National Conference on Sustainable Engineering And Management (NCSEM24)
ISSN 2349-6002

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

Reshma. S[1], Absa S.P[2], Riya S.R[3], Sreena V Wilson[4], Ms. R. Sajitha[5]
[1,2,3,4] *Department of computer science and engineering*
[5]*Assistant professor, Sivaji collage of engineering*

Abstract - A major part of law enforcement and public safety is the detection of violent events in surveillance systems. The speed, accuracy, and adaptability of violenceevent detectors across a range of video sources in various forms provide measures for measuring their success. A number of studies focused on speed, accuracy, or both when detecting violence, but they ignored to account for adaptability across various types of video sources. In this paper, we suggested a deep-learning based real-time violence detector. CNN serves as a geometric feature extractor in the proposed framework, while LSTM is used as a time-based relation learning technique with a focus on the three factors (accuracy, speed, and overall flexibility). The advised model reached 98%.

I.INTODUCTION

One of the most critical goals for Smart cities is law enforcement and citysafety; 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 commonly used datasets for comparing violence detection are the Violent-Flows Public Violence Dataset [4], the Pictures Dataset [1], and the Hockey Dataset [1]. 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 achievedin the Movies Dataset [1] while in the Hockey Fight Dataset [1] we still have a placefor improvement. Also, because the architecture in [6] uses Conv LSTM [13], we decided to explore a different approach, we experimented with Conv3d [12] and (CNN and LSTM). A few words may be used to describe this paper's contributionsinclude:

- CNN (Convolutional neural network) followed by LSTM (Long-short termmemory) 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.

II.RELATED WORK

Video violence event identification is an issue of defining dynamical features; if a model can accurately identify the dynamical features, it will work better.

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 (Mo SIFT) [17] and Space-time interest points (STIP) [18].

Mo SIFT uses local appearance and motion to detect distinctive local features so itcan encode and detects interest points local appearance and model local motion. The study in [1] Compared between STIP and Mo SIFT and was able to achieve anaccuracy of 91% on hockey dataset by using Mo SIFT. The study in [16] Compared between STIP and 2019 2nd Scientific Conference of Computer Sciences (SCCS),University of Technology - Iraq 105 SIFT found that STIP performance was muchbetter than SIFT.

Deep learning approach can be used to extract spatiotemporal features automatically. For example, the Architecture in [6] uses deep learning to capture the spatiotemporal features automatically. A set of layers of convolution, max-pooling operations to create classifier features, and a layer of convolutional long-short memory (conv LSTM) for storing frame-level changes define the architecture of [6], that characterizes violent scenes, existing in the video. For the hockey dataset [1], the architecture in [6] achieves 97.1% accuracy at the rate of 31 frames/sec, thefastest model speed in research.

In a method for deep learning, the most common techniques for collecting and learning spatiotemporal features are:

1.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 ANNs or any otherway of labeling and learning. This approach can benefit from transfer learning by using a pre-trained model in the CNN layer such as vgg19 [9], reset [10], and otherpre-trained models to extract the general spatial features. If there is a shortage of small data, a transfer learning method [11] is a very effective way to develop models with high accuracy.

2.Conv3D [12] Multiple studies show Conv3d's unique ability to learn temporal relations, above the results for both of the CNN and LSTM methods. Conv3D arranged the height, width, colors channel, and time (frame) in four dimensions. 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. Algorithms structure has been added to the LSTM model inConvlstm [13], allow it to be used in both state-to-state and input-to-state transitions.

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 forcomparison 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 colours channels)and since vgg19 work with the 3d shape of input we use (time Distribution techniques).

For all sets of tensors, the Time Supply function performs the same method. For all sets of tensors, the Time Supply function performs the same method. 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, colours]. Eachvideo (a group of tensors) get into the vgg19 [9] as a frame by frame each with the shape of [h, w, colours] the vg19[9] apply same weight same calculation for that group of tensors the calculation for that group of tensors the calculations changed once new group recived the 40-time step is 40 consecutive frames.

We take the full sequence prediction from the LSTM [8] units and not the lastprediction 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. Wealso 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 withthe 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.

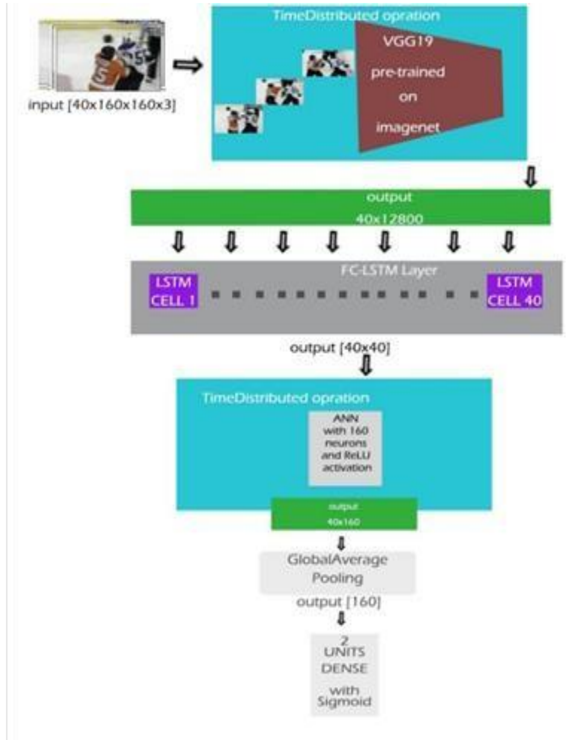

Fig. 1. The architecture of the base model

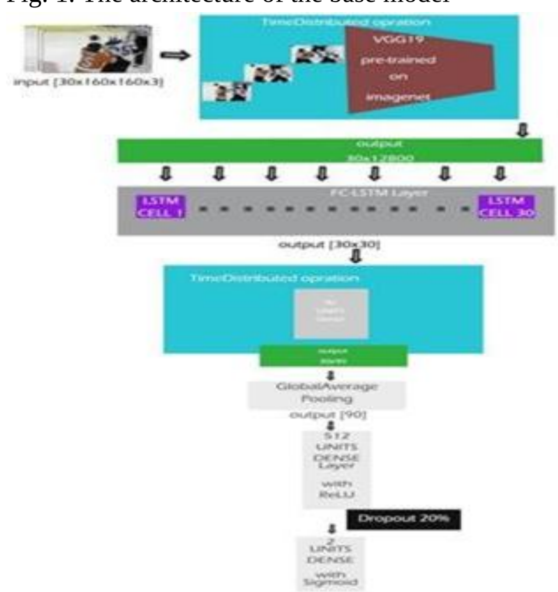

Fig. 2. The architecture of the proposed Model

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 inboth 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], weachieved an accuracy of 98 %, as shown in Fig. 3.

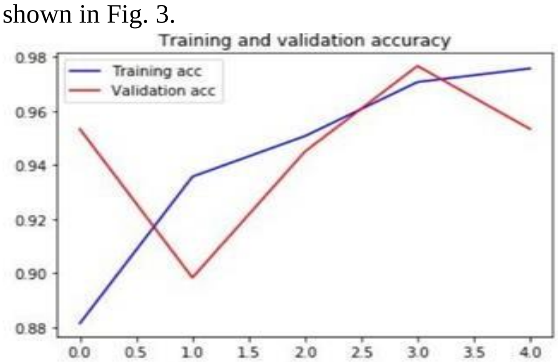

Fig.3. Test Vs. Training Accuracy Graph For The Base Model

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

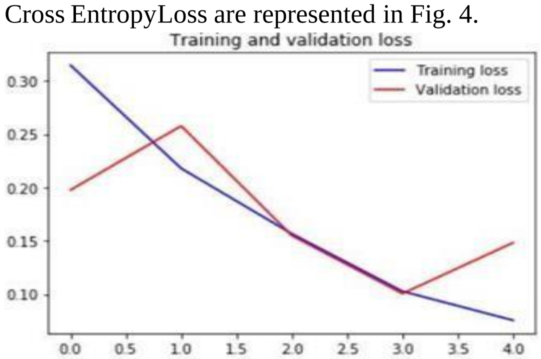

Fig. 4. Test Vs. Training Loss Graph For The Base Model

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 Violenceevent 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.

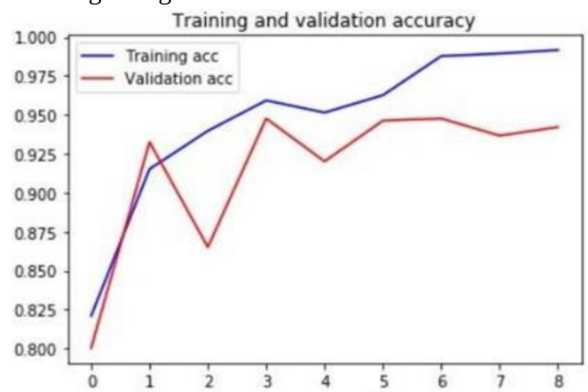

Fig. 5. Test Vs. Training Accuracy Graph For The Proposed Model

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

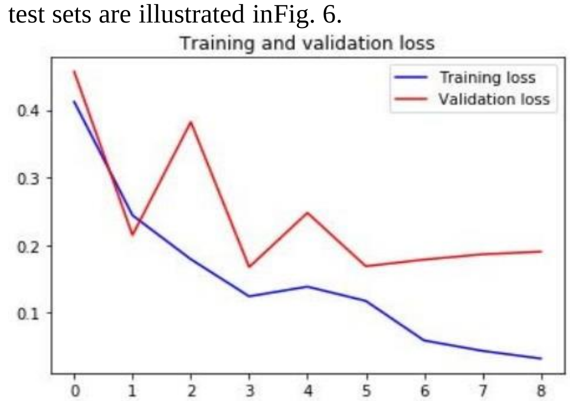

Fig.6. Test vs. Training loss graph for the proposed model

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 fromthe 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". PloSone, 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 ofviolent 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.