VIOLENCE DETECTION FROM SURVEILLANCE FEEDS

Namitha Mariyam George, Ruben Roy, Thomas Korah,
Tinil Tom Thomas
UG Students, Dept. of Computer Science and
Engineering
Saintgits College of Engineering
Kottayam, Kerala, India.
namithamg.1@gmail.com, rubenroy.manu@gmail.com,
thomaskorah198@gmail.com, tiniltom98@gmail.com

Aneena Ann Alexander
Asst.Prof. Dept. of Computer Science and Engineering
Saintgits College of Engineering
Kottayam, Kerala, India.
aneena.aa@saintgits.org

***Abstract*— Detecting violence in video footages through automated method is critical for law enforcement and analysis of surveillance cameras with the intention of maintaining public safety. Moreover, it is able to be a notable tool for protecting kids from accessing improper contents and help parents make a better-informed selection about what their kids ought to watch [1]. A convolutional neural network is used to achieve frame level features from a video. The frame level features are then gathered by the use of a variant of the long short-term memory that uses convolutional gates. The convolutional neural network together with the convolutional long short- term memory is able to capturing localized spatiotemporal features which enables the evaluation of local motion taking place in the video.**

In light of this, in this work we will investigate how to depict the idea of violence for a convolutional neural network. Initially by breaking it into more related concepts and objectives such as fights, explosions, blood, etc. and they are combined in a meta-classification to describe violence. We will also describe ways to represent time-based events for the network, since movements are key elements in violence. And finally, we will explore how to localize violent events, since many video streams are a combination of violent and non- violent acts.

***Index Terms*— Computer Vision, Deep Learning, Action Recognition, Violence Detection, Video Surveillance. CNN, LSTM**

I. INTRODUCTION

The usage of surveillance cameras has allowed researchers to analyse a large volume of data to ensure automatic monitoring. An increased security system in smart cities, schools, hospitals, and different surveillance domains is vital for the identification of violent or abnormal activities to avoid any problems that may cause social, economic, and ecological damages. Automatic detection of violence for fast actions is extremely important and might expeditiously assist the involved departments [1].

In terms of development, the technology for detecting objects and movements has gone a long way and allows us to merge these technologies to create a set-up that can identify potential violent activities that occur in our daily lives [2].

That's where our system comes because it proposes utilizing deep learning algorithms to automatically detect violent activity. It involves different process stages such as identification of objects action detection and classification of images. Research has been conducted by researchers in recent years, and many businesses are also looking to build a system that automatically detects violent activity through the videos [3]. There's been a lot of progress too. We are creating a system of technologies that can help us identify violent activity without the assistance of manual detection or a human presence.

We suggest methodologies in this system which will be able to recognize violent threats and events using deep learning methodologies. We used Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) together with Long Short-Term Memory (LSTM) and various methods that made our system validate its techniques for recognizing action.

Our system will be able to seamlessly detect violent activities from video streams or from recorded videos. Firstly, we need to take video inputs and bring these inputs through the Neural Network, and get an output using deep learning methods that tell us whether or not the acts are violent [5].We had to go through several trials and errors because it is still a major setback to identify actions and distinguish between peaceful acts and violent activities. But we've been trying to make that as accurate as possible.

For the aim of evaluation and to foster analysis on violence detection in video we tend to introduce a brand new video database containing around a thousand videos, 2 seconds each into two groups: fights and non-fights that is named the Hockey Fight Detection Dataset. Each 2 sec videos are converted into frames (30 frames per second).We will be using CNN for optical flow extraction and finally results are obtained. Fights can be identified by experimenting on this database.

II. RELATED WORK

Recognizing offensive content draws more attention nowadays given the rapid generation of data on the Internet [6]. Although only a few existing strategies touch on the problem of violence detection in still images ,taking in account of its possible use in violence webpage filtering, on-line public opinion observance and a few alternative aspects, recognizing violence in still images is worth being deeply investigated[7].

Baseline results for violence detection:

To this end, they first established a new database which contains 500 violence images and 1500 non-violence images. And they used the Bag-of-Words (BoW) model which is frequently adopted in image classification domain to discriminate violence images and non-violence images. The Bag-Of Words (BoW) model, is used to be an order-free document representation in Natural Language processing (NLP) and has been widely adopted as a main framework for computer vision tasks such as image classification [12]. The BoW model represents each image through a histogram over a bunch of *visual words* in a visual dictionary (codebook), which corresponds to the number of occurrences of particular image patterns in a given image [12]. While constructing the codebook, the visual words in it are usually defined as the cluster centers generated from the K-means clustering over a pool of low-level feature descriptors such as SIFT. The BoW model is favored by the image classification community due to its simplicity, computational efficiency and robustness to occlusion and within-class variance. The BoW model usually consists of three major procedures, which are feature extraction, feature coding and feature pooling. Among the four basic procedures in the BoW model, we pay the most attention to feature representation so as to assess the effectiveness of different features in classifying violence as non-violence images [12]. Four commonly used features are chosen as a comparison: SIFT, HOG [13], LBP [14] and color histogram.

Violence detection using computer vision:

Even though the detection of fights or generally, aggressive behaviors are less studied. Such actions is also very helpful in some eventualities like in prisons, psych-iatrical or aged centers or perhaps in camera phones. Once considering the previous approaches, we then check the well-known Bag-of-Words framework used for action recognition in fight detection, alongside two of the most effective action descriptors available: STIP and MoSIFT [10].

For the aim of analysis and detection of violence we then introduce a brand new video database which contains about one thousand sequences which is divided into two groups: fights and non-fights [10]. Experiments on this information and another one with fights from action movies show that fights may be detected with close to 90% accuracy

Violence detection for video surveillance system using irregular motion information:

In this proposed violence detection method in surveillance video system. The proposed method consists of 3 steps:

i) Detect the object region within a video image by background subtraction method and then after generating the adaptive background image to detect object region, if the difference value between the input image and the background image is larger than the previously threshold set value, the background image is updated and apply the morphology filter to reduce noise artifacts[8].

ii) In the violence event, the object's movement occurs irregularly. So, make an estimate of motion vector using the Combined Local-Global approach with Total Variation (CLG-TV) in the object region.

iii) Detect the violence event by evaluating the characteristic of motion vectors generated in the region of object by the use of the Motion Co-occurrence Feature (MCF).

The positive side of this method is that the estimated optical flow is used to pull out the MCF feature and detected the violence event using irregular motion information in MCF but the object detection is plausible only in a static environment [8].

Violence detection in crowded scenes using substantial derivative:

This paper presents a unique video descriptor supported substantial derivative, a crucial idea in fluid mechanics that captures the speed of change of a fluid property because it travels through a velocity field. In contrast to normal approaches that solely use temporal motion data, our descriptor exploits the spatio-temporal characteristic of considerable derivative. Above all, the spatial and temporal motion patterns are captured by severally the convective and local accelerations. When estimating the convective and local field from the optic flow, they followed the standard bag- of-word procedure for every motion pattern one by one, and that they concatenated the two ensuing histograms to create the ultimate descriptor. They extensively evaluated the effectiveness of the planned technique on five benchmarks, together with three standard datasets (Violence in Movies, Violence in Crowd, and BEHAVE), and two new video-surveillance sequences downloaded from YouTube [11]. The experiments show how the planned approach sets the new progressive on all benchmarks and the way the structural data captured by convective acceleration is crucial to find violent episodes in crowded eventualities [11].

III. PROPOSED MODEL

1. Model Architecture

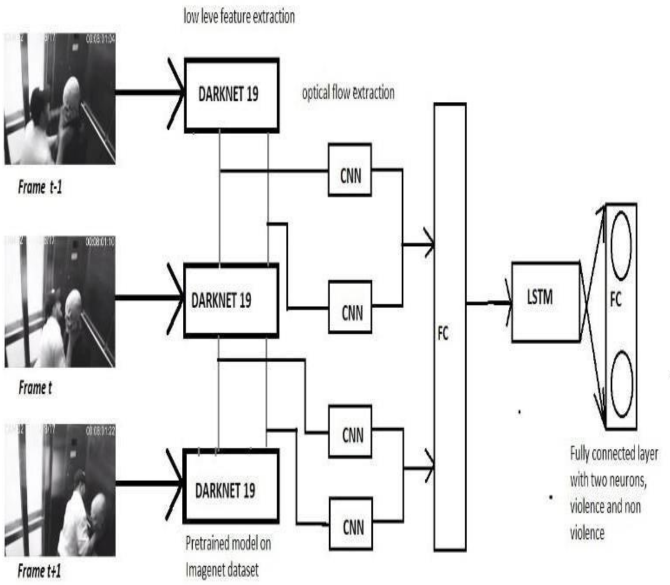
low leve feature extraction
DARKNET 19
optical flow extraction
CNN
CNN
CNN
CNN
DARKNET 19
DARKNET 19
FC
LSTM
FC
Frame t-1
Frame t
Frame t+1
Pretrained model on Imagenet dataset
Fully connected layer with two neurons, violence and non violence

Fig1: Architecture of Proposed Model

Video pre-processing- Dataset contains 1000 videos each of 2 second which is converted into frames (30 frames per second). The extracted frames are then sent to DARKNET19. Basically, the Darknet 19 is used for low-level feature extraction. Output of Darknet is send to convolutional neural network (CNN) for optical flow extraction. These are then send to fully connected layer.

A convolutional neural network (CNN, or ConvNet) is a category of deep neural networks that is most ordinarily applied for analysing visual imaging. They're conjointly called shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture which is employed in the sector of deep learning [1].LSTM unit consists of a cell, an input gate, an output gate and a forget gate. The cell remembers values over absolute time intervals and also the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited for classifying, processing and creating predictions which is supported by time series data, since there will be lags of unknown duration between important events during a time series [4]. LSTMs were developed to manage the exploding and vanishing gradient problems which will be encountered once training traditional RNNs.From LSTM it's once more send to a fully connected layer with two neurons, one for violence and different for nonviolence[8].

2. Algorithms

1. CNN

The **Convolutional Neural Network (CNN)** has been considered one of the most significant categories for image recognition, classification, object detection, face recognition etc. in neural networks [12]. CNN has major features of neural networks that is used rigorously in Computer Vision. Typically, CNN image classification lay hold of an image as an input and then processes it and then classifies the image in categories like Cat, Dog, Bear, and Tiger etc. The feedback the machine gets, it treats the image as a pixel array and it depends entirely on the resolution of the image. It sees input image height, width and dimension**.** Basically, CNN consists of different layers and input images are processed and categorized through these layers. These CNN layers usually consist of core layers, pooling layers, fully-connected layers and layers of normalization [12].

The convolution layer is the first layers that must be passed by the input image, and it is the initial layers that extracts different features from an input image. Using small squares of data, it learns about the picture features [16]. Normally the input image is taken by the convolution layer and a filter is used upon the input image which results in an output image so it take hold of the two images as an input and then produces a third as an input. If we need to describe this mathematically then the layer multiplies the signal with the kernel from the input to obtain a modified signal [11].
Usually the few image matrices that become a matrix known as the Kernel matrix are used for image processing.

The pixels also deal with progresses in CNN. Stride is simply the number of pixels moving over the matrix of inputs. If the stride is 2 then simultaneously the filters switch to 2 pixels [9]. Padding is needed to take the pictures input perfectly and process them perfectly. Over the input image, the filter requires it too perfectly. Then either cut out the image where it wasn't finished by the filter, or pad the image to zero where it wasn't done perfectly by the filter.

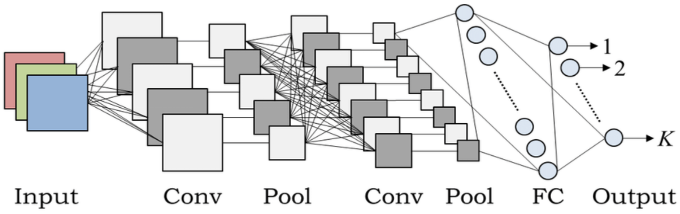

Fig 2: Architecture of Convolutional Neural Network [12]

In addition, pooling is needed which is a sample-based method of decentralization. Pooling reduces the dimensions of the input-images and allows to make assumptions about the features of the input image. There are various kinds of pooling strategies that can be used to make assumptions like Max pooling, Average pooling, and total pooling. Max pooling is the most efficient technique as it takes the largest element [12].

After the image input passes through the layer of pooling, the matrix becomes a vector and then it becomes a fully connected layer like a neural network. The framework then requires pooling to reduce the size of the dimensionality. After reducing the dimensionality, as many convolutional layers must be added until it’s satisfied.

2. RCNN

A **Recurrent Neural Network (RNN)** is a classification of artificial neural networks in which node-to-node connections create a directed-graph along a time continuum. This enables it to show temporary dynamic actions [10]. RNNs obtained from feedforward neural networks those of which are able to utilize their internal state (memory) to process input sequences of variable lengths. It makes them appropriate to activities such as unsegmented recognition, linked handwriting, or speech recognition.

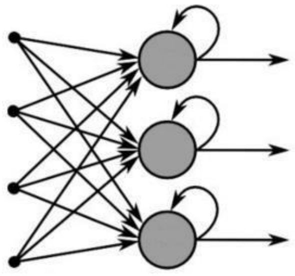

Fig 3: Simple Recurrent Neural network [12]

The word "recurrent neural network" is indiscriminately used to refer to two large groups of networks with a similar general structure, one of which is a finite impulse and the other is an infinite impulsion. All network types exhibit temporal hierarchical behavior [12]. A finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced by a purely feedforward neural network, whereas an infinite impulse recurrent network is a directional cyclic graph that cannot be unrolled.

All finite-impulse and infinite-impulse recurrent networks may have added stored states, and the neural network can regulate the storage directly. If time delays or feedback loops are implemented it is possible to replace the storage with another network or graph. These regulated states are called gated-state or gated-memory, and are part of long-term recurrent memory networks (LSTMs) and gated units. This is also known as the Neural Feedback Network [12].

3. LSTM

Long Short Term Memory Networks (LSTM) are usually a Recurrent Neural Network (RNN) extension, and are totally able of learning long-term dependencies. LSTM network expands the memory of RNN so it can remember things even more than the normal RNN can remember [16]. Each LSTM unit is used as building units for the layers of an RNN which is then also called an LSTM network. LSTM allows RNN to keep track of their inputs for a long time. This is because LSTM's hold on their information in a memory, which is similar to a computer's memory because the LSTM are capable of reading, writing and deleting information from its memory. This memory can be viewed as a gated cell, where gated means the cell decides whether or not to store or delete information (e.g. whether or not it opens the gates), based on the information's importance [13]. Importance assignment occurs through weights that the algorithm also learns. It simply means the information is important and has not been learnt over time.

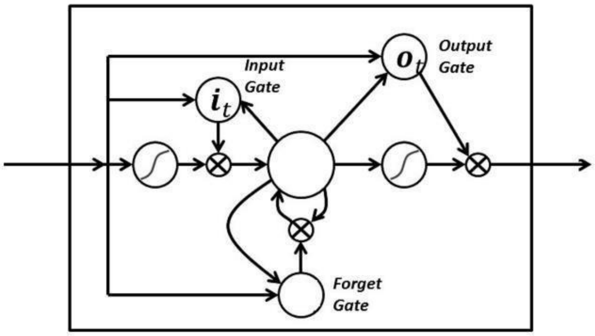

Fig 4: Different gates in LSTM [13]

The three gates in LSTM are: the gate of input, forgetting and output. These gates determine whether to allow new input (input gate) or not, delete the information as it is not significant (forget gate) or let it affect the output at the current time step (output gate). A picture of an RNN with its three gates can be seen above.
Equations for LSTM,
$i_t = \sigma(wi[h_{t-1}; x_t] + b_i)$ (3.1)[13]
$f_t = \sigma(wf\,[h_{t-1}; x_t] + b_f\,)$ (3.2)[13]
$o_t = \sigma(wo[h_{t-1}; x_t] + b_o)$ (3.3)[13]

i_t = represents input gate:
f_t = represents forget gate:
o_t = represents output gate:
σ = represent sigmoid functions
w_x = weight for the respective gate(x) neurons:
h_{t-1} =output of the previous LSTM block (at timestamp t-1):
x_t =input at current timestamp:
b_t =biases for the respective gates(x):

The first equation 3.1 is for the Input gate. This equation tells us the whether to allow a new input or not. The second equation 3.2 is for forget gate which tells us of the information is not important or not. And the third equation 3.3 is for output gate which tells us whether we let the unimportant information affect our output at the current time step [13].

IV. METHODOLOGY

The aim of creating this system is to identify potential threats and provide alerts in likely cases like when fights break out. An applicable scenario for example, would be school fights or bullying. A CNN takes video frames as the input and outputs the features to the Long Short-Term Memory (LSTM) to learn global-temporal features and finally classifies the features by fully-connected layers. This network can be implemented by the pre-trained models in ImageNet. It also has the ability to accept variable length videos.

The proposed model works by taking two video frames as input, the effect of optical should be mimicked. The pre- trained CNN processes the two input frames. The first neural network is a convolutional neural network aimed at extracting high-level image features and reducing input complexity. We are using a pre-trained Darknet model trained on the large visual recognition challenge ImageNet dataset. This is a conventional computer vision task, in which the models attempt to categorize whole images in 1,000 classes like "zebra", " cow" and "mop".[15]

VIOLENCE REPRESENTATION

The best solutions should have very clear and concrete concepts, such as well-defined objects, facial expressions and specific actions. On the other hand, the concept of violence is subjective and complex, posing the challenge of how to reliably represent it in a neural network. Our suggestion is to separate the concept of violence into practical definitions which involves action that cause harm to anything or anyone.

INPUT FRAMES AND CNN

Videos are picture sequences moving at more than 24 fps. In order for a system to identify if a fight is taking place between the people present in the video, it should be able to identify the locations of the people and understand how the movement of the said people is changing over time. Convolutional Neural Networks (CNN) are capable of giving each video frame a good representation. Since we are interested in spatial and temporal dimensional changes, LSTM will be an appropriate option. The LSTM will be able to cipher the spatial and temporal changes. This will give a better representation of the video being analysed.

TRANSFER LEARNING USING DARKNET

It is a conventional computer vision task, in which the models attempt to categorize whole images in 1,000 classes like "zebra", “cow" and "mop"[15]. Contemporary object recognition models have tens and thousands of parameters and therefore it will take weeks to finish their training [15]. Transfer learning is a technology that significantly improves many of this work with a fully trained model for a number of categories such as ImageNet and retraining for new classes from existing weights. There are two pre -trained models, one for images of 224x224 and one for images of 448x448. We chose the 224x224 model. Taking the 224x224 sizes would mean fewer overall parameters and thus less computational power. The final layer of the Darknet is connected to the layers which are trained using the video dataset, the layers of the Darknet are frozen during the training.

WORKING

The system works by inputting a video into it, the video is then divided into frames and analysed on the go, subsequent frames are fed into the model and checked for a binary classification with the trained model. For example, a video of 10 seconds being converted to frames with an interval of 0.2 seconds will generate 50 frames. There are a possible 49 combinations of dual frames (1-2, 2-3,3-4….49-50). These 49 sequences can be considered as 49 cases in which we can it classify it as violent or not and give a classification True/False for violent or nonviolent. In order to reduce the computation power, the videos are resized before testing, also areas in which motion occurs is only considered while checking violence as if there is no motion then there is no activity/violence.

V. RESULT

The proposed methodology uses CNN+LSTM in order to detect violent activities or events from video footages. In order to calculate the performance, we have used two sets of 20 videos violent and no violent. All the videos were consisting of only one class either violence or no violence, not both in the same video. The performance of the system was measured in Precision, Recalland F1 Score.

- Precision = TN / (FP + TN)
- Accuracy = (TP + TN) / (P + N)
- F1 score = 2* (precision*recall)/ (precision + recall)

The obtained values are:

- ★ Precision = 0.9947
- ★ Accuracy = 0.9939
- ★ F1 Score = 0.9943

TP- TRUE POSITIVE
FP-FALSEPOSITIVE
TN- TRUE NEGATIVE
FN- FALSE NEGATIVE
P- POSITIVE
N- NEGATIVE

VI. Conclusion

Violence is a significant threat to personal safety and to community stability in public places. Millions of equipment are currently being deployed in public places [1]. The automated identification of violence from the large quantities of surveillance video data is therefore of great importance. The main aim of the proposed model is to provide a better way to detect violence in the video data. The presented work will explore how to better describe the idea of violence for a classification and it will provide a better definition for violence. As mentioned previously, the aim of creating this system is to detect violent actions from videos while processing it in real-time. Violence detection from video data is a challenging problem because of the identification of complex sequential visual patterns [3]. The proposed system uses pre-trained model on Image-Net (Darknet19). The CNN is able to extract the frame features. In order to learn global-temporal features, the outputs from the CNN are then concatenated and passed to a fully connected layer and the LSTM cell. We also showed that if the network is trained on video frame difference as input, it gives better accuracy [15].
A three-staged end-to-end framework is proposed for violence detection in a surveillance video stream.

- In the first stage, persons are detected using an efficient CNN model to remove unwanted frames, which results in reducing the overall processing time.

- Next, frames sequences with persons are fed into a pretrained Darknet model for feature extraction.

- Next, this is sent to a LSTM-CNN network to classify a video as violence or non-violence.

VII. References

1. Ernesto L. Andrade, Scott Blunsden, and Robert B. Fisher. “Modelling crowd scenes for event detection” volume-1, IEEE, 20-24 Aug. 2006, 18th International Conference on Pattern Recognition.

2. Bruno Malveira Peixoto, Sandra Avila, Zanoni Dias and Anderson Rocha. “Breaking down violence: A deep-learning strategy to model and classify violence in videos”, CEUR-WS, 27-30 Aug 2018, Proceedings of International Conference on Avail-ability, Reliability and Security, Hamburg, GER.

3. Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory”, neural computation, vol. 9, IDSIA, 15 Nov 1997, IDSIA Switzerland.

4. Rupayan Chakraborty, Avinash Kumar Maurya, Meghna Pandharipande, Ehtesham Hassan, Hiranmay Ghosh, and Sunil Kumar Kopparapu. “Affective Impact of Movies and Violent Scene Detection”, volume 1436, TCS-ILAB, 14-15 Sep 2005, Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany.

5. Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Return of the Devil in the Details: Delving Deep into Convolutional Nets”,Computer Vision and Pattern Recognition, BMVC, 14-15 May 2014, BMVA Press Proceedings of the British Machine Vision Conference, London.

6. Ming-yu Chen and Alex Hauptmann, “MoSIFT: Recognizing Human Actions in Surveillance Videos”, CMU-CS-09-161, CMUCS, May 2009, School of Computer Science Carnegie Mellon University Pittsburgh,PA, USA.

7. Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, “Semantic Context Detection Based on Hierarchical Audio Models”ACM, Nov 2003, Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, NY, USA

8. Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang., “Detecting Violent Scenes and Affective Impact in Movies with Deep Learning” Vol -2 MediaEval, 14- 15 Sep 2015, Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany

9. Fillipe Dias Moreira de Souza, Eduardo Valle, Guillermo C´amara Chavez, and Arnaldo de Albuquerque Araujo, “Color-Aware Local Spatiotemporal Features for Action Recognition” CIARP, Nov 2011, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 16th Iberoamerican Congress, Chile

10. Fillipe D.M. De Souza, Guillermo C. Chavez, Eduardo A. do Valle Jr., and Arnaldo de Araujo, “Violence detection in video using spatio-temporal features” IEEE, Aug 2010, 23rd SIBGRAPI: Conference on Graphics, Patterns and Images Gramado, Brazil

11. Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Ionescu Bogdan, Vu Lam Quang and Yu-Gang Jiang, “Violent Scenes Detection ”MediaEval, Oct 2013,MediaEval 2013 Working Notes, Oct 2013

12. W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos”, Computer Vision and Pattern Recognition, IEEE, 31 Mar 2018, IEEE Conference on Computer Vision and Pattern Recognition

13. Wang D, Zhang Z, Wang W, Wang L, and Tan T, “ Baseline Results for Violence Detection in Still Images” IEEE 18 Sep 2012, Ninth International Conference on Advanced Video and Signal-Based Surveillance

14. D. Gordon, A. Farhadi, and D. Fox, \Re ^3: Re al-time recurrent regression networks for visual tracking of generic objects", IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 788{795, 2018.

15. A. J. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, \Bidirectional convolutional lstm for the detection of violence in videos", in ECCV Workshops, 2018.

16. S. Hochreiter and J. Schmidhuber, \Long short-term memory", Neural computation, vol. 9, no. 8, pp. 1735{1780, 1997.