Violence Recognition from Videos using Deep Learning Techniques

Mohamed Mostafa Soliman
Scientific Computing department, Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
mohamed.mostafa.std5@cis.asu.edu.eg

Mohamed Hussein Kamal
Scientific Computing department, Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
mohamadhussein@cis.asu.edu.eg

Mina Abd El-Massih Nashed
Scientific Computing department, Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
mina.abdelmassih@cis.asu.edu.eg

Youssef Mohamed Mostafa
Scientific Computing department, Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
youssefmostafa@cis.asu.edu.eg

Bassel Safwat Chawky
Scientific Computing department, Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
bassel.safwat@cis.asu.edu.eg

Dina Khattab
Scientific Computing department, Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
dina.khattab@cis.asu.edu.eg

***Abstract*—Automatic recognition of violence between individuals or crowds in videos has a broad interest. In this work, an end-to-end deep neural network model for the purpose of recognizing violence in videos is proposed. The proposed model uses a pre-trained VGG-16 on ImageNet as spatial feature extractor followed by Long Short-Term Memory (LSTM) as temporal feature extractor and sequence of fully connected layers for classification purpose. The achieved accuracy is near state-of-the-art. Also, we contribute by introducing a new benchmark called Real- Life Violence Situations which contains 2000 short videos divided into 1000 violence videos and 1000 non-violence videos. The new benchmark is used for fine-tuning the proposed models achieving a best accuracy of 88.2%.**

Keywords—violence recognition, deep learning, VGG-16, LSTM, fine-tuning.

I. INTRODUCTION

Violence as defined by the World Health Organization (WHO) is " the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either result in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation" [1]. Violence surrounds us and it has many types like self- directed violence, collective violence, warfare, non-physical violence and interpersonal violence. The latter is the kind of individuals or crowds fighting with fists, knives or sticks and it is the type of violence targeted by this research. Most traditional ways of violence recognition that are based on human attention i.e. traditional surveillance video systems, are not effective for many reasons. These include high salaries for the human watching camera's feed, human mistakes of not noticing violence actions when they occur which makes traditional ways unreliable to detect violence actions. Having an automated system with the ability to recognize the occurrence of violence in videos with real-time response will enable authority holders to increase safety and take appropriate decisions.

Before the deep learning era due to a shortage of data and relatively low computational power, classical computer vision methods were proposed for violence recognition [2, 3and 4]. These methods were based on the preprocessing phase such as detection and recognition of humans, followed by feature detection and extraction from detected human movements and ends with a classification phase that determines if the video has violent action or not. These previously mentioned phases have their own algorithms and setups. Also, these phases are specific to each problem. After the prosperity of deep learning, the amount of data massively increased and computational power gave the ability to build deep neural networks with a massive number of parameters. This allows deep learning to deal with the previously mentioned classical phases as a single entity in which the deep model receives input data, learns by itself to find appropriate features and gives the output(s) without the need to go through many steps. It also allows for better generalization.

Go through many steps. It also allows for better generalization.

This paper proposes a method to tackle the problem of violence recognition from videos based on deep learning methods. The following is a summary of the paper's contributions:

- A new model is proposed which takes a small video and extracts its RGB frames which are fed as input to an end-to-end deep neural network. The network structure consists of the convolutional layers of VGG-16 [5] that acts as a spatial feature extractor. These aggregated spatial features are fed into Long Short-Term Memory [6] that acts as a temporal feature extractor followed by a stack of fully connected layers for classification. The proposed model is trained on three public benchmarks of violence datasets including hockey fight [2], movie [2] and violent- flow [7] datasets.
- A new benchmark called Real-Life Violence Situations1 (RLVS) is contributed which consists of 2000 videos; divided as 1000 violence videos and 1000 non-violence videos. Unlike previous benchmarks, the new dataset includes videos with a high variety in gender, race and age which are collected from different categories.

The RLVS benchmark is used to fine-tuning the proposed model and makes it more reliable to real-life situations. The final achieved accuracies of the fine-tuning process are 86.2%, 88.2% and 84.0% on hockey fight, movie and violent-flow datasets respectively.

978-1-7281-3995-119$31.00 ©2019 IEEE

The remaining of the paper is organized as follows: Section II provides a survey on related works of violence recognition based on deep learning. Section III explains the proposed model. Section IV presents the used datasets, experiments and achieved results. Finally, section V provides our conclusions and future work.

II. RELATED WORK

Many researchers proposed several techniques to contribute to the violence recognition problem either by using classical computer vision techniques [2, 3and 4] or deep learning-based methodologies [8, 9, 10, 11, 12,and 13]. The following introduces a summary of the state-of-the-art methodologies that use deep learning techniques since they are more related to the proposed method.

Sudhakaran et al. [8] used AlexNet [14] as a spatial feature extractor and a ConvLSTM to extract temporal features. They used fully connected layers for classification. Their model achieved an accuracy of 97.1% on the hockey fight dataset, 100% on movie dataset and 94.57% on a violent-flow dataset. Zhou et al. [9] constructed a FightNet to represent complicated visual violence interaction. They used three kinds of input i.e. RGB images for spatial networks, optical flow between consecutive frames and acceleration images for temporal networks. Their method achieved an accuracy of 97% on the hockey dataset and 100% on the movie dataset.

Serrano et al. [10] assumed that the video sequence can be summarized in one image. The feature extraction step aims to obtain a representative image from each input video sequence. A 2D Convolutional Neuronal Network (CNN) was used to classify the representative image and obtain the final decision for the sequence. Their method achieved an accuracy of 94.6% on the hockey dataset and 99% on the movie dataset. Keçeli et al. [11] computed the optical flow to input video frames using the Lucas-Kanade method [15, 16], then several 2D templates were constructed with overlapping optical flow magnitudes and orientations. These templates were fed to a pre-trained CNN as input to extract high-level deep features. They used two classifiers; Support Vector Machine (SVM) [17] and subspace k- nearest neighbor [18]. Their method achieved an accuracy of 94.4% on the hockey dataset, 96.5% on movie dataset and 80.9% on violent-flow dataset respectively.

Li et al. [12] manually enhanced the MediaEval 2015[19] violence dataset labeling videos into ten subclasses. They extracted image features using VGG Net [5], GoogleNet [20] and GoogletNet4k [21] and for motion features extraction they used three trajectory-based descriptors, namely Motion Boundary Histogram (MBH), Histogram of Oriented Gradient (HOG) [22], and Histogram of Optical Flow (HOF). They finally used SVM as a classifier. Using the Average Precision metric (AP), they achieved AP of 0.275 on MediaEval 2015 without subclass and AP of 0.303 with a subclass. Dai et al. [13] trained a CNN network for violence detection, and then they adopted a specially designed two-stream CNN to extract features on both static frames and motion optical flows. Also, LSTM was applied on top of the two-stream CNN features to capture the longer-term temporal dynamics. The classification step was done using SVM and the achieved mean AP was 0.296 in the violence detection subtask.

III. PROPOSED METHOD

Fig. 1 shows the architecture of the proposed method. First pre-processing operations are applied to the input video frames. The next two consecutive stages of feature extraction are applied; a VGG-16 [5] stage which is responsible for spatial features extraction for each frame, and an LSTM [6] stage which works as temporal features extractor. Finally, the extracted features are fed to fully connected layers for classification. The following subsections highlight the detailed procedure for each phase.

A. Preprocessing

The datasets videos appear in shape of (V, F,…, 3) where V represents the number of dataset videos, F represents the number of video frames, represents the width of frame i and represent the height of frame i. These input videos go through a sequence of pre- processing operations as follows; First each frame of the input videos is resized into 224×224×3. Two data augmentation techniques are used to increase the size of the data e.g. blurring, vertical flipping and adding vertical lines. Fig. 2 shows a sample of the applied data augmentation process. Afterward, the hockey fight [2], movie [2] and violent-flow [7] datasets are shuffled separately and each dataset is split into 80% for training and 20% as a validation set. The output size of this step is (V, F, 224, 224, and 3).

B. Feature Extraction

In this step two kinds of features are extracted consecutively; the first features set consist of the frame spatial features which are extracted using only the convolution layers of a pre-trained VGG-16 [5] on ImageNet dataset [23]. The total parameters used from VGG-16 are 117,479,232 which are non-trainable parameters. The output size after this step is (V, F, 4096) where 4096 is the number of features extracted from each frame. Due to the violent action between humans are distributed along a sequence of frames, so LSTM [6] is used to extract temporal features as the second set of features to keep track of the changes along time. The used setup of LSTM consists of 128 units as the dimensionality of the output space and the tanh function as an activation function. Also, the hard sigmoid function is used as recurrent activation, glorot uniform‟ is used as a kernel initializer and the initial values of bias are set to zero. The output size of this step is (B, 128) where B is the batch size. A batch normalization layer is used after the LSTM layer. The LSTM parameters count and output shape are shown in Table 1.

TABLE 1. LSTM AND BATCH NORMALIZATION LAYER‟S PARAMETER

Layer Name	Output Shape	Parameters
Lstm_1	(None, 128)	2163200
Batch_normalization_1	(None, 128)	512

***None** word in Table 1 refer to batch size

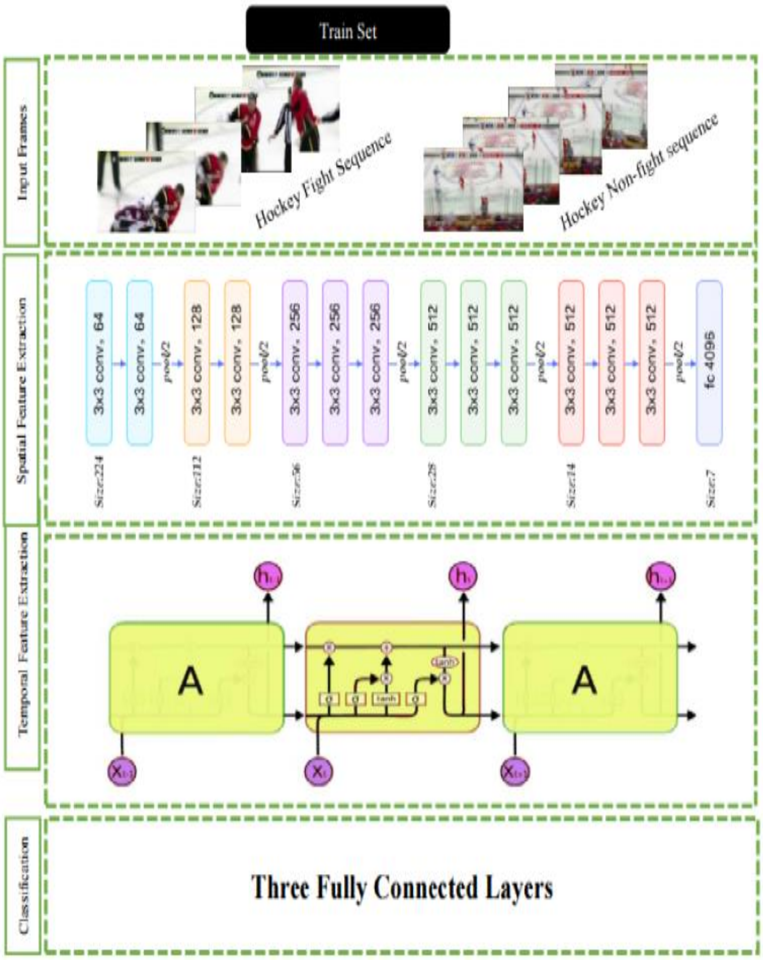
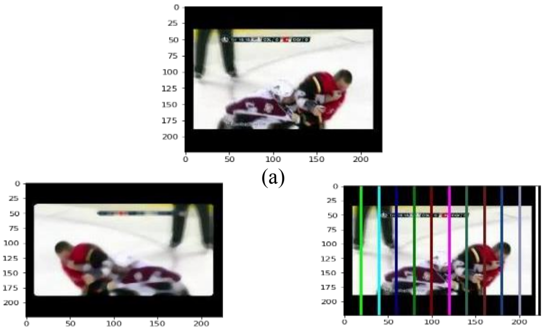


Fig. 1. Detailed block diagram of the proposed model. The input sequence is fed into convolution part of VGG-16 net with original weights. Aggregated Spatial features are fed into LSTM which keeps track the changes in input sequence over time. Finally, a group of connected layers are used for classification.


(a)

(b) (c)

Fig. 2. Data augmentation techniques used (a) original image, (b) Original image after applying median blurring and vertical flipping and (c) Original image after adding vertical colored lines.

C. Classification

In this step a sequence of three fully connected layers are used for the classification goal, where the first layer has 2048 neurons, the second layer has 1024 neurons and the last layer contains only 2 for violence and nonviolence classes.

The first and second layers used Rectified Linear Unit (RELU) activation function as in(1), while the last layer uses the soft-max activation function as in (2). The neurons are initialized using the Xavier uniform initializer.

To reduce the effect of over fitting two methods are used; first method is L2 norm which is used as regularization parameter in all fully connected layers with λ = 0.15 where λ is the regularization parameter.
In the second method the drop out is applied after the first connected layer only with probability of 20%.

$$f(x) = Max(0, x) \qquad (1)$$

$$f(x_i) = \frac{e^{xi}}{j\sum e^{xj}} \qquad (2)$$

IV. EXPEREMNTAL RESULTS

The proposed model is evaluated against three of the state-of-the-art benchmarks datasets including hockey fight [2], violent flow [7] and movie [2] datasets. In addition, a newly constructed benchmark of Real-Life Violence Situations (RLVS) is used for both testing and fine-tuning of the proposed model.

A. Datasets

1) The Hockey Dataset [2]: consist of 1000 videos divided into 500 violence and 500 non-violence videos. It was collected from hockey games of the National Hockey league where each video consists of 50 frames so that each frame has a size of 720 × 576. All videos share the same background, while Ice hockey players only appear in all videos. In the experiments performed, 20 frames were taken (after the tenth frame) from each clip as input to our proposed model to increase the probability that all the taken frames have violence actions.

2) Movie Dataset [2]: consists of 200 videos divided into 100 violence and 100 non-violence videos. The violence videos collected from movie scenes, while the non-violence videos were collected from other actions. Unlike hockey dataset, the movie dataset has different backgrounds. In the experiments performed, 15 frames were taken starting from the first frame as input to our proposed model.

3) Violent-Flow dataset [7]: consists of 246 videos which contain crowd scenes of fight between persons. The videos were collected from violent situations occurred in football matches. In the experiments performed, 20 frames were taken (after tenth frame) as input to our proposed model.

4) Real-Life Violence Situations (RLVS): Due to existing disadvantages in the previous datasets such as including the same environment (hockey fight dataset), having few numbers of videos and bad resolution videos (Movie and Violent Flow datasets), a new benchmark is created which aims to enhance all the previously mentioned flaws in all datasets. The RLVS benchmark consists of 2000 videos divided into 1000 violence clips and 1000 non-violence clips.

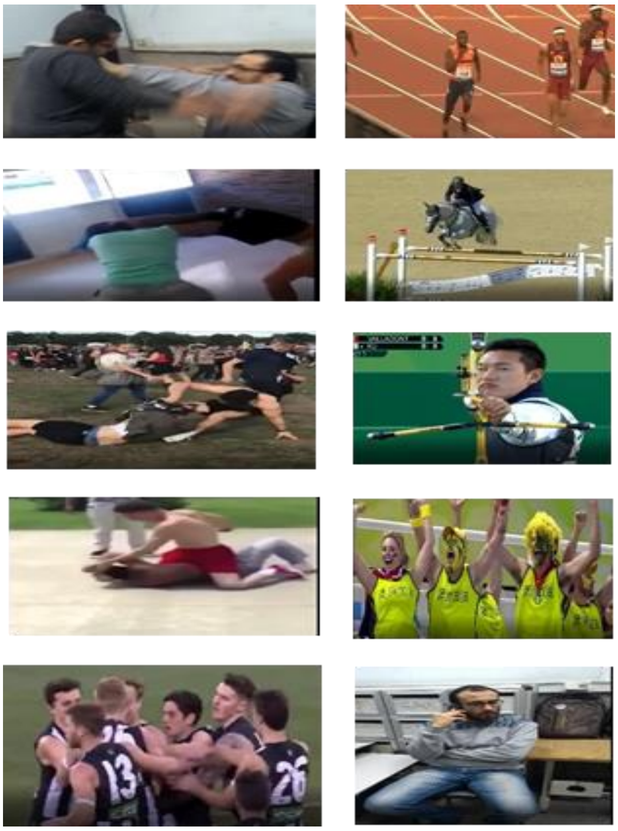
The violence clips involve fights in many different environments such as street, prison and schools. The non-violence videos contain other human actions such as playing football, basketball, tennis, swimming and eating.

Part of the RLVS dataset videos are manually captured, however to prevent the redundancy in persons and environment in the captured videos, other videos are collected from YouTube.

Long videos are cut into short length videos with maximum duration of 7 seconds, minimum duration of 3 seconds and average duration of 5 seconds. The collected videos are considered to have high resolution (480p – 720p) and to include a variety of people in race, age, and gender with different environments. The collected video frame width ranges between 224 and 1920, while the height of the frames ranges between 224 and 1080 with average video size of 397 × 511. Fig. 3 shows sample snips from the created dataset.

B. *Environment settings*

The proposed model is written in python using Keras library [24] with TensorFlow [25] backend and some helper libraries like OpenCV [26] and matplotlib [27]. The Stochastic Gradient Descent optimizer [28, 29, and 30] is used with learning rate equal to 0.06 without momentum. The categorical cross-entropy is used as the loss function to the proposed model. The batch size is set to 100 and the model is trained along 2000 epochs. Regarding the hardware used; the system is run using Kaggle [31] which have GPU NIVIDIA K80, and 2 CPU"s, 14 GB RAMs and 5 GB as hard drive.



(a) (b)

Fig. 3. Samples from the RLVS dataset, (a) violence samples and (b)non-violence samples.

C. *Results and Discussions*

In this section, the model performance and the dataset generalization capabilities are discussed. In the first experiment, three separate models are trained on Hockey Fight dataset M (H), Movie dataset M (Mov) and Violent- flow M (V), respectively.

Additionally, a 5-fold cross validation technique is applied; for each fold, the accuracy and loss are calculated and then the average accuracy is calculated for all folds. Table 2 compares the results of the validation accuracy for the proposed model of the three trained models (last row) and the state-of-the-art techniques on the three datasets. This comparison shows that the accuracies of the proposed models are comparable with the state-of-the-art techniques.

TABLE 2. ACCURACY COMPARISON BETWEEN THE PROPOSED MODELS AND THE STATE-OF-THE-ART TECHNIQUES

Method	Hockey fight	Movie	Violent-Flow
Sudhakaran et al [8]	97.01%	100.0%	94.57
Zhou et al [9]	97.00%	100.0%	-
Serrano et al [10]	94.60%	99.0%	-
Keçeli et al [11]	94.40%	96.5%	80.90%
Proposed	95.1%	..99%	90.01%

To test the generalization of the proposed models, two different sets of experiments are performed. The first experiment is based on testing the trained models M(H), M(Mov) and M(V) using the test set of the RLVS benchmark. Table 3 shows the achieved test accuracy of that experiment. The violence accuracy and non-violence accuracy in Table 3 refer to the percentage of videos that are correctly classified from each class with respect to each class videos.

TABLE 3. TEST ACCURACY ACHIEVED BY PROPOSED TRAINED MODELS ON THE RLVS TEST SET

Model	Violence accuracy	Non-violence accuracy	Overall Accuracy
M(H)	67.0%	31.7%	49.3%
M(Mov)	94.0%	12.3%	53.2%
M(V)	73.0%	70.0%	71.5%

One intuition is that a good dataset shall allow the model to efficiently learn to separate the violence and non-violence classes and to generalize in different domains. As shown in Table 3, it can be observed that the violence accuracy is relatively higher than the non-violence accuracy. Furthermore, the overall test accuracy is much lower than the validation accuracy achieved in Table 2 for all models. This indicates the inability of all models to generalize to other sequences of fight in other domains which is referred to over-fitting problem. This can be due to several reasons. Regarding the M (H) model, the nature of repeated fight pattern between persons and the stability of the background between the violence and non-violence classes indicate that the fast motion and the type of interaction between persons are considered the important features which help the network to identify such features in player's movements. Moreover, the small number of trained videos in all datasets allows the models to lose the ability to generalize.

To prove that the RLVS dataset can give the model the ability to generalize, in the second set of experiments, a new model M(RLVS) is trained and validated using the RLVS benchmark with an achieved training and validation accuracies of 99.9% and 94.5% respectively. The model is then tested using the other Hockey, Movie and Violent-Flow datasets without any further fine-tuning. Table 4 presents the test accuracies achieved by this test experiment.

By comparing the results of the overall accuracies in Tables 3 and 4, it is clear that the trained model on the RLVS dataset achieved better results in terms of generalization. Regarding the bad accuracy (5.2%) of non- violence accuracy in hockey dataset is due to that the non- violence videos of the RLVS dataset suffer from shortage of hokey actions.

TABLE 4. TEST ACCURACY ACHIEVED BY M(RLVS) ON ALL OTHER DATASETS

Dataset	Violence Accuracy	Non-Violence Accuracy	Overall Accuracy
Hockey	99.2%	5.2%	52.2%
Movie	66.9%	88.0%	77.6%
Violent-Flow	85.2%	60.0%	73.6%

D. Transfer Learning [32] (Fine-tuning)

In order to achieve more enhancements to the proposed model, the final experiment performs a fine- tuning process to the three models M(H), M(Mov) and M(V) by using the training set from the RLVS dataset (1200 videos). Some changes of the proposed model hyper-parameters are applied including decreasing the learning rate from 0.06 to 0.02 and the number of epochs to 400. Table 5 illustrates the achieved accuracy results of testing the fine-tuned model with the test set of the RLVS dataset.

TABLE 5. ACCURACY RESULTS OF FINE-TUNING MODEL

Model	Violence Accuracy	Non-Violence Accuracy	Overall Accuracy
M(H)	88.33%	84.0%	86.166%
M(Mov)	93.33%	83.0%	88.2%
M(V)	90.0%	78.0%	84.0%

As seen from Table 5, the performance of all models has been enhanced after fine-tuning. The over-fitting problem is eliminated by decreasing the high variance of accuracy between the violence and the non-violence classes. The overall accuracy is still comparable to the state-of-the-art.

V. CONCLUSION

In this research, we contribute in violence recognition problem by proposing an end-to-end deep learning model. The proposed model consists of the convolution part of VGG-16 which is pre-trained on ImageNet as spatial feature extractor, followed by LSTM as temporal feature extractor. Finally, the proposed model ends with a sequence of three fully connected layers for classification purpose. Three public datasets widely used by literature including Hockey Fight, Movie and Violent-flow are being tested. The performance of the proposed models on these datasets is comparable to current literature.

Another contribution is the new benchmark named Real Life Violence Situations (RLVS) which is considered the so far largest dataset for violence videos. It consists of 2000 videos divided into 1000 violence and 1000 non-violence videos. The proposed dataset has a variety in violence situations. Also, the non-violence class involves many classes from regular human actions. The new benchmark is used as test set for the previously proposed models. A fine-tuning process is applied to benefit from the knowledge gained, solve over-fitting and non-generalization problem and to enhance the testing accuracy of all models.

VI. REFERENCES

[1] https://www.who.int/violenceprevention/approach/definition/en/

[2] E. B. Nievas, O. D. Suarez, G. B. Garc, and R. Sukthankar, “Violence detection in video using computer vision techniques,” in International conference on Computer analysis of images and patterns , pp. 332–339, Aug 2011.

[3] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, “Violent video detection based on MoSIFT feature and sparse coding,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing 2014, pp. 3538–3542, May 2014.

[4] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, “Violence detection using oriented violent flows,” Image and vision computing, vol. 48, pp. 37–41, Apr 2016.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, Sep 2014.

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[7] T. Hassner, Y. Itcher, and O. Kliper-Gross, “Violent flows: Real-time detection of violent crowd behavior,” in 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6, June 2012.

[8] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6, Aug 2017.

[9] P. Zhou, Q. Ding, H. Luo, and X. Hou, “Violent interaction detection in video based on deep learning,” in Journal of Physics: Conference Series, vol. 844, no. 1, p. 12044, June 2017.

[10] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, “Fight recognition in video using hough forests and 2D convolutional neural network,” IEEE Transactions on Image Processing , vol. 27, no. 10, pp. 4787–4797, June 2018.

[11] AS. Keceli, A.kaya, “Violent activity detection with transfer learning method,” Electronics Letters, vol. 53, no. 15, pp. 1047– 1048, June 2017.

[12] X. Li, Y. Huo, Q. Jin, and J. Xu, “Detecting Violence in Video using Subclasses,” in Proceedings of the 24th ACM international conference on Multimedia, pp. 586–590, Oct 2016.

[13] Q. Dai et al., "Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning.," in MediaEval , Sep 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097– 1105, 2012.

[15] B. D. Lucas, T. Kanade, and others, "An iterative image registration technique with an application to stereo vision,", 1981.

[16] B. D. Lucas, "Generalized image matching by the method of differences.," 1986.

[17] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, Sep 1995.

[18] M. Hund, M. Behrisch, and F&amp;auml, "Subspace nearest neighbor search-problem statement, approaches, and discussion," in International Conference on Similarity Search and Applications, pp. 307–313, Oct 2015.

[19] O. Seddati, E. Kulah, G. Pironkov, and S. Dupont, "UMons at MediaEval 2015 Affective Impact of Movies Task including Violent Scenes Detection.," in MediaEval, Sep 2015.

[20] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9, 2015.

[21] S. Liao, X. Li, H. T. Shen, Y. Yang, and X. Du, "Tag features for geo-aware image classification," IEEE transactions on multimedia, vol. 17, no. 7, pp. 1058–1067, May 2015.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," June 2005.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, June 2009.

[24] https://keras.io/

[25] https://www.tensorflow.org/

[26] https://opencv.org/

[27] https://matplotlib.org/

[28] H. Robbins and S. Monro, "A stochastic approximation method," The annals of mathematical statistics, pp. 400–407, Sep 1951.

[29] J. Kiefer, J. Wolfowitz, and others, "Stochastic estimation of the maximum of a regression function," The Annals of Mathematical Statistics, vol. 23, no. 3, pp. 462–466, 1952.

[30] L. Bottou, "Optimization methods for large-scale machine learning," Siam Review, vol. 60, no. 2, pp. 223–311, May 2018.

[31] https://www.kaggle.com/

[32] L. Torrey and J. Shavlik, "Transfer learning," in Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI Global, 2010, pp. 242–264.