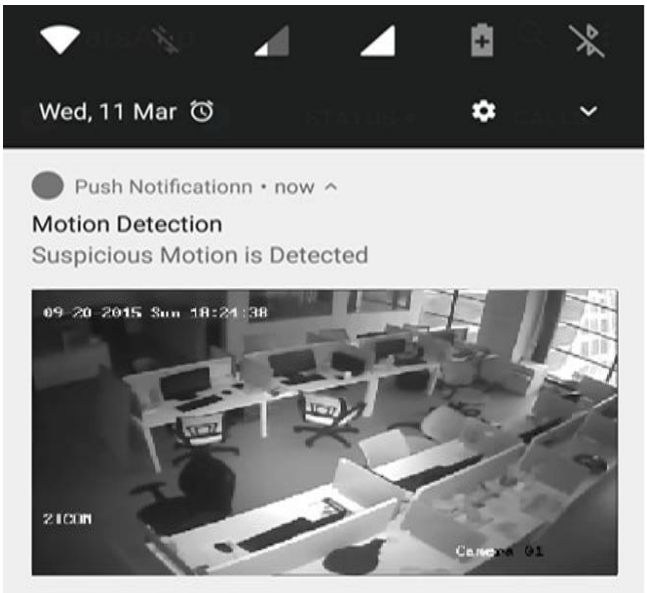

Figure 5. False positive prediction

Volume 3, Issue 1, pages 27–34

p-ISSN 2655-8564, e-ISSN 2685-9432

Discussion. Human Detection in video surveillance using deep learning techniques is the growing area in the field of computer vision. In general, Human Detection is the process of automatically finding the action in the sequence of videos. In this project, we are making use of Convolutional Neural Network. A convolutional neural network (CNN) is an artificial neural network architecture targeted at pattern recognition. In CNN is the methods require labels which are difficult to attain due to the video high dimension information. On this, human activity is detected using this algorithm and after that user can be alert through android application about the subjects.

Environment sensing is the process of detecting a change in the position of an object relative to its surroundings or a change in the surroundings relative to an object. The performance of the system can be enhanced by detecting the changes in its surrounding and it can adapt to the change at the same time. For example, if there is any moment in the shop after closing then the system will alert the user about suspicious activity by send alert message that can user take action on it. Our main goal was to detect human at low visibility due to night time.

The deconstruction of Implementation is as follows:

1. Initially, the input video is taken from video surveillance.
2. This video is processed by the video processing which is used to detect the human activity in the video by frame by frame.
3. The output video is provided to the network to identify the human detection using CNN model.
4. The output of model is sent to user to alert about human activity to take action via application.

4 Conclusion

The accuracy of actually catching a robbery is not calculated in the study but this will reduce the success rate. The purpose of project is to achieve goal to find techniques for behavioural identification. Various techniques for motion recognition based on deep learning. Today, human activity detection in video is the most popular hot space. For security purposes, behaviour recognition can be use in shop or mall. For example, the

use of CCTV, the cases of forced entries and robberies have decreased drastically. But the delay in response to such cases can cause problems. If the owner can get the notification of such events, the culprit can be caught red handed. It becomes important to alert the user by detecting what activity is been performed by the subjects.

Acknowledgements

We (the authors) thank Mrs. Sushama Khanvilkar who gave valuable suggestions and ideas when we were in need of them. She encouraged us to work on this project. We are also grateful to our college for giving us the opportunity to work with them and providing us the necessary resources for the project. Working on this project also helped us to do lots of research and we came to know about so many new things.

References

[1] R. Khurana and A. Kushwaha, "Deep Learning Approaches for Human Activity Recognition in Video Surveillance - A Survey." *First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 542–544, 2018.

[2] A. Khaleghi and M. Moin, "Improved anomaly detection in surveillance videos based on a deep learning method." *8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN)*, 73–81, 2018.

[3] L. Anishchenko, "Machine learning in video surveillance for fall detection." *Ural Symposium on Biomedical Engineering, Radio electronics and Information Technology (USBEREIT)*, 99–102, 2018.

[4] https://www.pyimagesearch.com/2019/07/15/video-classification-with-keras-and-deep-learning/

[5] https://towardsdatascience.com/introduction-to-video-classification-6c6acbc57356

[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5469670/