Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 5, May 2020, pg.51 – 59

SUSPICIOUS ACTION AND BEHAVIOR DETECTION USING CNN

B. U. Anu Barathi
Assistant Professor (SG)
Department of Computer Science & Engineering
Rajalakshmi Engineering College Thandalam, Chennai

P. Aadesh; R. Balajee; M. Balaji
UG Students
Department of Computer Science & Engineering
Rajalakshmi Engineering College Thandalam, Chennai

Abstract— Surveillance is the monitoring of behaviour, activities to manage or direct. This includes observation from a distance with the help of electronic equipment, such as closed-circuit television (CCTV). Most of the surveillance systems are still under human surveillance. Violence within the premises could degrade the decorum of the institution and need to be addressed with at most sincerity. Luckily, the recently emerging AI technique can automatically detect the anomalies. Such anomaly detection is fast and can be further used as the pre-processing mechanism to filter out the surveillance videos, and then forward the anomaly videos to perform the examinations by other highly accurate algorithms. One such algorithm is a convolutional neural network (CNN) that takes the input video frames and outputs the features to the Long Short-Term Memory (LSTM) to learn global temporal features and eventually classify the features by fully-connected layers. This network can not only be implemented by the pre-trained models in ImageNet but also have the adaptability to accept variable-length videos.

Keywords— Surveillance, Convolutional neural networks, LSTM, Image processing, Fight

I. INTRODUCTION

Surveillance is the monitoring of behavior, information, and activities to manage or directing. Closed-circuit television, a piece of electronic equipment is the best example of surveil. Many surveillance systems still require human supervision. The main objective of the project is to design a system with human action recognition for real-world surveillance videos. School fight is always a major issue, however, it is infeasible for the staffs to surveil all time. Violent activities such as bullying, school fights or even taking knives to school still persist in the school premises. The existence of violent behavior in a school leads to numerous serious outcomes.

Digital image processing consists of the manipulation of images using digital computers. Its use has been increasing exponentially in the last decades. Its applications range from medicine to entertainment, passing by geological processing and remote sensing. Multimedia systems, one of the pillars of the modern information society, rely heavily on digital image processing. There are two types of methods used for image processing namely, analogue and digital image processing. Analogue image processing can be used for the hard copies like printouts and photographs. Image analysts use various fundamentals of interpretation while using these visual techniques. Digital image processing techniques help in manipulation of the digital images by using computers. The three general phases that all types of data have to undergo while using digital technique are pre-processing, enhancement, display and information extraction.

II. LITERATURE SURVEY

This is an emerging subject in image processing, therefore a lot of papers have been published. In the beginning, violence detection can be done by detecting blood and flame.

Ersin ESEN et al [1] proposed a new method for the task of fight detection in surveillance videos. The proposed method relies on a novel motion feature, namely Motion Co-Occurrence Feature (MCF) that represents the co-occurrence of the magnitude and direction of motion vectors for the task of detecting fights in surveillance videos. At first, it extracts the motion vectors by block matching technique then it estimates direction and magnitude values. Then they obtained high accuracy by combining k-Nearest Neighbour classifier. Their motivation to use MCF for fight detection stems from two important aspects of the feature. Firstly, fight scenes contain relatively irregular motion with respect to normal actions such as walking, running or car movement. This fact constitutes a discriminating characteristic for fight scenes. It also points out that MCF with high frame history is not suitable for real-time applications.

Chunhui Ding et al [2] proposed a novel system to detect violence in video clips by applying 3D Convolutional Neural Networks architecture without using any prior knowledge. They found that they can easily get the spatial information by applying 2D convolution. So the method was extended to 3D, that is, they had also applied convolution on the temporal sequences. The authors compute convolution on a set of video frames, therefore, the motion information can be extracted from the input video. It states that future work is to use 3D Convolutional Neural Networks architecture to detect mid-level concepts, as well as to improve the accuracy.

Alex Krizhevsky et al [4] proposed a system that uses deep convolutional neural network to classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1000 different classes. On the test data, they achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art.

Kaiming He et al [5] presented a learning framework to ease the training of networks that are substantially deeper than those used previously. To improve the performance, pre-trained models such as ResNet were used. These models are trained on more than 1000 classes, which contains 15 million images. It is available as open-source and used to recognize objects in all types of object detection algorithms as a pre-trained model.

Preceding methods[7][8][9] were depending on the traditional way to extract features like spatiotemporal and motion scale-invariant feature transform(MoSIFT) and Space-time Interest points(STIP) which are widely used descriptors for action recognition. Spatiotemporal features can be extracted automatically using Deep learning method. The architecture in [11] consists of a series of convolutional layers followed by max-pooling operations for extracting discriminant features and a convolutional long-short memory(convLSTM) for encoding the frame-level changes, that characterizes violent scenes, existing in the video. The Architecture in [11] achieves 97.1% with a speed of 31 frames/sec, which is the highest accuracy and fastest model speed in the literature for the hockey dataset [10]. MoSIFT encodes and detects model local motion and interest points local appearance. It had achieved 91% accuracy on hockey dataset in [10].

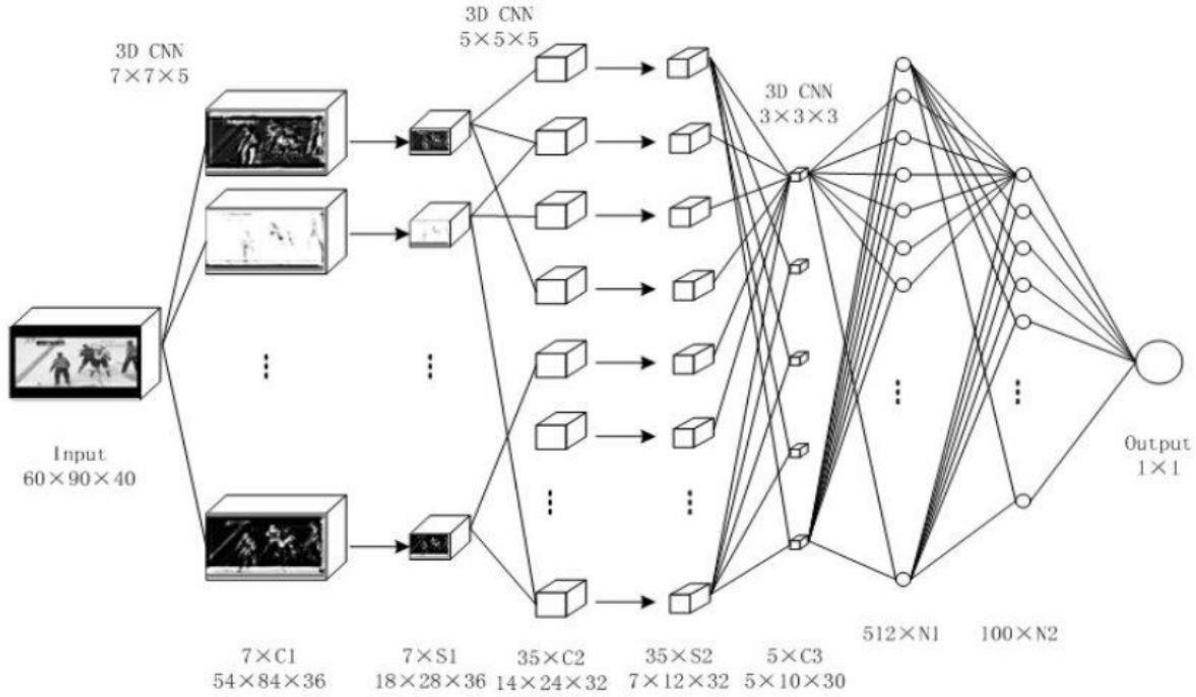

Figure 1 Architecture

III. METHODOLOGY

The proposed model makes use of a pre-trained algorithm called Darknet and it extends the previous work of Violence detection by 3D convolutional networks[2]. This model is majorly based on CNN. Our system comprises of three major processes. Initially, feature extraction is performed followed by the model training process and predicting the presence of violence.

At the beginning of the model, image frames are extracted from the input video stream from CCTV. Now since the image frames are obtained from the video, they are sent to darknet19. Because of its performance and accuracy on Imagenet, darknet19 is used in this system. The Darknet19 contains 19 convolutional layers. In addition to the pre-trained model, residual layers(CNN) are implemented to avoid the problem of degradation. The output frame from the darknet19 is fed to CNN. The learning process of CNN takes place by comparing two consecutive frames that are received from the previous layer. Once the comparison is over the output of CNN is passed to LSTM for learning the global temporal features.

Finally, a conclusion is made based on the output of the LSTM layer. Further, these outputs are categorized into two labels that are indicated by F and NF. If and only if the violence is detected a notification is sent via mobile to the person in charge.

IV. ARCHITECTURE

1. *Data Acquisition*

The proposed method uses three kinds of datasets, Hockey Fight Dataset, Movies Dataset and Violent-Flows.

a). Hockey Fights: During hockey matches, two players often enter into a fight. Thus this Dataset contains an equal number of non-violence and violence actions

b).Movies: This dataset was extracted from movies, General activities are categorized as Non-violence(NO FIGHT), whereas general action activities are categorized as Violence(FIGHT). The dataset consists of 123 violence and 123 non-violence videos. In contrast to the Hockey dataset, this dataset varies profoundly between samples.

c).Violent Flow: This dataset was collected from crowd gathering places which involve a large number of participants in the video. Violent events during football matches occupy a major part of this dataset and it contains 100 videos.

2. *Data Preprocessing*

This process consists of frame difference, dark edge removal, image cropping and image transpose.

Dataset Summary					
Dataset	Description	Total videos	True labels	False labels	total size
Hockey fights	hockey players	1000	500	500	214MB
Violent-Flows	big crowd videos	200	100	100	81 MB
Movies	movies clip	246	123	123	159 MB

Table 1 Datasets

a) Frame difference: This technique is used to find the presence of any difference between the two selected frames and then forwarded only when true or else ignored.

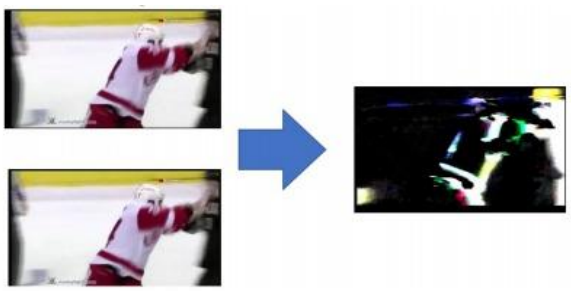

Figure 2 Frame Difference

Figure 3 Edge Removal

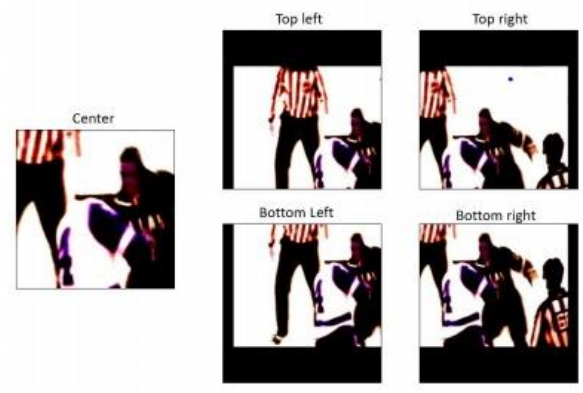

Figure 4 Image Cropping

b) Dark edge removal: Unwanted parts present near the border is removed for better enhancement of training.

c) Image cropping: a slicing of the image, done each time with a different anchor corner.

d) Image transpose: as a complement steps to the cropping process, a transpose was done, this step was done during the fit generator process.

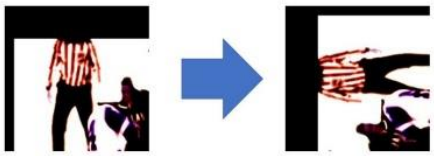

Figure 5 Image Transpose

3. *Data Catalog Creation*

Data catalogs that will tell the data manager where to load the videos, edit the file: tools/Train_Val_Test_spliter.py to specified the path to the dataset videos, the ratio to split the datasets into training, validation and test set. And run such scripts, we will get three data catalogs: train.txt, val.txt, test.txt.

4. *Model training*

In this process it makes use of the training catalogue data file, where we define the video path, starting time and ending time of violent footage portion present in the video. The in-between frames are considered as the frames, that contains violent behavior.

5. *Real-Time Detection*

It detects the presence of violence in the given footage and stores it for future use. And we have also defined the life cycle management policy to delete videos automatically after a particular period of time.

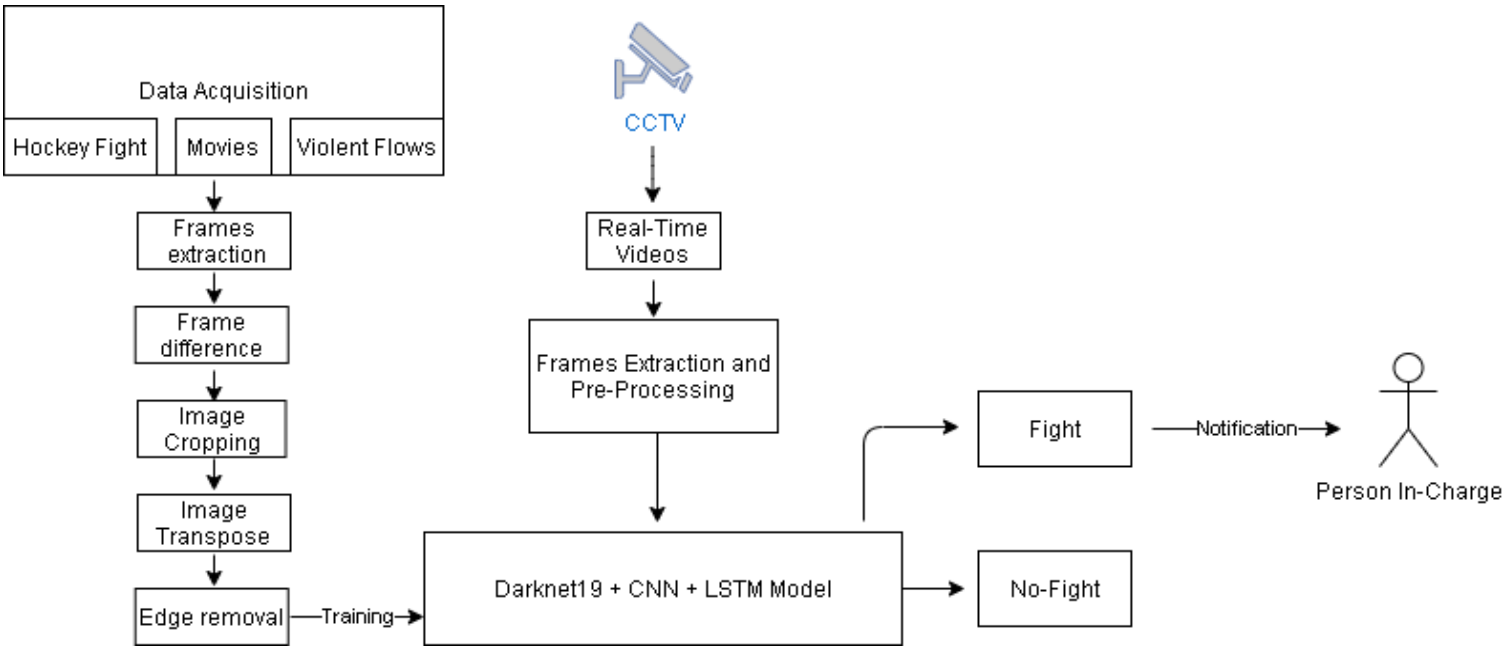

Figure 6 Main Architecture

V. IMPLEMENTATION

1. Convolutional Neural Network

In neural networks, Convolutional neural network (CNNs) is one among the most used categories to do images recognition, images classifications. Objects detections, recognition faces etc., are a number of the areas where CNNs are widely used.

The breakthrough in Computer Vision with Deep Learning is built and perfect over time, mainly by a specific algorithm called the convolutional neural network. The convolutional neural network (CNN) is a deep learning algorithm that can take an input picture, determine the importance of different elements / objects in the image (learnable weight and bias) and distinguish one another. The Pre-processing is required very low compared to other ones on ConvNet Classification algorithms.

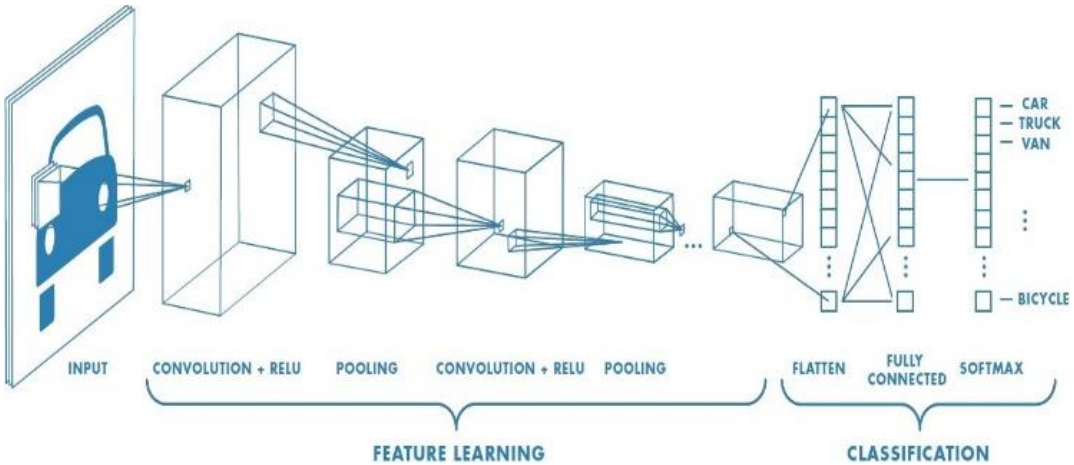

Figure 7 Convolutional Neural Network

Through the application of relevant filters Convolutional neural Network will be able to successfully grasp the Spatial and Temporal dependencies in an image. The architecture is best fit for the image dataset like this because the reduction in the number of parameters involved and reuse weights. In other words, the network can be trained to better understand the sophistication within the image. Convolutional Neural Network basically consist of three layer.

A convolutional layer consists of a set of filters whose parameters must be learned. The height and weight of the filters are smaller than the input volume. Each filter is combined with the input volume to calculate the activation map made with

neurons. In other words, the filter will slip in the width and height of the input, and the dot products between the input and the filter will be calculated at each spatial location. The stacking of activation maps of all filters along with the depth dimension gives the output volume of the convolutional layer.

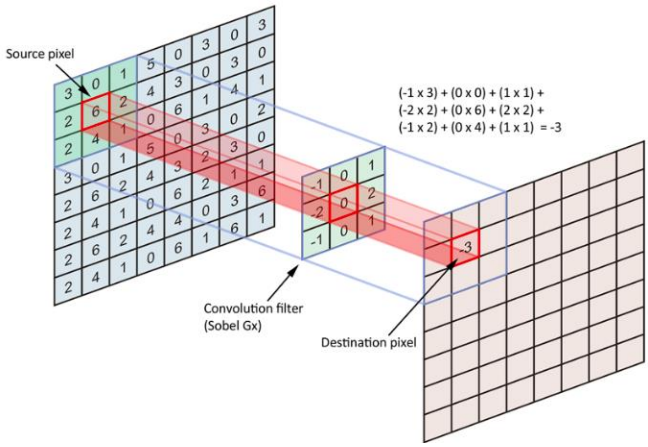

Figure 8 Convolutional Layer

ReLU or Rectified Linear Unit is a non-linear operation. ReLU acts on an elementary level. In other words, it is an operation which is applied per pixel and supersedes all the non-positive values of each pixel in the feature map by zero. It is basically a smooth approximation.

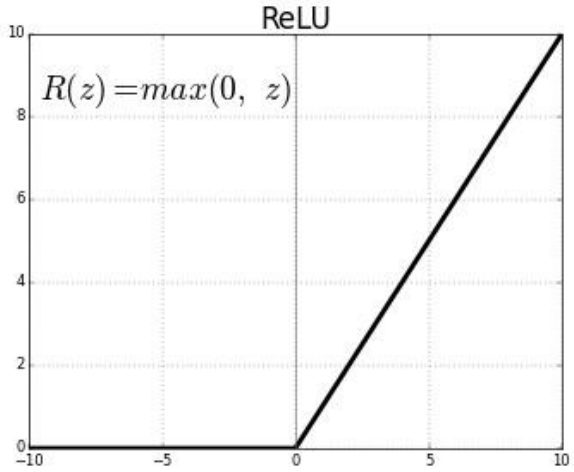

Figure 7 ReLU Function

This layer is periodically inserted in the ConvNets and its main function is to reduce the size of volume which reduces the computation speed of memory and prevents overfitting. Max pooling and Average pooling are two most common types of pooling layers . If we use a max pool with 2 x 2 filters and stride 2, the resultant volume are of dimension 16x16x12.

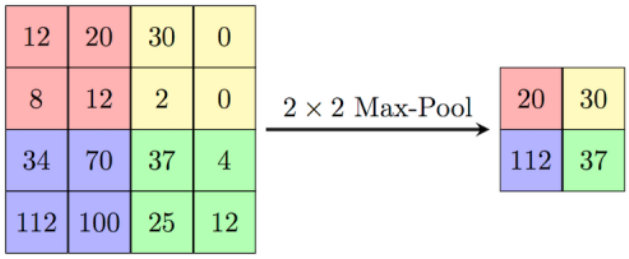

Figure 10 Max Pooling Layer

Softmax layer or fully connected layer is regular neural network layer which takes previous layer's input and calculates the class scores and outputs the 1-D array of size equal to the number of classes

2. DARKNET

DarkNet-19 is a convolutional neural network that's 19 layers deep. Pretrained version of the network trained on quite 1,000,000 images from the ImageNet database can be loaded. The pre-trained network can classify images into 1000 object categories, like keyboard, mouse, pencil, and lots of animals. As a result, the network has learned rich feature representations for a good range of images. The network has an image input size of 256-by-256.

We have used Darknet19 due to its accuracy on ImageNet dataset and it's real-time performance. In order to avoid the degradation problem, we have added additional CNN residual layers.

Large video datasets' lack of richness is an issue. Several studies show that the pre-trained model outperforms very well.

VI. RESULT

On implementation the system without convolutional residual layer the results were not quite accurate. However after adding the residual layers the system outputs were more accurate.

The Presence of Red border indicates Violent activity.

Figure 8 Result-Fight

The Presence of Green Border indicates Non-Violent Activity.

Figure 9 Result-No Fight

The starting learning rate had critical effect on the learning process, the lower starting learning of 0.0001 rate prove to increase the learning of the network compare to 0.001. We assume that the high learning rate cause extreme changes of the network weights and harm it's ability to converge to the right direction, the small learning rate force the network the update it's weights slowly but safely into the right direction of loss.

Unexpectedly, the single-frame model also gives better accurate video-accuracy. However, the per-frame-accuracy of the single-frame model is far less than the network that considers the temporal information. Moreover, the threshold of the amount of the continual positive signals is far larger than the network with the LSTM unit. This is often reasonable since the only frame model doesn't have any temporal information and therefore the only way that decreases the misjudgement is to extend the threshold of the continual positive signals.

The below graph shows the accuracy comparison between the previously developed models.

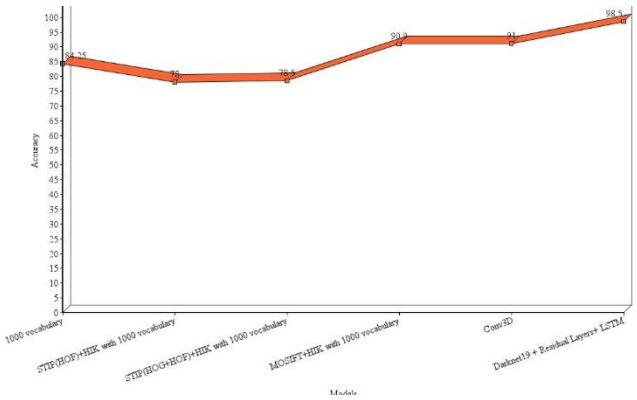

Figure 10 Accuracy Comparison

VII. CONCLUSION & FURTHER WORK

In this work we implemented deep learning model to predict violence in video data, We found our implementation to deal well with this task even though our GPU power was relatively low. The potential of deep learning models is high and can be used easily by law enforcements officers to identifying violence in the streets or in kindergartens. We found the smart data pre-processing of the video's frames play an important factor as well as some of the training parameters such as: CNN network, learning rate and data augmentation.

Looking forward to more complex violence scenarios and appliances it will take researchers to find creative solutions for data collection, advance generalization techniques and real-time optimizations.

In future, additional features could be added to improve its accuracy furthermore. By introducing parallelism processing speed can be improved.

REFERENCES

[1]. Ersin Esen, Mehmet Ali Arabaci and Medeni Soysal Tunitak Uzay, " Fight Detection in Surveillance Videos" in 2013 11th International workshop on content based multimedia indexing (CBMI),Veszprem, Hungary.

[2]. C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia.Violence Detection in Video by Using 3D Convolutional Neural Networks.

[3]. Christian Szegedy Wei Liu Yangqing Jia Pierre Sermanet Scott Reed Dragomir Anguelov Dumitru Erhan Vincent Vanhoucke Andrew Rabinovich *Computer Vision and Pattern Recognition (CVPR)* (2015).

[4]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., pp. 1–9, 2012.

[5]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[6]. L. Zelnik-Manor and M. Irani, "Event-based analysis of video," Proc.IEEE Conf. Comput. Vis. Pattern Recognition, pp. 123 –130, 2001.

[7]. I. S. Gracia, O. D. Suarez, G. B. Garcia, T.-K. Kim, "Fast fight detection". *PloS one*, vol. 10, no. 4, 2015.

[8]. O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim. Fast violence detection in video. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014.

[9]. T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR Workshops*, June 2012.

[10]. E. B. Nievas, O. D. Suarez, G. B. Garc´ıa, and R. Sukthankar. Violence detection in video using computer vision techniques. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 2011.

[11]. S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, pp. 1-6, 2017.

