Violence detection in videos using Conv2D VGG-19 architecture and LSTM network

J.V. Vidhya [a] and R. Annie Uthra [b]

[a] *SRM Institute of Science and Technology, Kattankulathur, Kancheepuram, Tamilnadu, 603203, India*
[b] *SRM Institute of Science and Technology, Kattankulathur, Kancheepuram, Tamilnadu, 603203, India*

Abstract

Violence identification from surveillance videos can be considered as a special form of Activity recognition, that targets at recognizing human actions in public places. A video sequence is a collection of consecutive frames that is sampled in both temporal and vertical directions and the given input video is converted into frames and the preprocessing is done at the frame level. For Feature extraction the 2D convolutional neural network (Conv2D) is used and it adapts the layers of VGG-19 net architecture with global average pooling and learns the spatial information in the given video. Those extracted features are then combined using Long Short Term Memory (LSTM) and it learns about temporal information from the video. The model is validated using the Hockey data set and a loss of 0.02 and accuracy of 98 is achieved.

Keywords

Violence identification, Activity recognition, Preprocessing, Feature extraction, VGG-19, Global average pooling, Long short term memory (LSTM).

1. Introduction

Video is modernizing the way it brings changes in the world. With the raising rates of the crime incidents, the use of popular security-enhancing technological devices called Closed-Circuit Television (CCTV) as an effective security measure is on the rise around the world. There are about 770 million CCTV cameras installed worldwide so far. Violence identification from surveillance videos can be considered as special form of Activity recognition [10], that targets at recognizing human actions in public places. Monitoring the suspicious activity of the human being throughout the day becomes a tedious task [4], [5]. This leads to the necessity of methods to detect abnormal human activity automatically. Video recognition of human behavior was carried out using machine learning techniques and computer vision techniques [1]- [3]

In the past years, several researches have been carried out over activity recognition and tested the model on quite simple datasets, which contains various actions simulated by actors in an environment.[6]-[9]. There are few factors that differentiate abnormal and violence activity. The activities which are unlike normal activity are termed as abnormal activity such as Beating, Stealing, Harassment, fighting are examples of violent activities [13]- [15]. Automatic recognition of violence in videos are becoming essential as it can reduce the time and labour consumption. There are many approaches and methods built to detect brutal events and other uncertain patterns in the videos [11], [12]. Conventional Feature mining methods with classifiers and deep learning framework can be used for this purpose. Machine learning and Deep learning provides a great way to detect the violence in the video and classify with high accuracy and less response time. Traditional methods used in earlier stages are STIPs [16]-[17] and MoSIFT [18]-[20]. Major machine learning algorithms used to recognize or classify the objects or persons are expected to over fit in the process of training data. Visual data are complex in nature. Due to complexity, the models incline to have input of high dimension and a lot of

Algorithms, Computing and Mathematics Conference, August 19 – 20, 2021, Chennai, India.
EMAIL: vidhyaj@srmist.edu.in (J. V. Vidhya)
ORCID: 0000-0002-9196-2867 (J. V. Vidhya)


parameters are to be used to fit in the model. Overfitting happens when the training data is low. As a result, the models cannot be generalized.

Deep Learning eliminates the need of hand-crafted features [21] [22]. The model can be generalized well using large training dataset. Deep Learning methods are successful for recognizing video based activities [23] [24]. ImageNet contains all the images of real world of various classes. It has about 14 million plus images organized in twenty thousand classes of objects or scenes. It used as a benchmark dataset for training the model in deep neural network [24].

Our work has been organized as follows, in section 2, several algorithms that were used to classify violent and non-violent actions in the earlier years are addressed. Section 3 shows the proposed approach implementation which comprises Pre-processing, Classification then prediction of videos to identify violence or nonviolence. Section 4 contains the experimental results of the model evaluated using Hockey dataset.

2. Related Works

A statistical technique based on optical flow to identify violent behavior in a crowd scene is proposed in [25]. Statistical characteristic of the optical flow descriptor (SCOF) was used to denote the video frame sequence. SCOF descriptor categorize the video as violence or non-violence using linear Support Vector Machine. The proposed model was tested on the Hockey dataset and Crowd Dataset. 86.9% and 86.37% accuracy is observed in Hockey and Crowd dataset.

Zhang et al. [26] suggested a novel method for localization and detection of violence in surveillance videos. Violence regions are extracted using the Gaussian model of optical flow (GMOF). Gaussian Mixture Model (GMM) is adopted to acquire the behavior of crowd features mined from the optic flow. Sampling the violent regions by means of a multi-scale scanning window, violence is detected. Histogram of Optical Flow (OHOF) descriptor is fed into a linear Support vector machine which classifies the event as violence or nonviolence. Performance of the algorithm is validated in BEHAVE, CAVIAR, and Crowd Violence dataset and an accuracy of 88.78%, 89.68%, and 86.59% is observed.

In [27], Optical flow and Harris 3D spatial temporal interest point detector are combined to detect violence in a frame of a video. Harris 3D considers base data as regions where both temporal and spatial domain changes. Pyramid LK captures the large motion whereas Lucas-Kanade optical flow algorithm captures the small motion in the video. These algorithms define the object movement intensities in a particular duration. Proposed method is testified using C270 Logitech camera. Motion intensity is estimated using motion coefficient. Threshold value depicts the violence occurrence precisely.

The work proposed in [28] presented a new Histogram of optical flow magnitude and orientation (HOMO) feature descriptor where optical flow between two successive frames is calculated. Six binary indicators that reflects orientation and magnitude deviations between consecutive frames is obtained. Histogram of binary indicators is combined to form the HOMO descriptor used to train the SVM classifier to detect violence or non-violence. Accuracy of 89.3% and 76.3% is observed in Hockey dataset and Violent flow dataset.

Zhou P et al. [29] proposed a violence detection using low level features. Based on optical flow fields regions with motion are segmented. In motion regions, dynamics and appearance of violent actions are mined using the low-level feature descriptors, what are Local Histogram of Optical Flow (LHOF) and Local Histogram Oriented Gradient (LHOG). Mined features are coded using Bag of Words (BoW) model. Support Vector Machine (SVM) classifies the vector as violence or not.

Ismael Serrano et al. [30] proposed Hough Forests model, provides for each class a weighted image by considering the relevant motion parts eliminating the noise and static background. Representative image is obtained from video sequence by accumulating frames associated with the temporal position.

A 2D convolutional Neural Network classifies the image frame as violence or non-violence. The proposed method is validated in Hockey Dataset and an accuracy of 94.6%is acquired.

The proposed model in [31] captures the spatial information using Convolutional Neural Network and temporal information using LSTM. An updated CNN(VGG19) where an additional dense layer is augmented to the final output layer is used as an alternative of adding global average pooling layer to the output layer. It acts as the spatial feature extractor to the LSTM cells. An accuracy of 96.33% is observed in Hockey Dataset using the above framework.

3. Proposed Model

In the proposed model, Raw videos are preprocessed and fed to Convolutional Network to obtain spatial information from the video frames. After obtaining spatial information the videos are further processed using LSTM (Long Short Term Memory) to analyze the temporal information in the video.

3.1. Dataset Used

Dataset used for implementation of the proposed model is "Hockey Fight Dataset". The dataset contains total of 1000 videos gathered from NHL (National Hockey League). Each clip is around 1.6 to 1.96 length in seconds. The dimension of the video segments is 720x576. Each image frame has a resolution of 360 x 288. Annotations are done in video level. Each video consisting of almost 50 frames are classified into fight and non-fight.

3.2. Preprocessing

The dataset comprises videos captured in hockey stadium. Video sequence consists of set of frames. The image frames are mined from the video and preprocessed before giving to the neural network. The image frames are initially in BGR (Blue Green Red) format, are then converted to RGB (Red, Green, Blue) format for further processing. Fig 1 shows the steps involved in preprocessing stage.

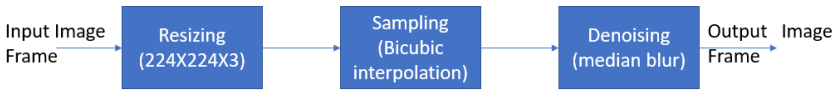

Figure 1: Preprocessing

3.2.1. Sampling

Sampling denotes the resizing of image. By sampling, a new image with high pixel can be attained with no loss of image quality. All image frames in video are transformed to a dimension of (224,224) which is the input shape of conv layer 1 of VGG19 model. The frames are up sampled using "Bicubic Interpolation" and considers only (4 x 4) 16 pixels in neighborhood at a time. It preserves fine details about the frames. Images resampled using Bicubic interpolation are smoother and have only few interpolation artifacts, yielding extensively better results.

3.2.2. Denoising

The noise present in the frames of the image, reduces the clarity of video which in turn affects the model performance. Denoising is done by using Median blur and is used for smoothing the frames. It smoothens the image using median filter with the kernel size of (3 x 3) aperture. Here, the central pixel

is replaced with the median of all pixels in the kernel window. It eliminates noise from the frame preserving the edges. After removing the noise image data is normalized using the highest value of the pixel data and fed to the convolutional neural network.

3.3. CNN

CNN architecture, VGG-19 trained on ImageNet dataset is used to train the model. VGG Network is a deeper network with small filters. VGG-19 architecture has 19 layers and a small filter of size 3 x 3 conv with periodic pooling all over the network model. It has 16 convolutional layers and 3 fully connected layers. The starting input layer has an image of size 224 x 224 with depth 3. Layer 1 and 2 of CNN Conv2D is of depth 64. Depth represents the number of filters used for generating feature map. Each filter corresponds to different pattern in the input convolving around the image and generate feature maps. Dot products of the kernel with each filter produces the feature map. Rectified linear Unit-ReLu is used in Convolutional 2D layer to make the model classify better and to improve computational time. MaxPooling2D is used as the pooling layer in the subsequent layer. Max pooling (2 x 2) with stride [2 2] and padding [0 0 0 0] has been adapted. Convolutional layer of depth 128 has been used in the next two consecutive layers. ReLu activation function is used in this layer. 128 feature maps are generated by this layers at each level. The succeeding layer is the MaxPooling2D layer with pool dimension 2 × 2. Convolutional 2D layer of depth 256 has been used as the next four layers of the network. Max pooling is applied on the final convolutional layer of depth 256. The resulting feature map is given as input to the next convolutional 2D layer of depth 512. Successive 4 layers are convolutional layers of depth 512. In the last layer Max pooling of stride 2 is applied. The resultant feature map is given to the next layer which has process similar to the last 4 layers of convolutional network of depth 512. After applying Max pooling the output is flattened to generate a 1D feature vector. The 1D feature vector is given to the Fully connected dense layer. Two Fully connected layers are adapted in this architecture. It has huge parameters because of dense connection. To get better results, Ensembling is done on the features of the final fully connected layer before going to the 1000 ImageNet classes. The Fully connected layer (FC2) of dimension 4096 is used for feature extraction as it represents the feature well. Global average pooling is applied to the output of the last convolutional block, and hence the final result of the model will be a 2D tensor. Fig 2 shows the architecture of VGG-19.

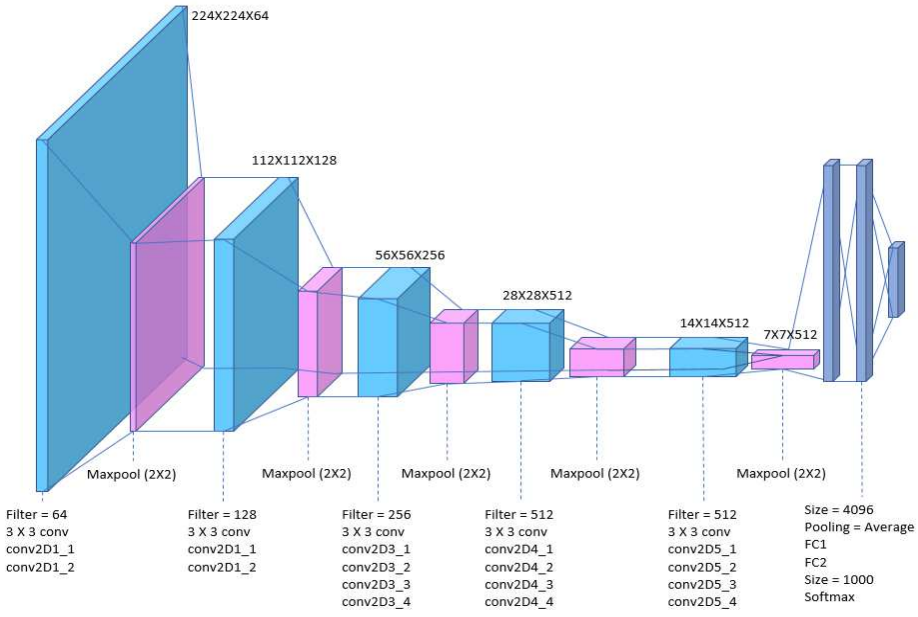

Figure 2: VGG-19 Architecture

3.4. LSTM

The extracted features are given as input to the Long Short Term Memory (LSTM) network as sequences. 20 frames per second are extracted from the video in the proposed model. Features extracted from these 20 frames are given to the LSTM layer at a time. The LSTM remembers the values over specific time intervals which helps our model to reminisce temporal features while making the required analysis on the given video.

Sequential model is created using LSTM. It comprises Dense layers and Activation layers. Initially to the model, the dense layer of 1024 classes is added with which the data is categorized. ReLU activation function is applied to the dense network. This enables the model to learn quicker and perform well by overcoming the vanishing gradient problems. The subsequent layer is a dense layer of 50 classes. Model uses Sigmoid function as an activation function in the next layer as it is very efficient. It is a probabilistic approach whose value ranges in between 0 to 1. Since the range is minimal the prediction would be more accurate. The last layer is the dense layer with 2 classes which predicts if the violence is there or not using the "softmax" activation function. It normalizes the output for each class amid 0 and 1, and divides by their sum providing the probability of the input video to have violence or not. The model uses 'Adamax' as the optimization function and 'Mean squared error' as a loss function. Figure 3 demonstrates the architecture of the LSTM model employed.

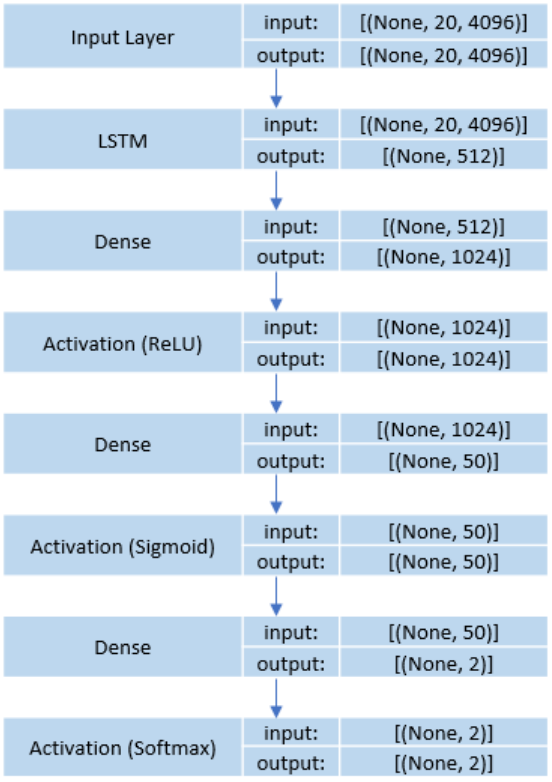

Figure 3: LSTM Architecture

4. Experimental Results

The proposed model attained an accuracy of 98 %, on the Hockey Fight Dataset. The accuracy is determined by evaluating how well the model detects violence or non-violence correctly. It is calculated using the formula

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP -> True Positive,
TN -> True Negative,
FP -> False positive
FN -> False negative

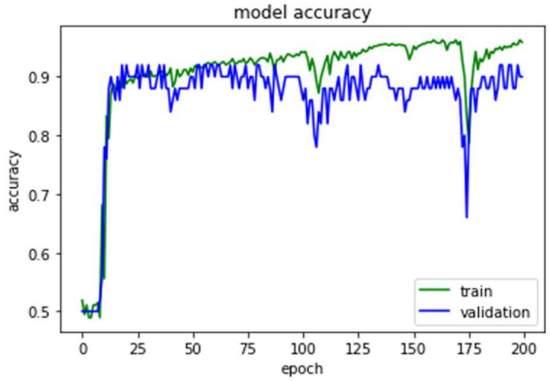

Figure 4: Training accuracy versus validation accuracy of the proposed model

The 'Mean squared error' loss of the model on the Hockey Fight Dataset is shown below. Mean squared error Loss observed using this model is 0.02.

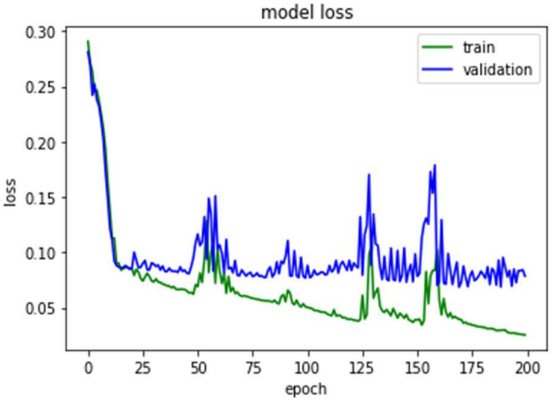

Figure 5: Training accuracy versus validation accuracy of the proposed model

5. Conclusion

In this paper, violence is detected in videos using modified Convolutional network and LSTM model. Videos are pre-processed by converting the image frames into the RGB format and resampling the video frame to the size(224x224x3). Median blur denoising is applied to the frames to remove the noise. The resultant preprocessed sequence of image frames is fed to the Convolutional neural network which uses VGG-19 architecture with global average pooling. Spatial information is learned using features extracted CNN. Extracted features are given to LSTM by which the temporal information about the video sequence are known. The proposed technique delivers an accuracy of 98 and mean square error of 0.02.

6. Acknowledgements

Mrs. J.V. Vidhya, is working as Assistant Professor and pursuing Ph.D in the Department of Computer Science and Engineering at SRM Institute of Science and Technology. Her research interests include Video Image Processing, Machine learning and Deep learning.

Dr. R. Annie Uthra is currently working as Associate Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology. Additionally, she serves as the Adjunct Associate Teaching Professor in the Institute for Software Research in the School of Computer Science at Carnegie Mellon University, Pittsburgh, USA. A graduate of SRM University's Master of Engineering in Computer Science and Engineering program, and has received Ph.D Degree from SRM University. Her research interest includes wireless sensor networks, Machine learning, Positioning and Navigation, IoT, Energy Aware Routing Techniques.

7. References

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[2] 1R. Poppe, "A survey on vision-based human action recognition," Image Vis. Comput., vol. 28, no. 6, pp. 976–990, 2010.Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).

[3] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," Computers, vol. 2, no. 2, pp. 88–131, 2013.

[4] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim, ''Fast fight detection,'' PLoS ONE, vol. 10, no. 4, Apr. 2015, Art. no. e0120448.

[5] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, ''Fast violence detection in video,'' in Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP), vol. 2, Jan. 2014, pp. 478–485.

[6] Barrett, D.P., Siskind, J.M.: Action recognition by time series of retinotopic appearance and motion features. IEEE Trans. Circuits Syst. Video Technol. 26(12), 2250–2263 (2015).

[7] Rodriguez, M., et al.: One-shot learning of human activity with an MAP adapted GMM and simplex-HMM. IEEE Trans. Cybern. 47(7), 1769–1780 (2017).

[8] Zhang, T., et al.: Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans. Circuits Syst. Video Technol. 27(3), 696–709 (2017).

[9] Wang, S., et al.: Anomaly detection in crowded scenes by SL-HOF descriptor and foreground classification. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE (2016).

[10] L. Tian, H. Wang, Y. Zhou, and C. Peng, ''Video big data in smart city: Background construction and optimization for surveillance video processing,'' Future Gener. Comput. Syst., vol. 86, pp. 1371–1382, Sep. 2018.

[11] C. Dhiman and D. K. Vishwakarma, ''A review of state-of-the-art techniques for abnormal human activity recognition,'' Eng. Appl. Artif. Intell., vol. 77, pp. 21–45, Jan. 2018.

[12] P. Zhou, Q. Ding, H. Luo, and X. Hou, ''Violent interaction detection in video based on deep learning,'' J. Phys., Conf. Ser., vol. 844, no. 1, 2017, Art. no. 12044.

[13] S. Chaudhary, M. A. Khan, and C. Bhatnagar, ''Multiple anomalous activity detection in videos,'' Procedia Comput. Sci., vol. 125, pp. 336–345,Jan. 2018.

[14] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, ''A new method for violence detection in surveillance scenes,'' Multimedia Tools Appl., vol. 75, no. 12, pp. 7327–7349, 2016.

[15] M. Alvar, A. Torsello, A. Sanchez-Miralles, and J. M. Armingol, ''Abnormal behavior detection using do minant sets,'' Mach. Vis. Appl., vol. 25, no. 5, pp. 1351–1368, Jul. 2014.

[16] Laptev I, Lindeberg T. Space-time interest points. In: 9th International Conference on Computer Vision, Nice, France. IEEE conference proceedings; 2003. p. 432–439.

[17] De Souza FDM, Cha´vez GC, Valle E, de Albuquerque Arau´jo A. Violence Detection in Video Using Spatio-Temporal Features. In: SIBGRAPI; 2010.Poker-Edge.Com, Stats and analysis, 2006. URL: http://www.poker-edge.com/stats.php.
[18] Yu Chen M, Hauptmann A. MoSIFT: Recognizing Human Actions in Surveillance Videos; 2009.
[19] Bermejo Nievas E, Deniz Suarez O, Bueno Garcı´a G, Sukthankar R. Violence detection in video using computer vision techniques. In: Computer Analysis of Images and Patterns. Springer; 2011. p. 332–339.
[20] Xu L, Gong C, Yang J, Wu Q, Yao L. Violent video detection based on MoSIFT feature and sparse coding. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE; 2014. p. 3538–3542.
[21] P. Bilinski, F. Bremond, Human violence recognition and detection in surveillance videos. in 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (IEEE, 2016), pp. 30–36.
[22] E.B. Nievas, O.D. Suarez, G.B. Garca, R. Sukthankar, Violence detection in video using computer vision techniques. in International Conference on Computer Analysis of Images and Patterns (Springer, Berlin, Heidelberg, 2011), pp. 332–339
[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. Adv. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, 2012, pp. 1097–1105.
[25] J. Huang and S. Chen, "Detection of violent crowd behavior based on statistical characteristics of the optical flow," 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, 2014, pp. 565-569, doi: 10.1109/FSKD.2014.6980896.
[26] Zhang, T., Yang, Z., Jia, W. et al. A new method for violence detection in surveillance scenes. Multimed Tools Appl 75, 7327–7349 (2016). https://doi.org/10.1007/s11042-015-2648-8
[27] Y. Lyu and Y. Yang, "Violence Detection Algorithm Based on Local Spatio-temporal Features and Optical Flow," 2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, Wuhan, 2015, pp. 307-311, doi: 10.1109/ICIICII.2015.157.
[28] Mahmoodi, Javad & Salajeghe, Afsane. (2019). A classification method based on optical flow for violence detection. Expert Systems with Applications. 127. 10.1016/j.eswa.2019.02.032.
[29] Zhou P, Ding Q, Luo H, Hou X (2018) Violence detection in surveillance video using lowlevel features. PLoS ONE 13(10): e0203668.https://doi.org/10.1371/journal.pone.0203668
[30] I. Serrano, O. Deniz, J. L. Espinosa-Aranda and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," in IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 4787-4797, Oct. 2018, doi: 10.1109/TIP.2018.2845742.
[31] A. R. Abdali and R. F. Al-Tuma, "Robust Real-Time Violence Detection in Video Using CNN And LSTM," 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, 2019, pp. 104-108, doi: 10.1109/SCCS.2019.8852616.