International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 4, August 2021, pp. 3374~3380
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i4.pp3374-3380

A fully integrated violence detection system using CNN and LSTM

Sarthak Sharma, B Sudharsan, Saamaja Naraharisetti, Vimarsh Trehan, Kayalvizhi Jayavel
Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

Article Info

Article history:

Received Sep 1, 2020
Revised Dec 20, 2020
Accepted Jan 13, 2021

Keywords:

Deep learning
LSTM
Mobile application
Smart cities
Transfer learning
Violence detection

ABSTRACT

Recently, the number of violence-related cases in places such as remote roads, pathways, shopping malls, elevators, sports stadiums, and liquor shops, has increased drastically which are unfortunately discovered only after it's too late. The aim is to create a complete system that can perform real-time video analysis which will help recognize the presence of any violent activities and notify the same to the concerned authority, such as the police department of the corresponding area. Using the deep learning networks CNN and LSTM along with a well-defined system architecture, we have achieved an efficient solution that can be used for real-time analysis of video footage so that the concerned authority can monitor the situation through a mobile application that can notify about an occurrence of a violent event immediately.

Corresponding Author:

Kayalvizhi Jayavel
Department of Information Technology
SRM Institute of Science and Technology
SRM Nagar, Kattankulathur, Chengalpattu District, Tamil Nadu-603203, India
Email: kayalvij@srmist.edu.in

1. INTRODUCTION

In this day and age, violence has been drastically increasing and is posing serious threats to humans, systems, and buildings. The situation becomes worse when violence takes place in public where most people are not accountable and cannot be held responsible without proof. Most of the heinous crimes take place in public due to their anomalous nature.

When we are talking about violent activities, it usually refers to an unusual physical interaction that happens between two or more people [1]. Monitoring the surveillance has led to a lot of difficulty for security personnel as they now have to painstakingly go through the footage to find the culprit specifically and track his moves from one camera to another or view it in real-time to detect violent activities and behaviour before or as they are occurring. A major constraint for this is also the quality of surveillance videos that they are provided with. Most snapshots from surveillance footage do not hold good in court as the defendant will deny their presence in the photo.

Since the violence in a city can occur at any time, relying on a human to monitor and detect violent events is not an efficient way to handle such cases. Such activities usually lead to very unpleasant scenarios which makes it very crucial for automatic detection of such events through real-time video footage to take place, so that the required, crucial decision can be made by the concerned authority. As a result, the idea of implementing systems and equipment has been introduced to detect such incidents using video retrieval and real-time monitoring. The prime focus is to get rid of the above-mentioned real-world constraints and decrease the crime rates significantly and efficiently.

Technology has revolutionized our world and daily life. We are now able to combine various technologies that are used to detect objects and various movements and create a system that will help in detecting such potential threats. In recent years, there has been a massive amount of work established on human action recognition [2, 3]. Specifically, employing deep learning for classifying the video sequences has been achieving better results than the existing handcrafted methods [4, 5]. The research related to violence detection usually involves the use of two popular datasets-Movies and Hockey [6].

Transfer learning is a very popular technique that is used in machine learning, which uses the features that it learns from a source domain to make predictions with a target domain. To deal with problems arising from huge volumes of data, many kinds of research choose the strategy of transfer learning to solve the problem. In our case, we are re-training the last 4 layers of the pre-trained Xception network and then utilizing it as a feature extractor for our custom classifier. With this technique, these models can be repurposed for any related work we require, from object detection for self-driving vehicles, action recognition to classifying video clips [7, 8].

Long short-term memory (LSTM) [9] network is a type of recurrent neural network used widely in sequence prediction problems where it can learn the order dependence. LSTM, as the name suggests, can retain huge amounts of information by default for a long period of time. CNN followed by LSTM has been proven to be the best architecture when the available data are small and the computing power resources are not very high for the task. Our proposed CNN + LSTM model can process the videos with a speed of 126 frames per second in our test environment.

2. RELATED WORK

The initial work for violence classification mainly revolved around audio-video correlation [10], detecting the presence of blood, vigorous degrees of motion, and identifying sound features such as screams [11, 12]. Carneiro *et al.* [13] focused to implement a hand-drawn high-level description and multi-stream-based learning model to solve the conflict detection problem in videos. One of the recent works in this field proposed [14] a system that works on the HOG features of video frames. The authors extracted HOG features from binary images and used the random forest classifier to identify the existence of violence in each frame. Finally, they employed the majority voting technique to classify the video clip into violence or non-violence. Although this system doesn't require a GPU for computations and establishes improved results compared to previous works, it suffers from low accuracy.

Recent research work on fight and violence detection shows the extensive implementation of deep learning architectures such as convolutional neural networks (CNNs) [15], long short-term memory (LSTMs), and two stream CNNs [16]. These automatic methods perform much better than the hand-crafted algorithms used for spatio-temporal feature extraction.

Mumtaz *et al.* [17] used transfer learning with GoogLeNet (Inception) [18] which consisted of 22 layers, over the two popular datasets-Hockey and Movies. Both of the datasets have their own complexity. The annotated videos were converted into labelled image frames, and the 1000 class layer is replaced with two categories (violence and non-violence). The result shows the accuracy of 99.28% and 99.97% in the Hockey and Movies dataset respectively. In [19], Perez *et al.* proposed a methodology of feature extraction with the two-stream based solution, consisting of two different 2D-CNN models. One is trained with the RGB frames of the video and the other one is trained with a stack of optical flows from the video frames.

Violence detection using CNN and LSTM has been employed in [20] where the author combines the three benchmark datasets. The methodology involves the use of a pre-trained CNN architecture VGG-19 [21] followed by LSTM which works with an input of 30 frames at a time. The results from CNN of each frame are grouped and then fed to the LSTM as a sequence. The model performs well, with an accuracy of 94.765% on their combined dataset.

The methodology proposed in [22] involves the use of CNN along with ConvLSTM for characterizing the videos. They have made use of the AlexNet model pre-trained on the ImageNet database as the CNN model for feature extraction. Their results have been proved with the three benchmark datasets, achieving the maximum 100% accuracy with the Movies dataset. The methodology from [23] focuses on creating a new localized guided fight action detection framework for realistic surveillance videos. A new dataset with 1520 videos was proposed and state-of-the-art models were trained over the dataset to achieve high levels of accuracies. They have used the pre-trained SSD-VGG16 for human detection and the pre-trained FlowNet 2.0 [24] model for estimating optical flow. A two-stream C3D network [25] is trained on the active regions extracted from the localization phase which are later combined for predictions.

Interestingly, Serrano *et al.* [26] proposed a method where a video sequence is summarized into a representable image, which can be used to classify the scene as violent or non-violent. Additionally, the methodology leverages those zones that could be important for the classification. So, the most important

parts of the sequence are given more weight, giving less importance to noise and static background. Finally, 2D CNN takes the representative image to classify it as needed. The result provided an accuracy of 99±0.5% for Movies and 94.6±0.6% for the Hockey dataset.

Seymanur *et al.* [27] solve the problem of detecting fights from surveillance cameras and explore the LSTM-based approach to unravel fight detection in video with the assistance of an attention layer. It's recognized that this method that has been proposed, confluences the Xception [28] model as well as BiLSTM. In [29], the violence detection system has been implemented using 3D ConvNet and keyframe extraction algorithm to produce an effective approach. They achieved 93.5% with the Crowd violence dataset. Keçeli *at al.* [30] use transfer learning again using AlexNet but has utilized the Lucas-Kanade method for finding the optical flows of the sequence of frames. From the optical flow values, templates are made which are then given as the input to pre-trained CNN AlexNet for feature extraction. Finally, two classifiers are employed support vector machines (SVM) and subspace k-nearest neighbors (SkNN). The best results have been obtained from the SVM classifier.

3. PROPOSED METHOD

The proposed architecture uses convolution neural networks as the spatial feature extractor followed by an LSTM network to perform sequence prediction on the feature vectors. For the spatial feature extraction, we have employed a transfer learning approach with CNN. The architecture of Xception [28] network has been considered with the pre-trained model on the ImageNet dataset [31]. Instead of training from scratch, we used a pre-trained Xception model as a feature extractor as it performed better than other pre-trained CNN models like VGG [21], LeNet [32] or ResNet. [33]. And we fine-tuned it by keeping the initial layers intact and retraining the last 4 layers on respective datasets

For the datasets, we have considered Hockey, Movies, and the UCF Crime dataset. Hockey and Movies are the well-known standard benchmark datasets. UCF Crime dataset [34] consists of CCTV footage of various categories of violence. We have used 4 categories, namely-fighting, assault, abuse, and arrest, forming the violence category, and the normal category as the non-violence category. The violence videos were manually trimmed to contain only the scenes having violence in them.

In the architecture, we are dealing with sequential input of shape (15x200x200x3) which corresponds to (frame x H x W x channels). In real-time considering a 30 fps feed, for every 90 frames, every 6^{th} frame is taken to form a sequence of 15 frames. Here we are using a sequence of 15 frames where each frame is of an RGB format and with 200x200 size. But since the Xception network can only accept 3D inputs we used the time distributed layer. Time distributed layer wrapper applies the convolution model to every temporal slice of the input. Time Distributed is a special type of layer wrapper present in the Keras library. It applies the same layer for a list of chronological input.

In our model, the input consists of 15 chronologically ordered frames, so the time distribution operation applies the same Xception model for every frame. These layers share the same weights. For 15 images, the weights are not tweaked 15 times, but only once, and distributed to every block defined in the current time distributed layer. We use this technique to apply the Xception network to all 15 inputs to get feature maps with 2048 channels. These feature maps are then flattened to a 2D tensor of shape (15, 100352) which is fed into the LSTM having 512 cells that try to learn time relations between 15-time steps. Finally, we take the 1D predictions from LSTM and feed it to a series of dense layers to get the output predictions. Softmax activation is used in the output layer. The basic architecture is shown in Figure 1 (see in appendix).

For training, we used cross entropy loss as the loss function. A batch normalization layer added before the dense layers helped to reduce the loss and speed up the training process by a bit. We also used dropout [35] layers to reduce the overfitting of the model on training data.

4. SYSTEM ARCHITECTURE

Our system aims to provide real-time violence detection that can decrease labor and help the concerned police authorities to efficiently monitor their locality and quickly tend to any site if violence has been detected. The architecture is integrated with various cloud services that are highly scalable and robust as shown in Figure 2. As more and more localities and police departments will be added, scalability plays a crucial role. On the detection of any violence, the details of the occurrence such as the camera ID, location, timestamp, and few snapshots of the occurrence are saved in the database, followed by an alert that is issued by the cloud message service to all the concerned authorities of that particular locality. We have used PostgreSQL for designing our database schema.

The mobile application allows the police authorities to continuously monitor their locality all the time. The app provides the live feed of the surveillance that the authority can choose to supervise. Every alert

issued can be marked as attended by the police when they start taking any kind of action. This makes sure that everyone else in that locality's department is updated about the status of the alert and who's taking care of it.

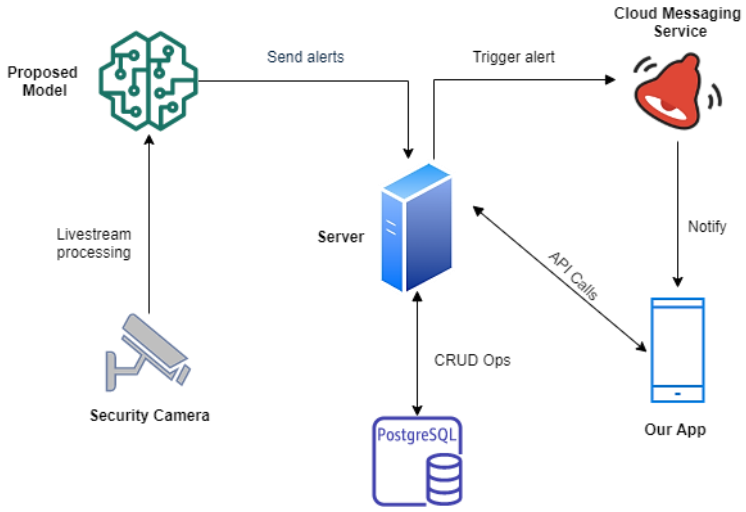

Figure 2. System architecture

5. RESULT AND ANALYSIS

The mobile application has been successfully developed and integrated with all the essential services, to provide a robust violence detection system as shown Figure 3. It would cover all the security cameras that are installed in the authority's locality under control. During any occurrence of violence, the officials governing that locality are immediately notified. The alert received provides all the crucial information required to take any necessary action immediately. It also provides a few snapshots of the area that can help them make better decisions on how to bring the situation under control. Every functionality is tested in real-time and has been connected to work with our proposed deep learning model.

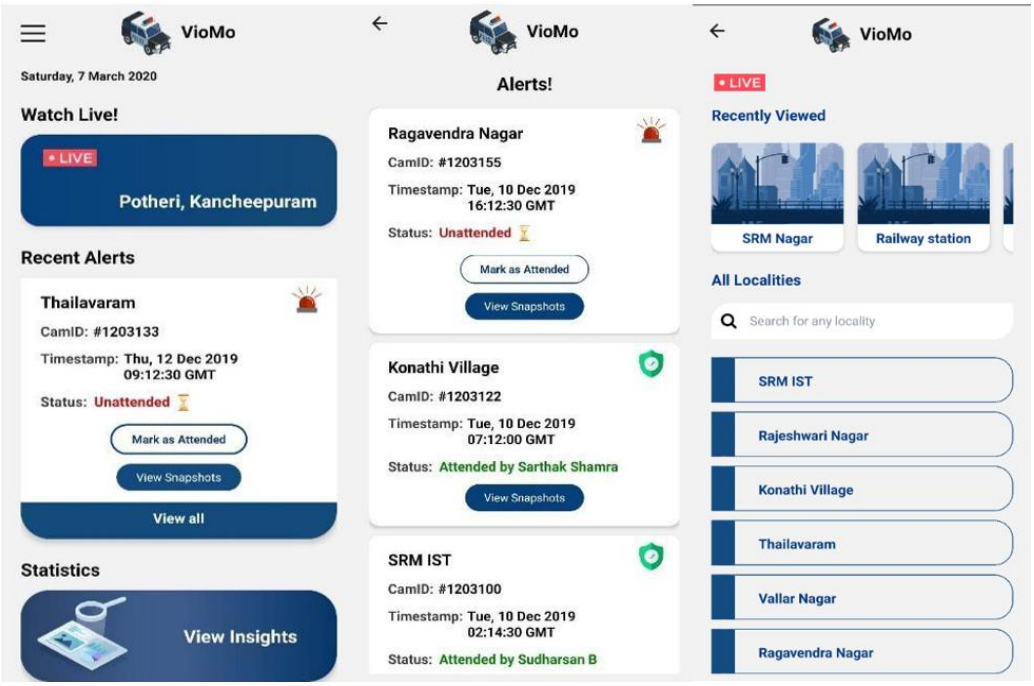

Figure 3. Mobile application (VioMo)

Our model based on Xception and LSTM was first trained on the two benchmark datasets, namely-Hockey and Movies. The Movies dataset consists of 200 videos, and the Hockeys dataset consists of 1000 videos. The datasets have an equal number of Violence and Non-Violence videos. The code has been written in Python using the Keras library. The training was done with a 5-fold cross validation technique having 35 epochs for each fold. The results of test accuracies are displayed and compared in Table 1.

Table 1. Accuracies reported on the benchmark datasets-Movies and Hockey

Algorithm	Year	Movies dataset-accuracy	Hockey dataset-accuracy
MoSFIT+HIK [6]	2011	89.50%	90.90%
Optical flows+CNN+SVM and SkNN [23]	2017	96.50%	94.50%
Inception+Transfer learning [17]	2018	99.97%	99.28%
FAST Detector+Hough forests with CNN classifier [26]	2018	99%	94.60%
Multi-stream VGG-16 [13]	2019	100%	89.10%
VGG19+LSTM [20]	2019	100%	96.33%
HOG+Random Forest [27]	2019	--	86.00%
Proposed: Xception + LSTM	2020	98.32%	96.55%

We observed that the model trained on the benchmark datasets do not work accurately with the real-time CCTV footage. It is mainly due to the fact that the videos are unrealistic and do not aptly depict the real-world scenarios. These videos differ a lot from the actual CCTV ones in terms of the camera angle too. To overcome this and validate our architecture for real-time analysis, the UCF Crime dataset was taken which makes our model perform better in real-time. Our modified UCF Crime dataset consists of an equal number of 160 trimmed violence and non-violence videos. We trained the model on this for 50 epochs, with a train/test split scheme with 80% for training and 20% for testing. The results from the same are shown in Figure 4. We achieved results from the testing set of accuracy of 98.87%, which is better than the results we obtained using the benchmark datasets and the other existing systems. The loss depicted by both train and test datasets are calculated using categorical cross-entropy.

The proposed model trained takes approximately 7.89 milliseconds per frame for the classification (without including the time for preprocessing). Overall, the proposed method takes approximately 0.28 seconds for processing of a 3-second video clip at 30 fps. The use of the UCF Crime dataset and fast processing time makes it suitable for violence detection in real-time video processing applications. The entire experiment was performed on a 12 GB NVIDIA Tesla K80 GPU.

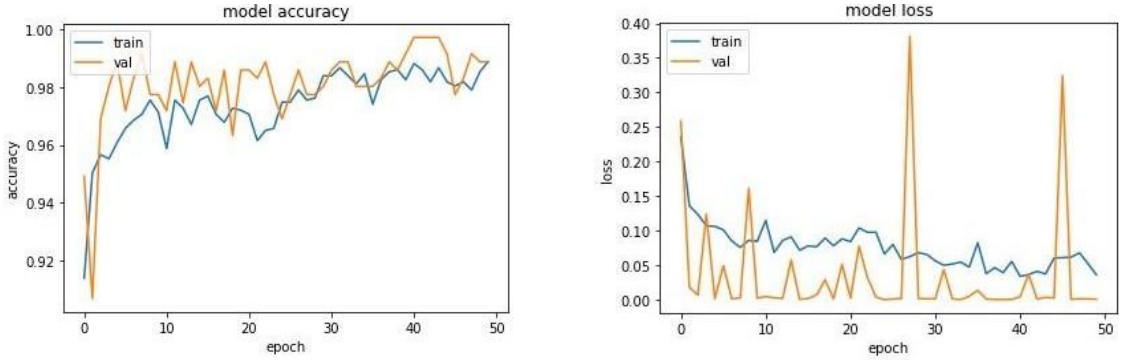

Figure 4. Accuracy and loss graphs of our model for the modified UCF crime dataset

6. CONCLUSION

Nowadays, the rate of violence around us is increasing drastically, acting as a threat to humans, buildings, and systems. There has always been a need for a better system that can aid the police in monitoring the violence, which is usually hard to handle as it is a group activity and the process of elimination to find the culprit is time-consuming. The results from our experiment demonstrate the effective use of CNN+LSTM architecture, for training the model over two popular Hockeys and Movies dataset, and our modified UCF Crime dataset. Our work provides a fully integrated system that can help the police authorities monitor their area under control. It efficiently utilizes the existing cloud services to deliver a robust mobile application that can enhance the current functions of the police department.

Our model that has been trained over the benchmark datasets has shown very powerful results. The excellent results over the UCF Crime dataset with a good speed as compared to other approaches makes our model function accurate in real-life scenarios. Apart from producing a novel model for violence detection, we have constructed a fully functioning system that supports this model to work well in the real-world, making our approach unique, detailed, and advanced from the existing solutions. The system precisely alerts the police through the app during any occurrence of violence and allows the police to take action accordingly. The authority can use our system to manage the crimes around them in a much efficient manner.

Our proposed system can be improved to perform better in many aspects. The system can be improved by specifying how severe the detected violence is. This can help the authorities make better decisions. The proposed Deep Learning architecture can be altered by modifying the hyperparameters, to improve the performance. Also, the system can be extended to serve the purpose of other types of detection such as fire accidents and burglary.

APPENDIX

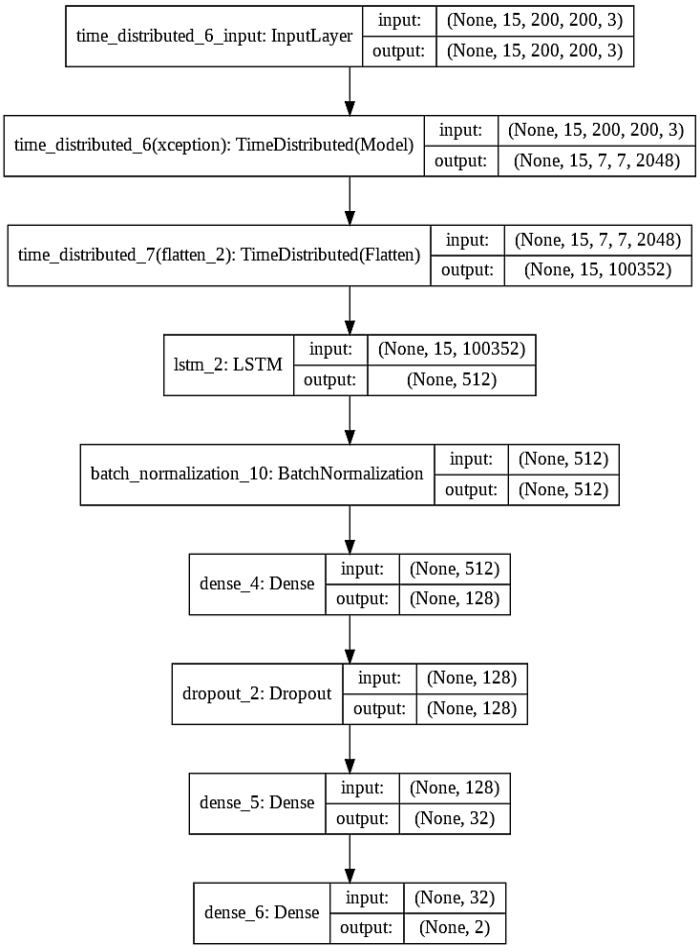

Figure 1. Model architecture

REFERENCES

[1] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2009.

[2] S. R. Ke, H. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo, and K. H. Choi, "A Review on Video-Based Human Activity Recognition," *Computers*, 2013.

[3] H. Kim, S. Lee, and H. Jung, "Human activity recognition by using convolutional neural network," *International Journal of Electronics and Communication Engineering (IJECE)*, vol. 9, no. 6, pp. 5270–5276, 2019.

[4] D. Wu, N. Sharma, and M. Blumenstein, "Recent Advances in VideoBased Human Action Recognition using Deep Learning: A Review," *International Joint Conference on Neural Networks* (*IJCNN)*, 2017, pp. 2865-2872.

[5] A. Sargano, P. Angelov, and Z. Habib, "A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition," *Applied Sciences*, vol. 7, no. 1, 2017, Art. no. 110.

[6] E. N. Bermejo, O. D. Suarez, G. B. Garcıa, and R. Sukthankar, "Violence detection in video using computer vision techniques," *Computer Analysis of Images and Patterns (CAIP)*, vol. 6855, pp. 332–339, 2011.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *2014 IEEE Conf. of Computer Vision and Pattern Recognition (CVPR*), 2014, pp. 1725–1732.

[8] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 463–469.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," *Pacific-Rim Conference on Multimedia*, 2008, pp. 317–326.

[11] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," *Hellenic Conference on Artificial Intelligence*, 2006, pp. 502–507.

[12] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," *Proceedings 1998 International Conference on Image Processing (ICIP98)*, 1998, pp. 353-357.

[13] S. A. Carneiro, G. P. da Silva, S. J. F. Guimara˜es, and H. Pedrini, "Fight Detection in Video Sequences Based on Multi-Stream Convolutional Neural Networks," *32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2019, pp. 8-15.

[14] S. Das, A. Sarker, and T. Mahmud, "Violence Detection from Videos using HOG Features," 2019 *4th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, 2019, pp. 1–5.

[15] G. Sakthivinayagam, R. Easawarakumar, A. Arunachalam, and M. Pandi, "Violence Detection System using Convolution Neural Network," *SSRG International Journal of Electronics and Communication Engineering (IJECE)*, vol. 6, no. 2, 2019, Art. no. 102.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[17] A. Mumtaz, A. Bux and Z. Habib, "Violence Detection in Surveillance Videos with Deep Network using Transfer Learning," *Electrical Engineering and Computer Science (EECS)*, pp. 558–563, 2018.

[18] C. Szegedy *et al*., "Going Deeper with Convolutions," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 1–9.

[19] M. Perez and A. C. Kot, "Detection of real-world fights in surveillance videos," *International Conference on Acoustics, Speech, and Signal Processing*, 2019, pp. 2662-2666.

[20] Al-Maamoon R. A. and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," *Student Conference on Conservation Science*, 2019, pp. 104-108.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv: 1409.1556*, pp. 1–14, 2014.

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

[23] Q. Xu, J. See, and W. Lin, "Localization Guided Fight Action Detection In Surveillance Videos," *IEEE International Conference on Multimedia and Expo*, 2019, pp. 568-573.

[24] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," *Computer Vision and Pattern Recognition (CVPR*), vol. 2, pp. 1-16, 2017.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," *Proceedings of the IEEE Int. Conference on Computer Vision*, 2015, pp. 4489-4497.

[26] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in video using Hough Forests and 2D Convolutional Neural Network," *IEEE Trans, on Image Processing*, vol. 27, no. 10, pp. 4787-4797, 2018.

[27] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras," *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, Turkey, 2019.

[28] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 1251-1258.

[29] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks," *IEEE Access*, vol. 7, pp. 39172–39179, 2019.

[30] A. S. Keçeli and A. Kaya, "Violent activity detection with transfer learning method," in *Electronics Letters*, vol. 53, no. 15, pp. 1047–1048, 2017.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, pp. 1–9, 2012.

[32] LecunY., Bottou L., Bengio Y *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[33] 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.

[34] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *arXiv: 1801.04264*, 2018.

[35] S. Wang and C. Manning, "Fast dropout training," *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 2, pp. 118–126, 2013.