2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS) | 979-8-3503-3583-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICISCOIS56541.2023.10100407

# Physical Violence Detection in videos using Keyframing

Bineeshia J
PSG College of Technology
Coimbatore, India
bineeshia.joel@gmail.com

Dhanush Ram A
PSG College of Technology
Coimbatore, India
adhanushram2000@gmail.com

Vinoth Kumar B
PSG College of Technology
Coimbatore, India
bvk.it@psgtech.ac.in

Chidambaram G
PSG College of Technology
Coimbatore, India
chidambaramg5@gmail.com

***Abstract*—Smart surveillance has got tremendous traction in recent years. An important component in smart surveillance is physical violence detection. All existing violence detection models have a pre-processing step where required frames are extracted from a video. The existing interval sampling method is simple and the frames it picks sometimes lose the action in the video. In this literature we have introduced a novel key framing method based on encoding and clustering as a pre-processing step for frame extraction. In the violence detection model, we have used ResNet50 as image encode. Our model is trained on Mix dataset. A dataset formed by combining Hockey fights, Movies and Violent flows dataset. Our model is compared with start-of-the-art models on a test dataset. The test dataset is formed by collecting 20 videos from YouTube and annotating manually. Our key framed model performs better than the state-of-the-art models even with 360 frames less for a 1 minute video clip.**

***Keywords—Violence detection, Keyframe extraction, Spatial-temporal CNN.***

## I. Introduction

Safety and security are an increasing concern in modern society. There is a steady increase in public violence. Most of these cases go unnoticed or it is too late to apprehend the perpetrator. There is lack of immediate notification methods to authorities that can help in the easement of the process. Considering the above mentioned, we have proposed an application that can detect physical violence in public using deep learning method. The proposed application takes feed from CCTV cameras as input and monitors physical violence and guns in public and trigger alarm mechanisms. We also implement key framing algorithms to select suitable frames that can effectively reduce the computational power and load on the server. The pre-trained convoluted model can help in better feature extraction.

Any violence detection system can be divided into the following modules preprocessing, Frame feature extraction, Action recognition, Action classification. In most of the literature's interval sampling is used as preprocessing step. Video have redundant frames. Preprocessing step aimed at filtering these redundant frames and passing only significant frames that represent action. One of the preprocessing steps used is interval sampling. Number of frames required from the video are taken at regular intervals. This helps to eliminate redundant frames because redundant frames are close to each other in a video. If action sequence is concentrated in small duration in the video, Frames representing action are not equally distributed in the video. In such situations interval sampling fails.

In this work, we propose a ResNet50 [1] convLSTM [2] architecture to detect violence. We introduce a new preprocessing method for frame extraction from videos based on keyframing. Videos as a whole contain redundant frames processing all the frames in the video is unnecessary and inefficient. Many existing methods use interval sampling to mitigate this problem. Interval sampling means taking desired number of frames from the video at equal interval. This brings a new set of problem. The frames extracted may not represent the action in the video. The action frames may be concentrated at one segment of the video. It is always better to take frames representing the action in a video. If there is no action there can't be any violence.

The objective of key framing is to extract desired number of frames in the video which better represents the action in the video. The keyframing method we propose makes use of MobileNet V3 [3] encoder and k medoids [4] clustering. The keyframing approach encodes all the frames in the video and applies k means clustering. Clusters of frames equal to the number of frames required are formed. Center from each cluster is taken as the representative frame. Frames in the same cluster are similar to each other and have minimal action. Only one frame per cluster is taken as the keyframe. The keyframes collected are feed to our violence detection model. Violent or non-violent label is returned from the model.

Our novel keyframing is used as a preprocessing step in both training and testing phases. Keyframing introduces significant improvement in violence detection in any model that uses interval sampling. In this literature, we have compared the performance of our proposed model with existing interval sampling and our novel keyframing as preprocessing methods. For training the models we have mixed Hockey fights, movie and violent flows datasets and created a dataset called Mix dataset. For testing purpose, we have collected 20 videos from YouTube and manually have annotated.

This paper is outlined as follows. Section 2 discusses on existing works. Section 3 provides more detail about the model architectures we propose. Section 4 describes the datasets used in this work. Section 5 summarizes the training methodology and experimental results.

979-8-3503-3583-5/23/$31.00 ©2023 IEEE

## II. Related Work

In [5], violence is detected using CNN, Bidirectional LSTM cells and Dense layers. A video byte is fed to the model as input. The model takes frames at evenly spaced time frequencies. Ten frames are filtered from the video. The model passes each filtered frame through a CNN to filter the information in the frame that was passed. After that, Bidirectional LSTM is used to identify the chronological flow of events. Lastly, a dense layer and a classifier determines the presence of violence in the frames.

The model in research [6] has the following components, a 3-D convolution based spatio-temporal encoder and classification layers. Video bytes are given to the network and it creates the first 64 feature maps using the basic convolution layer. Following that, N new feature maps are produced by each layer, where N is a tunable hyperparameter. Every dense layer uses a bottleneck design with pre-activation. Any two dense blocks are separated by a transition layer. To link the network for classification purposes a global average pooling layer was used.

In paper [7], the proposed 3D ConvNet structure is comprised of 13 layers. The first 10 layers are convolution and pooling layers. It is then followed by 2 FC (Fully-connected) layers and a SoftMax layer for assigning values to the determined classes. The proposed approach mainly focuses on a gray-centroid key framing. Distinctive frames are taken based on the proximity to the center. The closest frame is taken to be representative frame. Random sampling is used to get the desired number of frames.

In study [8], a VGG13 network is used for frame feature extraction, a BiConvLSTM is used for action recognition, and four fully connected (1000, 256, 10, 2) dense layers are used for classification. The deviation between adjacent frames are used by the network.

In study [9], four fully linked (1000, 256, 10, 2) dense layers are utilized for classification, a Convolution LSTM (ConvLSTM) is used for temporal analysis and action recognition, and an Alex-Net network with Image-Net training is used to extract frame features. The network uses the deviation between adjacent frames as input.

The work done in [10] takes a different approach, RGB images are fed into the first stream that performs spatial analysis. The second stream focusing on temporal analysis is given stacked grayscale images over 3 channels. Both the streams make use of CNN that can effectively discard patches that contain no useful information. Xception performs the mentioned operation. Both streams' output are merged using class score and classified as videos containing either violence or non-violence.

Paper [11] focuses on low-cost CNN to detect violence. These low-cost CNN can run on edge devices and can detect violence fast. Models like sqeezenet, mobilenetv1, mobilenetv2 and nasnet mobile are trained and compared. Sigmoid function is used to classify violence, non-violence and uncertain detections.

In paper [12] persons are detected by using yolov3 model. Pose information is extracted from the persons using Carnegie Mellon's Open Pose. Instead of sequence of frames, sequence of persons in frame pose information is feed to a custom CNN model. The pose information as feature vector helps the model to predict better with few layers.

In all the papers discussed, dataset used are wither from hockey matches, movies or web-scrapped videos. In paper [6] these datasets are mixed and used for training. Testing is done on validation dataset set aside during training. Best model is chosen by five-fold cross validation. Accuracy is used as the metric for comparison.

## III. Model Architecture

The proposed violence detection system in Fig. 1 takes a video clip as input and returns a violent / non-violent label. It consists of four modules. The key framing module (pre-processing) takes the video clip as input, extract frames and gives 10 prominent frames from the clip. These frames represent the content of the clip better and reduce redundancy. Key framing is used to reduce the amount of frames to the required size of the violence detector model. The 10 input frames are reduced to 5 by taking difference between consecutive pairs of frames. The 5 frames are fed to individual ResNet50 encoders for image feature extraction. The encoders outputs are passed to ConvLSTM action detection model for action recognition. Its output is passed to a fully connected classification model which returns a violence / non-violence label.

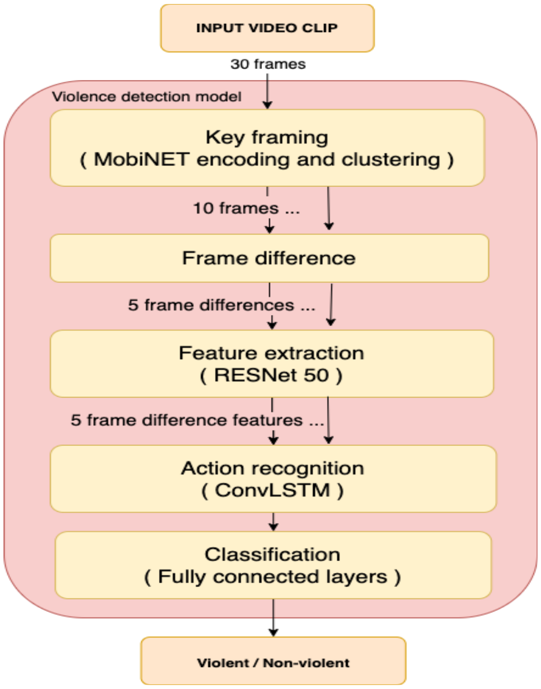


Fig. 1. Architecture of a violence detection model

Key framing in Fig. 2 helps in extracting the best representative frames when only few frames are needed from the video. Traditional sampling methods like interval sampling, samples required number of frames at regular time intervals. This leads to redundant and static frames to be selected especially in situations where actions are dispersed in video. The key framing module first extracts all the frames from the video clip. The extracted frames are encoded using

MobileNet V3 small pretrained on ImageNet. The encoded matrices of each frame are combined and converted into vector point. K medoids clustering is applied on the vector points. The number of clusters to be formed is set as the number of frames required as cluster centers are taken as the key frames. K medoids clustering takes cluster centers as one of the vector points. The centers of the clusters formed are taken and are mapped back to the frames.

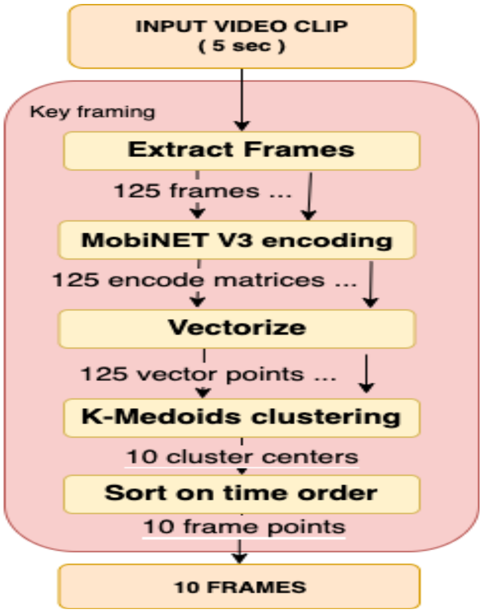


Fig. 2. Key framing architecture

The frames in the same cluster have similar features so, only one representative frame (cluster center) is taken. K medoids is used because cluster centers are always one of the points in the vector set. The frames obtained are sorted and returned to the violence detection model

The violence detector module in Fig. 3 is trained on the Mix dataset. It is composed of 4 modules. The key frames extracted are given to the module as the input. The frame difference module reduces the input number of frames into half by taking difference between consecutive pairs of frames. The frame differences well represent the action than individual frames. The set of frame differences is given as input to the encoder. The encoder module uses ResNet50 model. ResNet50 is chosen because of its robust architecture pre-trained on ImageNet. It is an efficient and powerful image encoder. The encoder encodes individual frame differences and send it to the action detection module. The action detection module is a ConvLSTM with 256 hidden layers. LSTM is a RNN model which recognizes temporal changes. Convulsion LSTM is used for analyzing temporal changes of images. ConvLSTM is chosen because it is better at handling spatiotemporal correlation. Learning spatiotemporal correlations is important for video analysis. The last module classification has 6 layers (Flatten, Batch normalization, Dense 1000, Dense 256, Dense 10, Dense 2). The output of the model is a tuple with two values which ranges from 0 to 1. The tuple maps to non-violence and violence. And, the highest value is taken as result and the corresponding label is returned.

The violence detection system in Fig. 4 extracts keyframes from the video clip and sends it to the violence detector model. The violence detector outputs a tuple indicating the percentage of violence / non-violence. The highest percentage label is returned.

## IV. Data

### A. Hockey Fights dataset

The videos consists of bytes from hockey matches. There are 1000 clips in which Violent and non-violent has 500 clips each. The mean duration is around 1 second. As it covers only hockey fights, all the sceneries and subject in the clips are similar.

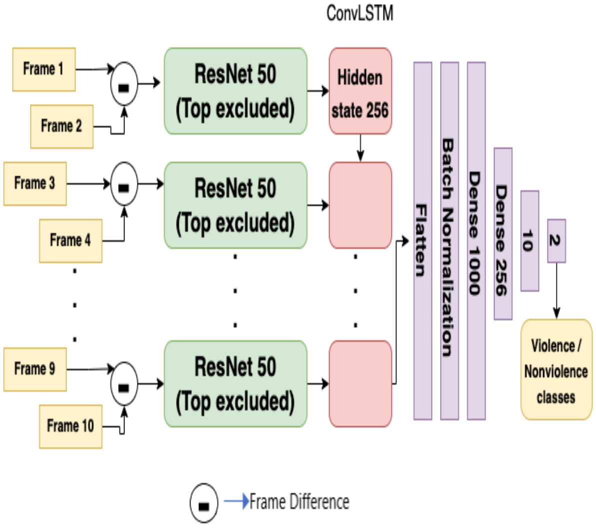


Fig. 3. Violence detection model architecture (after key framing)

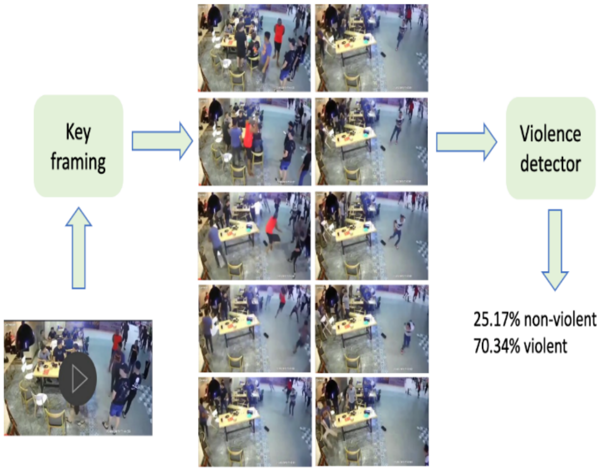


Fig. 4. Violence detector model data flow

### B. *Movies dataset*

It has clips from movies. There are 200 images and violent and non-violent labels have 100 each. Violent clips are from movies and non-violent clips are from open action detection datasets. Clips have different backgrounds and subjects.

### C. *Violent Flows dataset*

Videos in violent flow dataset are taken from YouTube. These videos are taken in real life scenarios and represent crowd violence. In total there are 247 videos. The videos are of varying length. Their average duration is 3.60 seconds.

### D. *Mix dataset*

Videos from the above discussed datasets are combined together to form a dataset called Mix dataset. Our model is trained and validated on this dataset as it represents crowd, real life, sports and movie violence.

### E. *Violence Test dataset*

For testing purpose, 20 videos both violent and nonviolent are collected form YouTube. These videos are manually examined and annotated.

## V. Result Analysis

### A. *Training*

Google colab is used as the platform for training. The models are trained on the mix dataset. Along with the proposed model (Key framed ResNet50 ConvLSTM) three other models are trained for testing. The three models are Interval sampled CNN BiLSTM, Key framed CNN BiLSTM and Interval sampled ResNet50 ConvLSTM. The mix dataset contains videos labelled as violent / non-violent. A series of 10 frames that are labelled as either violent or non-violence are taken as input. The dataset is pre-processed to required form based on the model. 30 percent of frames in the video are extracted using interval sampling / key framing and sequences of 10 frames are labelled as violent / non-violent to generate the required datasets.

The following hyperparameters are selected for each model that is trained. The loss calculation method is based on sparse categorical cross-entropy. As the optimizer, SGD is utilised. The rate of learning is set at 0.001. The metric for training is accuracy. The four models are trained till the peak and stabilisation of their validation accuracy.

The validation accuracy for Interval scaled and key framed CNN BiLSTM models goes to one as the models are small and tend to overfit. In ResNet50 – ConvLSTM models, Batch normalization prevents overfitting. In the validation accuracy between interval sampled and key framed ResNet50 ConvLSTM models' key framed model has higher accuracyas shown in Fig 5 and 6.Table 1 shows the validation accuracy comparison for interval and key framed models.

### B. *Testing*

The models are tested on the violence test dataset collected. For comparison, along with the proposed model (Key framed ResNet50 ConvLSTM) three other models are also trained and tested. The three models are Interval sampled CNN BiLSTM, Key framed CNN BiLSTM and Interval sampled ResNet50 ConvLSTM. The violence test dataset contains videos labelled as violent / non-violent. The models' input is a sequence of 10 frames. The dataset is transformed to required form based on the model. 30, 10, 5 percent of the video is extracted using interval sampling / key framing and sequences of 10 frames are generated as the input for testing the different models.

The results shows that a model with key framing performs better even with a smaller number of frames extracted from video clip are used to test. In CNN - BiLSTM models, with 30% of the frames extracted from the videos, key framed performs equal to interval sampled. Similarly for 10% and 5% frames extracted it performs 1% better.

As shown in Table 2, in ResNet50 – ConvLSTM models, with 30% of the frames extracted from the videos, key framed performs 8% better than interval sampled. Similarly for 10% and 5% frames extracted it performs 6% and 19% better. From these results we infer that interval sampled violence detection models' performance degrades with less no of frames extracted from the videos. Our key framed ResNet50 – ConvLSTM on an average performs 19% better than interval sampled CNN BiLSTM.

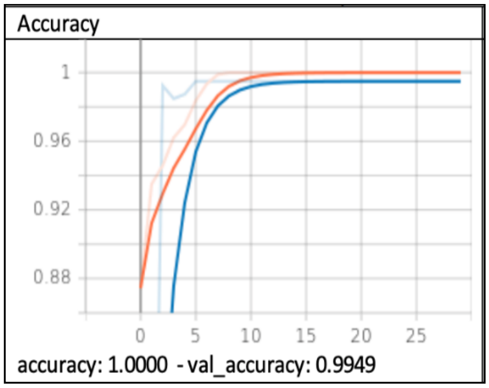


Fig. 5. Resnet50 ConvSTM – Interval sampled - Accuracy Graph

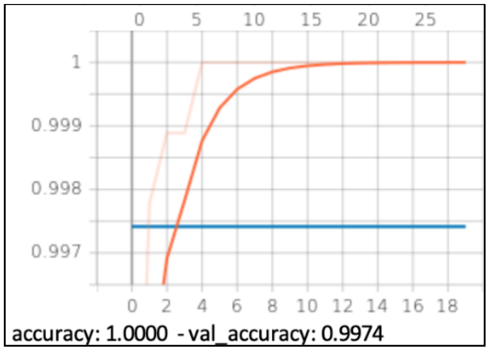
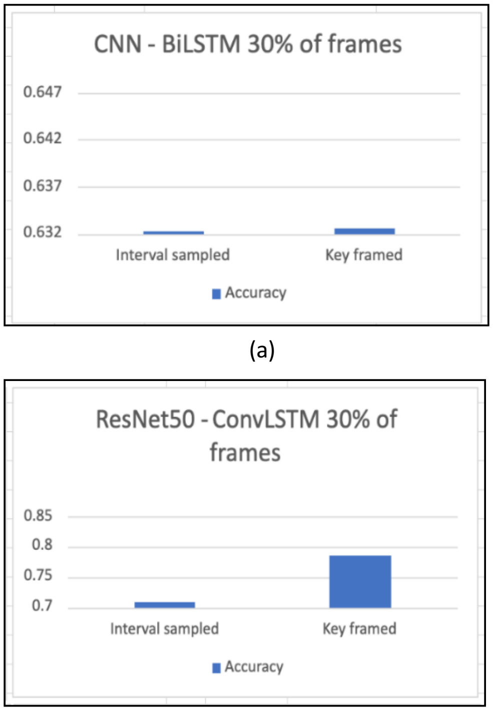


Fig. 6. Resnet50 ConvSTM – Key framed - Accuracy Graph

TABLE I. VALIDATION ACCURACY COMPARISON FOR ALL FOR MODELS

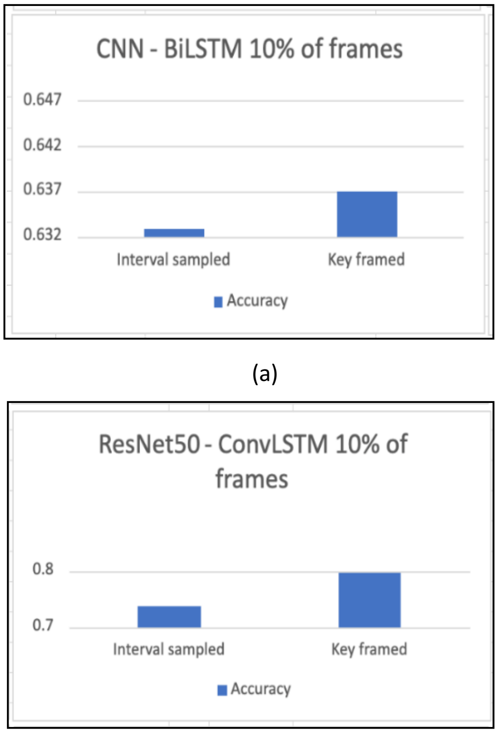
| Validation accuracy | | |
|---|---|---|
| | ***CNN - BiLSTM*** | ***ResNet50 - ConvLSTM*** |
| ***Interval Sampled*** | 1 | 0.9949 |
| ***Key framed*** | 1 | 0.9974 |



(a)

(b)

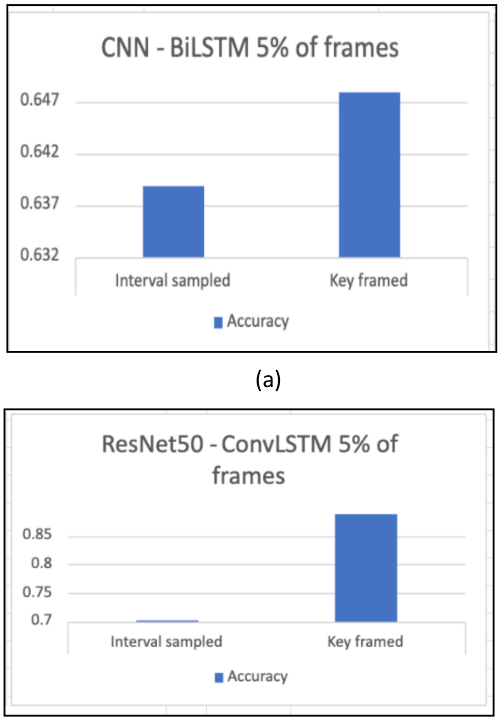
Fig. 7. Testing accuracy when 30% of the frames extracted from the videos (a) CNN- BiLSTM (b) Resnet50 ConvSTM



(a)

(b)

Fig. 8. Testing accuracy when 10% of the frames extracted from the videos (a) CNN- BiLSTM (b) Resnet50 ConvSTM



(a)

(b)

Fig. 9. Testing accuracy when 5% of the frames extracted from the videos (a) CNN- BiLSTM (b) Resnet50 ConvSTM

## VI. Conclusion

Most of the existing violence detection models use interval sampling as a pre-processing step to extract frames from video. Our violence detection model introduces key framing as a pre-processing step. Our model was able to perform better with the introduction of key framing with respect to both accuracy and lesser computational power. The model is trained in such a way that it does not overfit. Key-framed violence detection performs better with a smaller number of frames. The proposed model with key framing performs 19% better than with interval sampled with only 5% of the frames extracted from the video. And our key framed ResNet50 – ConvLSTM on an average performs 19% better than interval sampled CNN BiLSTM. The future work for violence detection would be to collect better datasets as existing datasets are small and limited in size and are skewed to a specific context of violence. Our key framing approach as a pre-processing method can be used and tested on other violence detection methods.

*TABLE II.* TESTING LOSS & ACCURACY COMPARISON FOR ALL FOR MODELS WITH DIFFERENT PERCENT OF FRAMES EXTRACTED FOR VIOLENCE DETECTION

| **Test 30 % frames** | | | | |
|---|---|---|---|---|
| | ***CNN BiLSTM*** | | ***ResNet50 - ConvLSTM*** | |
| | ***Interval sampled*** | ***Key framed*** | ***Interval sampled*** | ***Key framed*** |
| ***Loss*** | 3.36 | 3.78 | 1.39 | 0.69 |
| ***Accuracy*** | 0.63 | 0.63 | 0.71 | 0.79 |
| **Test 10 % frames** | | | | |
| | ***CNN BiLSTM*** | | ***ResNet50 - ConvLSTM*** | |
| | ***Interval sampled*** | ***Key framed*** | ***Interval sampled*** | ***Key framed*** |
| ***Loss*** | 3.38 | 3.79 | 1.32 | 0.68 |
| ***Accuracy*** | 0.63 | 0.64 | 0.74 | 0.80 |
| **Test 5 % frames** | | | | |
| | ***CNN BiLSTM*** | | ***ResNet50 - ConvLSTM*** | |
| | ***Interval sampled*** | ***Key framed*** | ***Interval sampled*** | ***Key framed*** |
| ***Loss*** | 3.31 | 3.70 | 1.36 | 0.37 |
| ***Accuracy*** | 0.64 | 0.65 | 0.70 | 0.89 |

## REFERENCES

[1] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[2] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: a machine learning approach for precipitation nowcasting. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 802–810.

[3] Howard et al., "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140.

[4] Jin X., Han J. (2011) K-Medoids Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

[5] Halder, R., Chatterjee, R." CNN-BiLSTM Model for Violence Detection in Smart Surveillance." SN COMPUT. SCI. 1, 201 (2020).

[6] Li, J., Jiang, X., Sun, T. and Xu, K., "Efficient Violence Detection Using 3D Convolutional Neural Networks," 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, pp. 1-8, doi: 10.1109/AVSS.2019.8909883.

[7] Song, W., Zhang, D., Zhao, X., Yu, J., Zheng, R. and Wang, A.,"A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks," in IEEE Access, vol. 7, pp. 39172-39179, 2019, doi: 10.1109/ACCESS.2019.2906275.

[8] Hanson, A., PNVR, K., Krishnagopal, S., Davis, L., (2019) Bidirectional Convolutional LSTM for the Detection of Violence in Videos. In: Leal-Taixé L., Roth S. (eds) Computer Vision – ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science, vol 11130. Springer, Cham.

[9] Sudhakaran, S. and Lanz, O., "Learning to detect violent videos using convolutional long short-term memory," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

[10] Mehmood, A., "Efficient Anomaly Detection in Crowd Videos Using Pre-Trained 2D Convolutional Neural Networks," in IEEE Access, vol. 9, pp. 138283-138295, 2021, doi: 10.1109/ACCESS.2021.3118009.

[11] J. C. Vieira, A. Sartori, S. F. Stefenon, F. L. Perez, G. S. de Jesus and V. R. Q. Leithardt, "Low-Cost CNN for Automatic Violence Recognition on Embedded System," in IEEE Access, vol. 10, pp. 25190-25202, 2022, doi: 10.1109/ACCESS.2022.3155123.

[12] K.B. Kwan-Loo, J. C. Ortíz-Bayliss, S. E. Conant-Pablos, H. Terashima-Marín and P. Rad, "Detection of Violent Behavior Using Neural Networks and Pose Estimation," in IEEE Access, vol. 10, pp. 86339-86352, 2022, doi: 10.1109/ACCESS.2022.3198985.

[13] Yuan J, Liu Z, Wu Y. Discriminative subvolume search for efcient action detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009; pp. 2442–2449. IEEE.

[14] Itcher Y, Hassner T, Kliper-Gross O. Violent fows: Real-time detection of violent crowd behavior. In: 3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012.

[15] Deniz O, Serrano I, Bueno G, Kim T-K. Fast violence detection in video. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), 2014. volume 2, pp. 478– 485. IEEE.

[16] Bilinski P, Bremond F. Human violence recognition and detection in surveillance videos. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 30–36. IEEE, 2016.

[17] Abdali A-MR, Al-Tuma RF. Robust real-time violence detection in video using cnn and lstm. In: 2019 2nd Scientifc Conference of Computer Sciences (SCCS), 2019; pp. 104–108. IEEE.

[18] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning," 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), 2018, pp. 558-563, doi: 10.1109/EECS.2018.00109.

[19] N. Honarjoo, A.Abdari, and A. Mansouri, "Violence detection using pre-trained models," 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2021, pp. 1-4, doi: 10.1109/IPRIA53572.2021.9483558.

[20] M. Gadelkarim, M. Khodier, and W. Gomaa, "Violence Detection and Recognition from Diverse Video Sources," 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1-8, doi: 10.1109/IJCNN55064.2022.9892660.