Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 5 Issue 1, January 2023***
(https://fcihib.journals.ekb.eg)

Violence Detection Enhancement in Video Sequences Based on Pre-trained Deep Models

Yahia Reda Elkhashab[1,*], Wessam H. El-Behaidy[1]
[1]*Department of computer science, College of computer science and Artificial intelligence, Helwan University, Egypt*
yahiaelkhashab17@gmail.com, w_behaidy@yahoo.com

***Abstract* — Violence detection is one of the challenging applications in the field of Human activity recognition (HAR). The enhancement of violence detection, through surveillance cameras, will help our societies to be more controllable and safer. In this paper, a proposed two-layer deep model is built for classifying video sequences into violent and non-violent actions. The first layer extracts the space features of video frames using the pre-trained model DenseNet-121. Then, the extracted features are fed to a long short-term memory (LSTM) network. LSTM captures the temporal features by learning the dependencies between frames, which links all frames of a video as one action. The proposed model is experimentally evaluated on two datasets. The recognition rate has improved up to 96%, which is better than those of most existing similar models over the open HOCKEY dataset and up to 92% over the real-live violence situations (RLVS) dataset. The implementation of the proposed model is available at:** https://github.com/YahiaElkhasahb/Enhancing-Violence-Detection-in-Video-Sequences-Based-on-Deep-Learning-Techniques.

Keywords— Violence detection, CNN, LSTM, Deep Learning, Human Action Recognition (HAR).

I. INTRODUCTION

Violent behavior can be defined as behavior that threatens or harms others. These behaviors can be just verbal threats but they can turn into physical harm. Based on a study, conducted in 2012 by the World Health Organization (WHO) [1], 475,000 people were killed globally as a result of societal violence. Violent incidents threaten the safety of everyone and require immediate detection and response to reduce these acts. For this reason, automated monitoring and detection are recently introduced for identifying violence.

Violence detection is considered a human action recognition (HAR) problem. HAR is one of the most important fields in computer vision, which has a huge number of applications in many areas, such as health care, controlling in smart homes, gameplay controls, and bad habit detection [2].

Violence detection has two main steps; encoding action into a feature vector, and classifying action based on a feature vector [3]. Firstly, the action representation (i.e., the feature vector) could be hand-crafted features that extract local descriptors of motion and human appearance. These descriptors are used to capture the spatial and temporal (i.e., spatiotemporal) features within a short duration. The commonly used descriptors are space-time interest points (STIPs), histograms of gradients (HOG), histograms of optical flow (HOF), motion boundary histograms (MBH), and scale-invariant feature transform (SIFT). To capture long-duration features, local descriptors are first extracted, then they are tracked to perform a feature trajectory [3]. The trajectories can combine or average different descriptors between consecutive frames to get a more robust feature trajectory. Secondly, the action classifier is used. The traditional classifiers can be used like k-nearest neighbor (k-NN) and support vector machine. But they lack considering the temporal features into account. Sequential classifiers like conditional random fields (CRFs), hidden Markov models (HMMs), and structured support vector machines (SSVMs) can also be used. These classifiers consider the temporal features into account but they are sensitive to background noise.

However, the deep architectures had great success in action recognition problems to their ability to extract relative features automatically. The action representation can be one of three temporal modeling deep learning approaches; space-time networks, multi-stream networks, and hybrid networks [3]. The space-time networks extend the 2D convolutions to 3D convolutions to capture the temporal features. The multi-stream networks make a fusion of two streams; one for spatial and the other for temporal features. The hybrid networks, which are used in this paper, combine CNN models and LSTM.

For enhancing violence detection accuracy, in this paper, we propose a hybrid network of two deep layers that:

1- Extracts frames from video sequences and feeds them to a pre-trained model, as a first layer, to extract space features from each frame of the video.

2- Implements the long short-term memory (LSTM) network, as a second layer, that links all the image frames to deal with them as a single action.

Then, a classification process is performed to determine if the current state is violent or non-violent.

In section **II**, the related work is discussed, whereas section **III** clarifies the proposed model. Section **IV** specified the dataset used and the results, whereas section **V** is our conclusion section.

II. RELATED WORK

In this section, recent research papers are represented, which are related to violence detection or to human action recognition (HAR).

Tao Zhang *et al.* [4] introduced a model named the Gaussian model of optical flow (GMOF). It was intended to extract possible areas of violence and distinguish violent events by feeding the proposed novel descriptor, orientation histogram of optical flow (OHOF) into a linear support vector machine (SVM) for classification. To exhibit the overall performance of the proposed algorithm, experiments had been carried out on three public datasets: the BEHAVE dataset [5], the CAVIAR dataset [6], and the Violent Flows dataset [7]. The results were $85.29 \pm 0.16\%$ on BEHAVE dataset, $86.75 \pm 0.15\%$ on the CAVIAR dataset, and 82.79 ± 0.19 % on the Violent Flows dataset.

A. Ullah et al. [8] have proposed a technique to detect human actions in videos by utilizing frame-level deep features of the AlexNet [30] and processing it through deep neural network bidirectional long short-term memory (DB-LSTM). The proposed technique is experimentally evaluated and the results are applied to three action recognition datasets: UCF101 [9], Action YouTube [10], and HMDB51 [11]. The accuracy of the proposed technique was 92.84 % for Action YouTube, 87.64% for HMDB51, and 91.21% for UCF101.

Qing Xia et al. [12] proposed a model for violence detection which is based on SVM and a two-channel convolutional neural network (CNN). The two-channel network is trained for choosing the best features. In addition, for improving the accuracy of detection, an effective fusion between the motion and appearance data was used. The experiments were tested on two datasets: Hockey Fight [13] and Violent Flows [14]. The highest accuracy achieved was 95.9±3.53 on Hockey Fight and 93.25±2.34 on Violent Flows.

Chen et al. [15] collected data using smartphones from 30 volunteers of different ages. A long short-term memory (LSTM) network is used. The input to LSTM is the complementary output from a single-layer feedforward neural network (SLFN) and handcrafted features. The best accuracy was 97.7% which was achieved with their proposed approach.

Wang et al. [16] presented an efficient and powerful pooling scheme. The proposed framework has cast this pooling scheme to learn, from each sequence, multiple instances with useful decision boundaries on the frame-level features against noise features. In addition, the framework is extended to work with end-to-end CNN training and nonlinear decision boundaries. Multiple experiments had been executed on: HMDB-51 [11], Charades [17], and NTU-RGBD [18]. Their model proved highly efficient as it achieved 81.3% on the HMDB-51 Dataset, and outperformed the state-of-the-art approaches by 1 to 4%.

Sultani et al. [19] propose a technique to detect real-world anomalies in surveillance videos with a deep learning approach. They fed the model with a weakly-labeled dataset using a deep machine learning framework. They use a new large-scale dataset, covering 13 anomalies in the real world, using clips captured by surveillance videos. Where these types were selected based on their belief that they have a strong influence when they occur on public safety and security. The proposed model for violence detection is experimentally evaluated on that dataset and shows that it significantly outperforms the results of baseline methods which achieved 75.41%.

Huaijun Wang et al. [20] propose a scheme based on deep learning that can identify activities and their transitions. The dataset is collected and built by sensors. The approach extracted the automatic features from the data using the CNN network. Whereas, the dependency between frames is extracted using the LTSM network, plus the comparison of local features fusion and identification of the patterns of the two motions. The open HAPT dataset was used in experiments to show the overall performance of the proposed method. The proposed model made improvements in results; the recognition rate is improved the up to 95.87% and higher than 80% for transitions.

From previous literature, we can conclude that the CNN as a feature extractor with a combination of LTSM or its variants gives an improvement to violence detection. However, the recent pre-trained models like denseNet121 usually outperform the costumed CNN layers, as feature extractors in many applications.

III. PROPOSED MODEL

The proposed model is clarified in this section. The model firstly preprocesses the frames of a video and sent them to a space feature extraction network. The output of this network is passed to the long short-term memory (LSTM) network for analyzing the temporal features of the video. Finally, the models train these features for classification. Fig.1 shows the architecture of the proposed model.

A. Pre-processing

Initially, the dataset is acquired as a group of videos, where each video contains a full frameset of (27-50) frames. As a preprocessing step, the first 20 frames were extracted per video, to reduce the processing time without affecting the accuracy of the system. Furthermore, the selected frames are resized to 224x224x3 to match the input size of the subsequent stage.

B. Space Feature Extraction Network

To extract the space features of frames, different feature extraction algorithms were used. The pre-trained models VGG16 and DenseNet121 have been experimented as a feature extractor alone. Furthermore, the hand-crafted Rotated BRIEF (ORB) and Oriented FAST were used in combinati–on with pre-trained models.

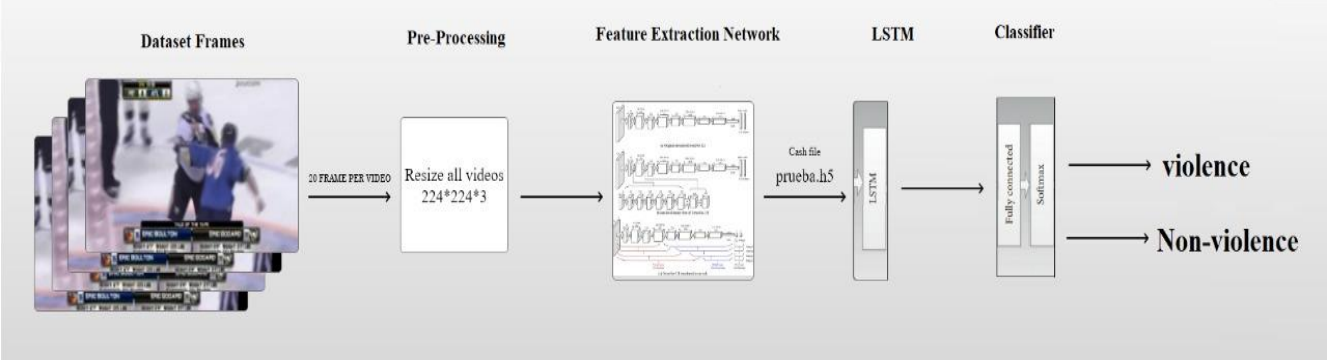

Fig. 1. The Proposed Model Architecture

i) ***Oriented FAST and Rotated BRIEF (ORB) [21]:*** ORB is a fusion based on both the FAST key point detector and the BRIEF descriptor. It is an efficient and fast alternative to SIFT [22] or SURF [23] algorithms, used to extract features in a comparison between costs and matching performance. For those reasons, ORB is chosen as our hand-crafted feature extraction.

ii) ***VGG16 Network [24]:*** is a network model with excellent classification performance in convolutional neural networks. VGG16 is chosen as it is an improvement of Alex Net, which gave good results in the work of A. Ullah et al. [8]. In this experiment, we used 200 epochs for the model fitting with a 500-batch size and an Adam optimizer.

iii) ***Dense Convolutional Network 121 (DenseNet121) [25]:*** is an architecture that focuses on going deeper into the network than the regular deep learning architectures. As well as, increasing the efficiency of the training, by using shorter connections between the layers. For that, DenseNet121 is chosen. In this experiment, we used the same number of epochs and batch size in VGG-16 while using the Adam optimizer.

Both pre-trained models are used as feature extractors, and for that, we saved the transfer values into a cache file. These values are just prior to the final classification layer of both models. These values are the extracted features that will be passed to the next stage.

C. *Long Short-Term Memory (LSTM)*

LSTM networks are a type of recurrent neural network (RNN) which able to learn order dependence in sequence prediction problems. It consists of a set of memory blocks, every memory block contains one or more than one recurrent memory cell and every cell of them contains three gates (input gate, forget gate, output gate). These memory blocks analyze the dependency between frames. During implementation, LSTM receives two parameters; RNN size (512 in our case) and input shape. The input shape consists of a number of chunks (i.e., number of frames), and chunk size (i.e., features vector size or the transfer values). In our case, the chunk size is 1024 in ORB, 4096 in VGG16, and 1024 in DenseNet121.

D. *Classification*

This is the last stage in our model that is used to train the extracted spatiotemporal features. This stage consists of three layers of a fully connected (FC) network followed by a softmax classifier. The layers of FC are of size 1024, 50, and 2 respectively. After training, the model should be able to detect violence. If violence is detected in any of the video frames, the entire video will be categorized as violent; otherwise, it will be categorized as a non-violent video.

IV. EXPERIMENTAL RESULTS

This section presents the datasets used in these experiments and the results obtained by the proposed model. In addition, a comparison between our results and other related research is performed.

A. *Datasets*

We used two datasets; the Hockey fight dataset and the Real-Life violence Situations dataset.

i) Hockey fight dataset

The hockey fights dataset [13] consists of 1000 clips that it is manually labeled as “violent” or “non-violent”. Each clip consists of 50 frames of 720×576 pixels”. They are collected from the National Hockey League (NHL)’s hockey matches, with the presence of camera motion, as in Fig. 2 This dataset is created to detect violence in sports footage. We used 800 clips for training and 200 for testing.

Fig. 2. Samples of Hockey Fight Dataset

ii) Real-Life Violence Situations dataset

RLVS or Real-Life Violence Situations [26] consists of 2000 clips, 3 hours in length, that are collected from YouTube. These clips show violent and non-violent situations in different situations and places, as in Fig. 3. These clips are divided into 1,000 violent and 1,000 non-violent action videos. The violent action videos are near-distance fights containing videos of bare hands fights, non-projectile weapons, and abuse fights using knives, sticks, and big throw-able objects. Whereas, the non-violence action videos contained videos of talking in a café, shaking hands with friends, eating, riding a horse, walking in a park, etc. Not all clips are equal but the average frames were 24-30 fps of 1920×1080 if the clip is manually captured, and 204×360 if it was added from YouTube. We used 1600 clips for training and 400 for testing.

Fig. 3 . Samples of RLVS Dataset

B. *Results and Discussion*

Two main experiments are performed on our proposed model. The two experiments differ in the space feature extraction stage; all subsequent stages of LSTM and classification are unchanged.

i) The first experiment

Here, the pre-trained models VGG16 and DenseNet121 are used as feature extractors. The extracted features from each model are passed to subsequent stages.

Using VGG16 on Hockey Fights Dataset, our model shows an accuracy of 93.90%, whereas it achieves 79.4% on RLVS Dataset. However, DenseNet121 on Hockey Fights Dataset shows an accuracy of 96.40%, whereas it achieves 92.05% on RLVS Dataset, as shown in Fig 4 and Fig. 5, respectively.

ii) The second experiment

Here, a fusion between ORB features with those extracted from the pre-trained model is performed, hopefully getting higher results.

On the Hockey Fight Dataset, using VGG16 with ORG achieves 92.0% accuracy, whereas using DenseNet121 with ORG achieves 94.49% accuracy, as shown in Fig 6 and Fig. 7, respectively.

Both accuracies are lower than what was reached in the first experiment, for that we did not apply the second experiment on RLVS Dataset.

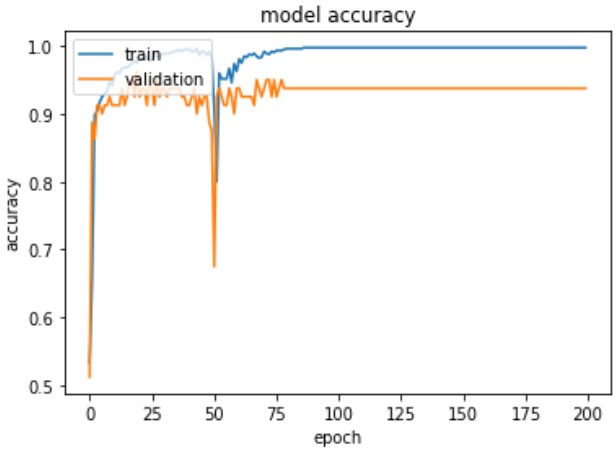

Fig. 4. DenseNet-121 Results on Hockey Fight Dataset

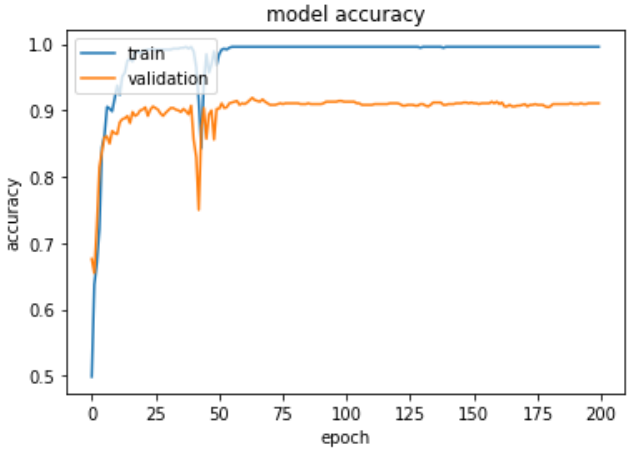

Fig. 5. DenseNet-121 Results on RLVS Dataset

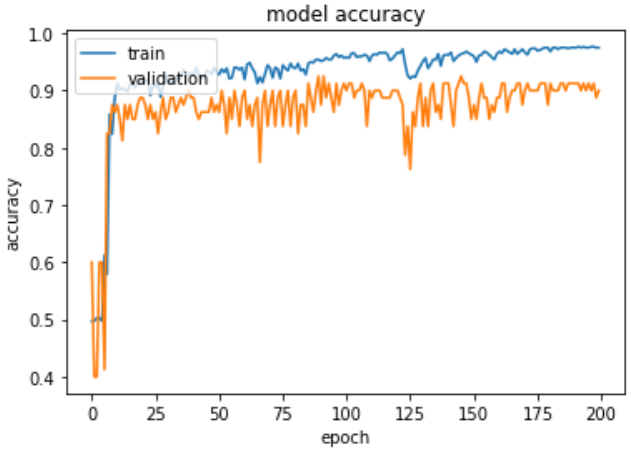

Fig. 6. VGG-16 Results on Hockey Fight Dataset

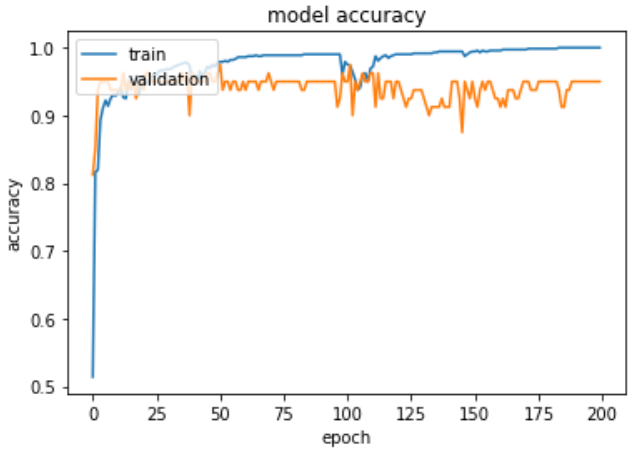

Fig. 7. ORB+DenseNet-121 Results on Hockey Fight Dataset

Table 1 and Table 2 summarize the results achieved from the different combinations on Hockey Fights Dataset and RLVS Dataset, respectively. The best accuracies were achieved by using extracted features from DenseNet121 alone on both datasets. For that, it is considered our best model.

Table 1. Performance accuracy of the proposed model on the Hockey Fight Dataset

Method	Accuracy	Loss
DenseNet-121	96.40%	3.50%
VGG-16	93.30%	5.60%
ORB+DenseNet-121	94.49%	4.84%
ORB+VGG-16	92%	7.24%

Table 2. Performance accuracy of the proposed model on the RLVS Dataset

Method	Accuracy	Loss
DenseNet-121	92.05%	7.37%
VGG-16	79.4%	16.3%

Table 3 shows a comparison between our best model and five previous related works of literature. Our model is the highest on Hockey Fight Dataset ranging from 0.5% to 3.59% improvement, with around 2% improvement on average. However, our model, on RLVS, has lower accuracy than Soliman et al. [26], and a similar accuracy to Moaaz et al. [8].

Table 3. Comparison between performance accuracy of our best model with related literature

Method	Hockey Fight Dataset	RLVS Dataset
Gao et al. [28]	92.81%	-
Bilinski et al. [29]	93.7%	-
Qing Xia et al.[12]	95.9%	-
Soliman et al. [26]	95.1%	94.4%
Moaaz et al. [27]	94.5%	92%
Our best model	96.40%	92.05%

V. CONCLUSION

We introduced a simple, efficient model for detecting violence. It is a combination of two deep learning models; DenseNet121 and LSTM. This combination reaches the spatiotemporal features required for detection and classification. The proposed model proves its efficiency on Hockey Fight dataset with a 0.82% improvement on average in comparison to other models. However, the fusion model of ORB with DenseNet121 has not reached the expected performance.

ACKNOWLEDGMENT

We deeply thank Mostafa Samir Mostafa, Nadeen Kadry Saeed, Yahia Hosam Rasmy, and Youssef Ayman Abd Ellatif for their help in the implementation of this model.

REFERENCES

[1] World Health Organization. Global status report on violence prevention 2014. World Health Organization, 2014.

[2] Wang, Huaijun, et al. "Wearable sensor-based human activity recognition using hybrid deep learning techniques." Security and Communication Networks 2020 (2020).

[3] Kong, Yu, and Yun Fu. "Human action recognition and prediction: A survey." arXiv preprint arXiv:1806.11230 (2018).

[4] Zhang, Tao, et al."A new method for violence detection in surveillance scenes." Multimedia Tools and Applications 75, no. 12 (2016): 7327-7349.

[5] S. J. Blunsden, R. B. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification" , Annals of the BMVA, Vol 2010(4), pp 1-12.

[6] https://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

[7] T. Hassner, Y. Itcher, and O. Kliper-Gross, Violent Flows: Real-Time Detection of Violent Crowd Behavior, 3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Rhode Island, June 2012

[8] Ullah, Amin, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. "Action recognition in video sequences using deep bi-directional LSTM with CNN features." IEEE access 6 (2017): 1155-1166.

[9] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 (2012).

[10] Liu, Jingen, Jiebo Luo, and Mubarak Shah. "Recognizing realistic actions from videos “in the wild”." 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.

[11] Kuehne, Hildegard, et al. "HMDB: a large video database for human motion recognition." 2011 International conference on computer vision. IEEE, 2011.

[12] Xia, Qing, et al. "Real time violence detection based on deep spatio-temporal features." Chinese Conference on Biometric Recognition. Springer, Cham, 2018.

[13] Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril,

H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011).

[14] Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: Computer Vision and Pattern Recognition Workshops, pp. 1–6 (2012)

[15] Zhenghua Chen, Le Zhang, Zhiguang Cao, and Jing Guo. "Distilling the knowledge from handcrafted features for human activity recognition." IEEE Transactions on Industrial Informatics 14, no. 10 (2018): 4334-4342.

[16] Wang, Jue, et al. "Video representation learning using discriminative pooling." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[17] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In ECCV, 2016.

[18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In CVPR, 2016.

[19] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6479-6488. 2018.

[20] Wang, Huaijun, et al. "Wearable sensor-based human activity recognition using hybrid deep learning techniques." *Security and Communication Networks* 2020 (2020).

[21] Rublee, Ethan, et al. "ORB: An efficient alternative to SIFT or SURF." 2011 International conference on computer vision. Ieee, 2011.

[22] Zhu Daixian, "SIFT algorithm analysis and optimization," *2010 International Conference on Image Analysis and Signal Processing*, 2010, pp. 415-419, doi: 10.1109/IASP.2010.5476084.

[23] Bay H., Tuytelaars T., Van Gool L. (2006) SURF: Speeded Up Robust Features. In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg.

[24] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[25] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[26] M. Soliman, M. Kamal, M. Nashed, Y. Mostafa, B. Chawky, D. Khattab, " Violence Recognition from Videos using Deep Learning Techniques", Proc. 9th International Conference on Intelligent Computing and Information Systems (ICICIS'19), Cairo, pp. 79-84, 2019.

[27] Mostafa Mohamed Moaaz; Ensaf Hussein Mohamed. "Violence Detection In Surveillance Videos Using Deep Learning". FCI-H Informatics Bulletin, 2, 2, 2020, 1-6. doi: 10.21608/fcihib.2020.42233.1003.

[28] Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using oriented violent flows. Image and vision computing, 48, 37-41.

[29] Bilinski, P., & Bremond, F. (2016, August). 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.

[30] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105.