Robust Real-Time Violence Detection in Video Using CNN And LSTM

Al-Maamoon R. Abdali
Computer Sciences Department
University of Technology
Baghdad, Iraq
mamonrasoolabdali@gmail.com

Rana F. Al-Tuma
Computer Sciences Department
University of Technology
Baghdad, Iraq
110016@uotechnology.edu.iq

***Abstract*— Detection of a violence event in surveillance systems is playing a significant role in law enforcement and city safety. The effectiveness of violence event detectors measures by the speed of response and the accuracy and the generality over different kind of video sources with a different format. Several studies worked on the violence detection with focus either on speed or accuracy or both but not taking into account the generality over different kind of video sources. In this paper, we proposed a real-time violence detector based on deep-learning methods. The proposed model consists of CNN as a spatial feature extractor and LSTM as temporal relation learning method with a focus on the three-factor (overall generality - accuracy - fast response time). The suggested model achieved 98% accuracy with speed of 131 frames/sec. Comparison of the accuracy and the speed of the proposed model with previous works illustrated that the proposed model provides the highest accuracy and the fastest speed among all the previous works in the field of violence detection.**

Keywords*— CNN, LSTM, Violence Detection, Smart Cities, Deep Learning

I. INTRODUCTION

One of the most critical goals for Smart cities is law enforcement and city safety; from that goal, came the importance of surveillance systems and violence detection. Since the violence in the city can happen at any time and relying on a human to monitor and detect violence event is not an efficient way to handle such cases. Many studies focus on building an automated way to identify violence in the video correctly and achieved a great result in term of accuracy and response time, the studies [6] [12] shows that deep learning can lead to better accuracy and speed from methods that are relying on handcrafted methods for action recognition in videos.

The most popular datasets for violence detection benchmarking is Hockey Dataset [1] and Movies Dataset [1] and Violent-Flows Crowd Violence Dataset [4]. These three datasets have been used widely in the field of violence detection as a separate dataset for benchmarking, and there is no study have combined them into one dataset to build one robust generic model that can work well with different data sources. In this paper, we wanted to achieve two goals ,the first one is to surpass the previous works highest result using the same benchmarking method in the literature with one of the discussed datasets, and the second goal is to explore the idea of combing the three datasets in one dataset and build a model that can generalize well over different data sources.

Since the highest result achieved for Hockey Dataset is 97.1% [6] and 100% [6] for Movies Dataset [1] we picked the Hockey Fight Dataset [1] to evaluate our base model on it, that because there is no room for improvement can be achieved in the Movies Dataset [1] while in the Hockey Fight Dataset [1] we still have a place for improvement. Also, because the architecture in [6] uses ConvLSTM [13], we decided to explore a different approach, we experimented with Conv3d [12] and (CNN and LSTM). The contributions of this paper could be summarized as follows:

• CNN (Convolutional neural network) followed by LSTM (Long-short term memory) has been proved to be the best architecture when data are small and low computing power available for the task in hand

• A robust, accurate, and fast model to recognize violence in the video using deep learning has been developed.

The remaining parts of the document organized as the following. Section II discusses some of the most common approaches for performing violence event recognition based on deep-learning, section III discusses the proposed model in detail, section IV contains the results and discussion of the experiment and the conclusion in section V.

II. RELATED WORKS

Violence event recognition in video is a problem of spatiotemporal features classification once a model can recognize the spatiotemporal features correctly; it can achieve a good result. Methods for extracting these features are either handcrafted or automated.

Most of the previous works [1] [2] [3] [4] were relying on a handcrafted way to extract spatiotemporal features the most common action descriptors is motion scale-invariant feature transform (MoSIFT) [17] and Space-time interest points (STIP) [18].

MoSIFT uses local appearance and motion to detect distinctive local features so it can encode and detects interest points local appearance and models local motion. The study in [1] Compared between STIP and MoSIFT and was able to achieve an accuracy of 91% on hockey dataset by using MoSIFT. The study in [16] Compared between STIP and

978-1-7281-0761-5/19/$31.00 ©2019 IEEE

SIFT found that STIP performance was much better than SIFT.

Deep learning approach can be used to extract spatio-temporal features automatically. For example, the Architecture in [6] uses deep learning to capture the spatiotemporal features automatically. The architecture in [6] 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 [6] 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 [1].

The most common ways in deep-learning approach to capture and learn spatiotemporal features are:

- CNN and LSTM: - it uses the Convolutional neural network [5] as a spatial features extractor, as suggested in [7] CNN considered the best spatial features extractor that outperformed almost all the kind of handcraft methods for spatial features extraction, then the extracted features feed into LSTM Layer [8] to learn the temporal relation than using any classification layer such as ANN or any other approach for learning and classification. This approach can benefit from transfer learning by using a pre-trained model in the CNN layer such as vgg19 [9], resnet [10], and other pre-trained models to extract the general spatial features. The transfer learning approach [11] is a very effective method to build a model with high accuracy, especially when there Is limited small data.
- Conv3D [12] several studies show the excellent ability for Conv3d to learn spatiotemporal relation, and it was able to outperform the (CNN and LSTM) approach. Conv3D convolved on four dimensions the time(frame) and height and width and colors channel. It is simple, fast, and more straightforward to train then (CNN and LSTM), the study [12] shows that Conv3D with enough data is the best architecture for action recognition.
- Convlstm [13] it extends the LSTM model to have a convolutional structure in both input-to-state and state-to-state transitions. ConvLSTM can capture spatiotemporal correlations consistently. This model shows a promising result, and it has been used in [6] to achieve an accuracy of 97.1% on Hockey Dataset [1].

III. The Architecture Of The Proposed Model

The goal of this study was to build two models:-

- The base model which is evaluated base on hockey dataset [1] and built for comparison with highest accuracy and speed achieved on that dataset.
- The Proposed Model which is trained and evaluated on our proposed approach for combining the three datasets[1][4] of violence detection.

We experimented with Conv3d and (CNN and LSTM), and since we have a small dataset, we need to use transfer learning in order to achieve high accuracy, So we do not have much flexibility to build good architecture for Conv3d, Therefore our base model was built on top of CNN (pre-trained vgg19) as spatial feature extractor followed by LSTM cells. Each item in our base model was with the shape of (40x160x160x3) which correspond to (frame x H x W x RGB colors channels) and since vgg19 work with the 3d shape of input we use (time Distribution techniques).

The Time Distribution operation applies the same operation for each group of tensors. The tensor here represents one frame, in the base model; the group of tensors consists of 40 consecutive frames represented with a shape of [frames, h, w, colors]. Each video (a group of tensors) get into the vgg19 [9] as a frame by frame each with the shape of [h, w, colors] the vg19[9] apply same weight same calculation for that group of tensors the calculations changed once new group received. The output of the time distributed vgg19 [9] is a 2d tensor with a shape of [40x12800] which is feed into the LSTM [8] layer here our LSTM [8] layer is consist of 40 Cell which means we try to learn a time relations between 40-time steps each Cell represent a time step each time step is a frame, and the 40-time step is 40 consecutive frames.

We take the full sequence prediction from the LSTM [8] units and not the last prediction for example in the base model this stage will result in 40x40 tensor as output, and then we apply a neural network layer with 160 neurons to it with time distributed fashion, and then we take global average pooling.

A previous study [14] shows that the global average pooling is an excellent method to achieve a generalized model. With more robust to spatial translations and used as a replacement for the fully connected layer, its output feeds directly into the output layer. The architecture of the base model illustrated in Fig. 1.

In the base model, we feed the output of the global average pooling directly into our final output layer, but in the second model, we found adding a dense layer into the architecture can help us achieve higher results in the combined dataset. We also used Adam optimizer [15] with a learning rate of 0.005, and we monitor the test loss to save only the best model, and we also reduce learning rate by a factor of 0.5 when the test loss is not decreasing. For the combination dataset, we end up with the architecture that is shown in Fig. 2.

In general, the steps we followed are the following:

- Read sequence of frames in 4d tensor (frame, H, W, RGB)
- Apply pre-trained CNN for each frame
- Group the result from the previous step and flatten the tensor to be a 2d shape (frames, SP) where SP is (H*W*RGB) and represent a spatial feature vector for one frame.

- Use the previous step output as feature vector input to LSTM where SP represent input and Frame represent time step ex for 30 frame input we have (SP1, SP2 .. SP30) each goes in a time step of LSTM.
- Take full sequence prediction from LSTM and feed it to a dense layer in a time distributed manner.
- Take the global average of the previous step output to get the result as a 1d tensor.
- Feed the output of the previous step into the output layer (dense layer with sigmoid activation which represents the probability of violence existence in the given video).

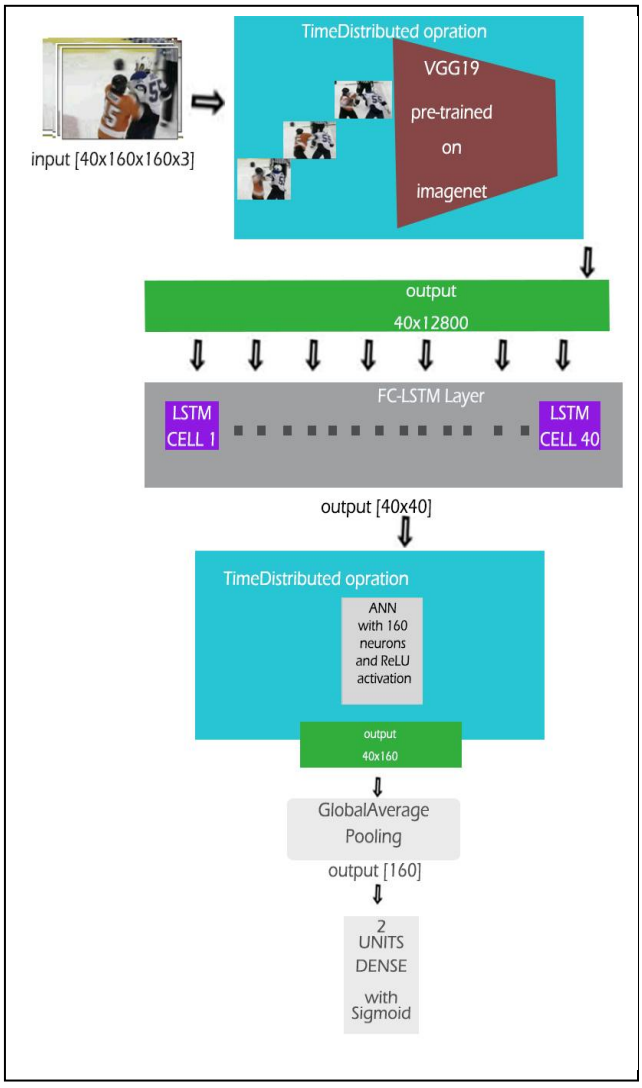

Fig. 1. The architecture of the base Model

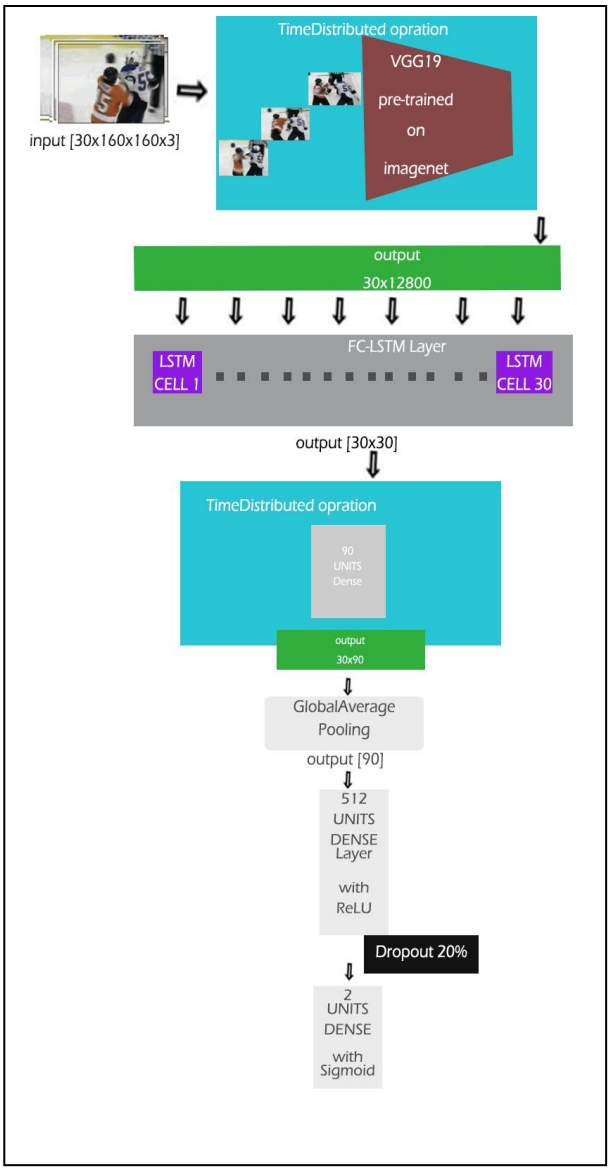

Fig. 2. The architecture of the proposed Model

In the base model, we use 700 videos from the hockey dataset [1] for the training set, and 300 videos as our test set, and the batch size was 20 with a shape of [40,160,160,3] for each video, while in the combined dataset the total training set videos was 896, and for the test set, we have 363 videos with the shape of [30,160,160,3] for each video.

We also take separated videos for each original datasets that do not exist in both training and test set to be used for result analysis and validation. We have used Cross Entropy Loss as our Loss function.

IV. Result And Discussion

After training the base model on the Hockey Fight Dataset [1], we achieved an accuracy of 98 %, as shown in Fig. 3.

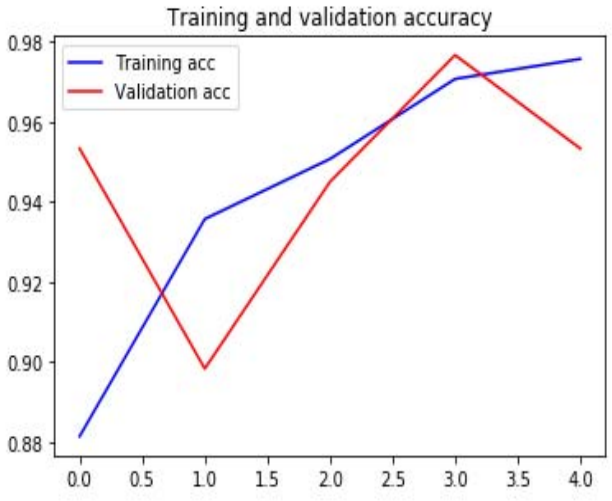

Fig. 3. Test Vs. Training Accuracy Graph For The Base Model

The loss of both train and test sets calculated using Cross Entropy Loss are represented in Fig. 4.

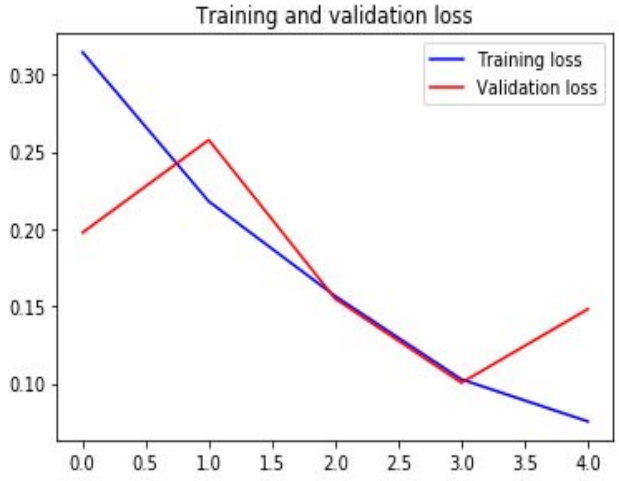

Fig. 4. Test Vs. Training Loss Graph For The Base Model

Both figures indicate that the best training weights that achieve both best accuracy and more generality (no overfitting) was in the third epoch Since we save only the best model base on test loss, the model weights that are saved was for the third epoch which is 98% this accuracy outperformed the highest known accuracy 97.1% [6].

The speed of the base model is 131 frames per second, which is four times faster than the fastest known model [6] in the field of Violence event detection.

In order to build a realistic model that is more robust for real-world cases where data can come from different sources and distributions, we considered using the same architecture of the base model with some tuning and train it on a combination of the three datasets.

After experimenting with the combination dataset The highest result we achieved on the test set for that combination dataset was 94.765%.

The accuracy of the proposed model for each epoch is showing in Fig. 5.

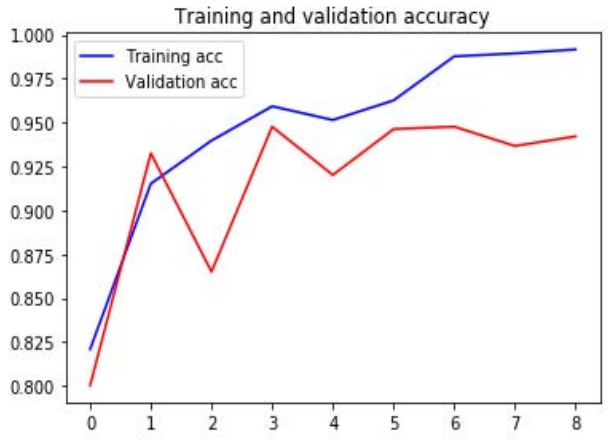

Fig. 5. Test Vs. Training Accuracy Graph For The Proposed Model

Also, the loss (Cross Entropy loss) of both train and test sets are illustrated in Fig. 6.

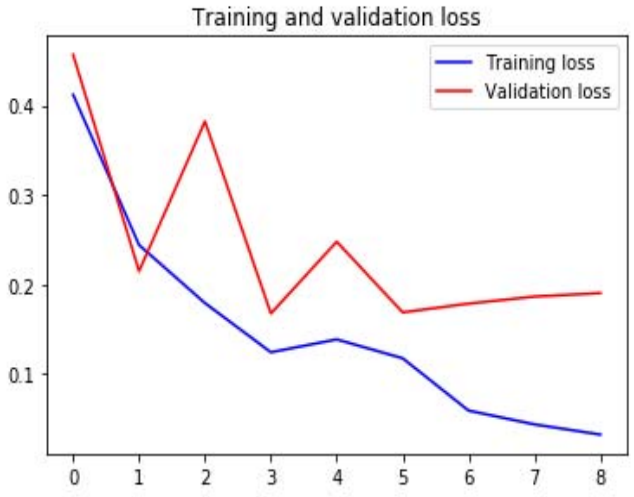

Fig. 6. Test Vs. Training Loss Graph For The Proposed Model

We also created a separated data for validation that contain separate items for each of the three datasets the proposed model achieved 100% on the Movies Dataset [1] validation set and 96.33% on the Hockey Fight Dataset [1] validation set and 85.71% for the Violent-Flows Dataset [4] validation set. After investigating the Violent-Flows Dataset [4], we found that the violent scene was very different from the other two datasets.

The Violent-Flows Dataset [4] contain violence scenes for crowds which in term contain multiple peoples that participants in the violent action and since we have small data for that kind of event in the total dataset the model was able to identify

violence scene between two people more accurate than crowd violence.

We believe if there were more data for such kind of event, the model would perform better in both overall detection accuracy and crowd violence detection. We also validate our model on random YouTube videos with a different format; the model shows 100% accuracy of detection for that random YouTube videos.

The proposed model speed on a NVIDIA GTX1060 laptop GPU was 131 frames per second, which is four times faster than the fastest known model [6] in the field of Violence event detection.

V. CONCLUSION

This work shows that (CNN and LSTM) with use of transfer learning is the best approach to achieve accurate, robust and fast model in the task of violence detection with a limited dataset and computing resources.

The proposed base model is evaluated on a benchmark dataset and resulted in improved performance compared to the highest accuracy achieved on the previous works for that dataset. In addition, the proposed method was faster than the previous works.

We believe that there is still a place for improvement and we suggest for feature work to explore or create a new well-balanced large data set with different video sources for violence detection with more class to detect the violence action itself not just the existence of the violence or not.

REFERENCES

[1] 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

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

[3] 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

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

[5] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition", *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[6] 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.

[7] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* pp. 806-813,2014.

[8] S. Hochreiter, and J. Schmidhuber. "Long short-term memory". *Neural Computation*, 9(8):1735–1780, 1997.

[9] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In *ICLR*, 2015

[10] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition*, pp. 770-778, 2016.

[11] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *in IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri."Learning spatiotemporal features with 3D convolutional networks". *In Proc. ICCV*, 2015.

[13] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, W. C. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", *NIPS*, 2015.

[14] M. Lin, Q. Chen, and S. Yan. "Network in network". *arXiv:1312.4400*,2013.

[15] D. Kingma and J. Ba. "Adam: A method for stochastic optimization".*ICLR*, 2015.

[16] F. D. De Souza, G. C. Chavez, E. A. do Valle Jr, and A. d. A.Araujo. "Violence detection in video using spatio-temporal features". In *Conference on Graphics, Patterns and Images (SIBGRAPI),* 2