# Integrating Pretrained VGG19 and Bi-LSTM for Violence Detection in Video

Pablo Negre[1](✉), Ricardo S. Alonso[2,3], Javier Prieto[1], and Paulo Novais[4]

[1] BISITE Research Group, Universidad de Salamanca, Salamanca 37007, Spain
{pablo.negre,corchado,javierp}@usal.es

[2] AIR Institute, Av. Santiago Madrigal, 39, Salamanca 37003, Spain
ralonso@air-institute.com

[3] UNIR (International University of La Rioja), Av. de la Paz, 137, Logrono, Spain
ricardoserafin.alonso@unir.net

[4] Departamento de Informática, Universidade do Minho, Braga, Portugal
pjon@di.uminho.pt

**Abstract.** Physical aggression presents a pervasive challenge worldwide, disrupting various aspects of individuals' lives and societal functioning. This phenomenon stems from difficulties in emotional regulation, interpersonal conflicts, and socio-economic factors. Women and minors are particularly vulnerable demographics, facing high rates of violence in both intimate and public settings. Despite concerted efforts, violence detection remains a crucial issue with implications for public safety and social well-being. In this study, the challenge of violence detection in videos is addressed utilizing a combination of pre-trained VGG19 and Bi-LSTM layers. While promising results have been demonstrated in previous research utilizing VGG16, the potential effectiveness of VGG19 in this context has not been thoroughly investigated, although a larger number of convolutional layers should mean a better understanding of the scene. Moreover, the use of Bi-LSTM layers has been shown to be superior to the use of LSTM layers by up to 3%. A broad range of hyperparameter combinations is explored to optimize model performance. Positive results are obtained through experiments, with accuracies of 97%, 90%, and 73% achieved on the Hockey Fights dataset, Violent Flow Dataset, and RWF-200, respectively. Although the model does not surpass state-of-the-art approaches utilizing VGG16, it exhibits promise when compared to other proposals within the field. Overall, our study contributes to advancing violence detection methodologies, emphasizing the importance of leveraging deep learning techniques for improving public safety and social well-being.

**Keywords:** Violence detection · Video Surveillance · Pretrained VGG19 · Bi-LSTM · Fine-tuning

R. Chinthaginjala et al. (Eds.): DCAI 2024, LNNS 1259, pp. 139–149, 2025.
https://doi.org/10.1007/978-3-031-82073-1_14

# 1 Introduction

Acts of physical aggression pose a significant challenge in our society, exerting a global influence. This phenomenon disrupts various aspects of life, affecting not only the immediate targets and their psychological well-being [15], but also reverberating through families, communities, and the overall functioning of nations (including mobility, tourism, and commerce) [13]. The origins of aggressive conduct stem from difficulties in emotional regulation, interpersonal conflicts, and the socio-economic and demographic characteristics of communities [7].
Among the most affected demographics are women and minors. Concerning women, as per a report by the World Health Organization (WHO), roughly one in three women globally has encountered either physical or sexual violence from an intimate partner or non-partner during their lifetimes [19]. Research conducted by ActionAid revealed that 79% of Indian women, 86% of Thai women, 89% of Brazilian women, and 75% of women in London face harassment or violence in public spaces [25]. Regarding children, findings published by [10] in the American Academy of Pediatrics indicate that in 2015, at least half of children in Asia, Africa, and North America experienced violence, with over 1 billion children globally, aged 2–17, being subjected to such treatment.

In summary, physical aggression presents a pervasive challenge worldwide, impacting nearly every aspect of individuals' lives and entire societies. Protecting the physical well-being of people globally should be a fundamental priority and a universal right.

# 2 Violence Detection State of the Art

In this Section it will present the most widely used datasets of violence videos in the state of the art, as well as the different types of violence detection algorithms that have been used in the state of the art.

Various approaches to addressing violent behaviors within societies have been explored. Beginning with more indirect and enduring strategies, numerous studies endeavor to comprehend the circumstances or contexts that precipitate acts of violence, with the aim of rectifying them [15]. Conversely, other investigations, exerting a more immediate impact, scrutinize the correlation between crime rates and the urban settings in which they manifest [24]. However, the final line of defense for victims of violence lies in the real-time detection of such occurrences to promptly alert authorities for conflict resolution, as well as the capture of video or imagery that can subsequently aid in identifying the individuals involved [2,11,23].

Violence detection algorithms in videos typically follow the same basic steps [18]; these steps are shown in Fig. 1. First, the input video is processed to extract key frames, which are representative images of different moments in the video. Then, preprocessing is done on the frames before introducing the data to the algorithm. Next, the violence detection algorithm employs feature extraction techniques to capture relevant information from the frames and a training process on violent and non-violent videos. The algorithm classifies each frame as

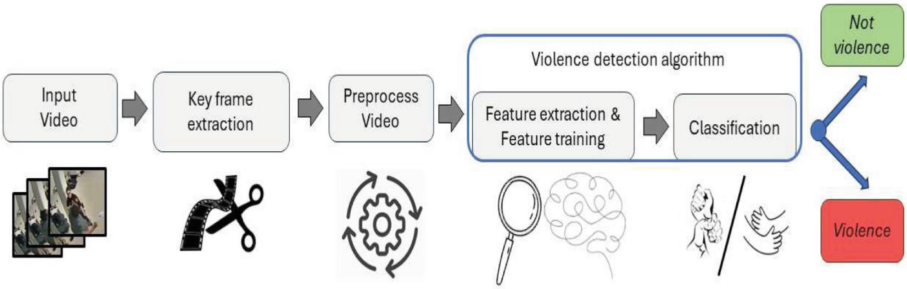

**Fig. 1.** Basic violence detection steps

"violent" or "non-violent" based on the extracted features and the trained classification model, thereby providing violence detection in the video. Considering that identifying violence constitutes an atypical task and that instances of physical aggression can vary significantly (in terms of participant count, type of aggression, severity level, etc.), ensuring an ample volume of data for algorithm training is paramount. Within the realm of contemporary research, numerous datasets have surfaced in recent years, albeit a select few are consistently utilized to enable comparison with findings from prior studies. Some of the most frequently employed datasets include [18]: *Hockey Fights* [4], which comprises footage from the National Hockey League (NHL); *Action Movies* [4], housing sequences from action-oriented films; *Violent Flow* [9], alternatively known as *Violent Crowd*, encompassing YouTube videos depicting authentic crowded scenarios with accompanying audio; and *Real World Fight-2000 (RWF-2000)* [5], comprising altercations captured by surveillance cameras in real-world.

A wide variety of algorithms is employed for violence detection, traditionally categorized into: *Conventional Methods*, which rely on manual feature extraction and traditional Machine Learning algorithms; and *Deep Learning Methods*, which leverage Deep Learning techniques, [16,23]. However, such categorization proves overly broad; hence, we present recent endeavors in video violence detection using artificial intelligence, classified by the algorithms' nature.

- **CNN**: approaches based on CNNs for violence detection extract spatial features from video frames and utilize them for classification decisions. The CNN adjusts its weights during training to extract features. Several studies demonstrate improved accuracy by employing pre-trained CNNs on extensive image datasets and utilizing fine-tuning techniques, leveraging the pre-trained network weights [2,11].
- **LSTM**: methods based on LSTMs extract temporal features of violence, capturing temporal patterns. LSTMs modify their weights during training to extract features. The combination of CNNs and LSTMs stands as the most prevalent approach in recent literature [18].

- **Manual Feature**: algorithms wherein feature extraction or training, or both, are conducted on mathematical principles unrelated to machine learning or deep learning [12].
- **Skeleton-based (Deep Learning or Manual)**: these algorithms aim to identify body positions in videos and infer violence based on these positions, employing mathematical or deep learning techniques [20].
- **Transformer**: this category employs Transformer-based architectures for violence detection in video content [3].
- **Audio-based (Deep Learning or Manual)**: encompassing studies utilizing video audio for violence detection, although not exclusively dependent on audio as the algorithm's input, emphasizing its significance within these works [14].

The combination of CNN and LSTM for violence detection is the most widely used architecture in the recent state of the art [18]; where some of the works have shown outstanding results [1,16].

## 3 Objectives and Model Arquitecture

This section presents the objectives and architecture of the model proposed for this work.

One of the most widely used pre-trained CNNs is VGG-16, obtaining excellent results in violence detection in combination with LSTM and Bi-LSTM layers [1,16]. The use of VGG-19 is lower, and the results have not been as good [17], although the use of a greater number of convolutional layers should provide a greater capacity to extract complex patterns. It is also the case that violence detection studies do not use the same datasets, so comparing their results is costly [1]. Finally, in the works that combine the use of CNN and LSTM, it is not highlighted that a certain number of neurons or layers of LSTM or Fully Connected Layers leads to better results in the prediction of violence [8]. Therefore, the aim of this work is the use of pre-trained VGG19 combined with Bi-LSTM layers; the model will be trained and tested with 3 of the most used datasets in the state of the art, and the optimal number of neurons for the Bi-LSTM and Fully Connected Layers will be searched for in a wide range of hyperparameters.

The architecture of the model proposed for this work is shown in Fig. 2 First of all, the frames are pre-processed, since VGG-19 expects to receive an array of shape (224,224,3), which indicates 224 pixels wide and high for each RGB colour channel. VGG-19 consists in going through a series of convolutional blocks, which are made up of consecutive convolutional layers. Between each convolutional block there are interspersed Max Pool Layers. Once the image has passed through all the convolutional blocks, it is passed to three densely connected layers that act as a classifier; in total 16 convolutional layers, 5 pooling layers and 3 fully connected layers. VGG19 pre-trained on the ImageNet dataset is used, taking advantage of its weights for spatial feature extraction. After feature extraction by VGG19 pre-trained features that are obtained from the output of its last convolutional layer, they are pooled and passed through a 2D Global Average Pooling layer. They are then fed into two Bi-LSTM layers that extract spatial features and finally into three Fully connected layers that act as a classifier between whether the video is violent or not.

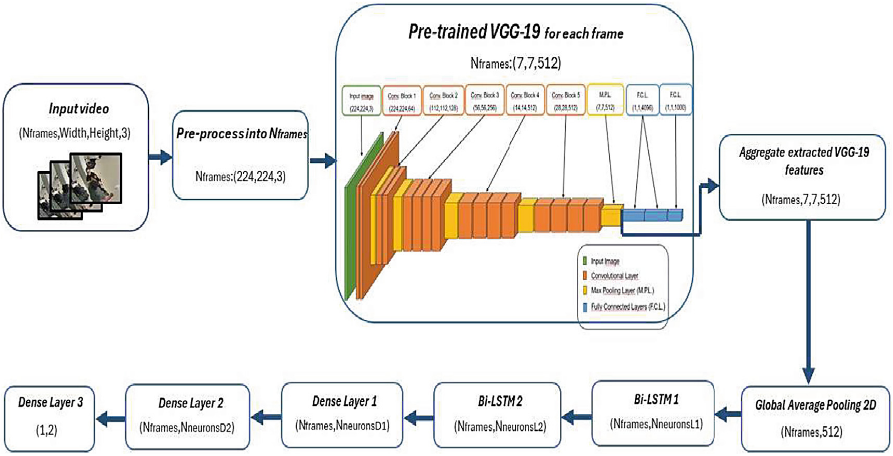

**Fig. 2.** Proposed Pretrained VGG19 and Bi-LSTM violence detection model

## 4 Experiments and Results

The results obtained by training and testing the violence detection model shown in Fig. 2 are presented. The model is trained and tested with three of the most widely used physical aggression datasets in the literature, already exposed in Sect. 1: *Hockey Fights* [4] with 1000 videos and 50 median frame number, *Violent Flow* with 247 videos and 107 median frame number [9] and *Real World Fight-2000 (RWF-2000)* [5] with 2000 videos and 150 median frame number. Only Violent Flow has videos with different lengths and since our model requires a fixed length of the analysed video, we take the median as a fixed value of the videos, repeating the last frame in case the video in question has less frames and trimming from the end of the video in case the original video has more than 107 frames.

### 4.1 Violence Detection Model Training

Since VGG-19 is already pre-trained on ImageNet, all that remains is to train the two Bi-LSTM layers and three fully connected layers. These layers are trained by introducing the extracted spatial features, after grouping them for each video in the form $(Nframes, 7, 7, 512)$ and after applying the Global Average Pool.

During the training, multiple hyperparameter values will be tried. The learning rate tried are: $10^{-2}$, $10^{-3}$, $10^{-4}$, Bi-LSTM and fully connected layers neuron number tried values are: 64, 128, 256, 512 and 1024. It is permitted during training to experiment with varying the number of neurons in the first and second layers of Bi-LSTM and fully connected layers, in order to assess how altering the number of neurons affects the results.

The three datasets will undergo training for a total of 150 epochs. The optimizer utilized is *Adam*, and the loss function is *Binary Cross Entropy*. Due to the extensive search range, the number of potential combinations is significantly high. Keras Tuner is employed to handle this task. A total of 200 combinations are evaluated, representing approximately 10% of the total 1875 possible combinations. Given that our aim during training is to maximize the Validation Accuracy metric, the top three models identified by Keras Tuner in its hyperparameter optimization process are preserved. This is crucial because among models achieving similar Validation Accuracy, the one selected may not necessarily generalize best to unseen test data.

The results of the three best models obtained for the three selected datasets are shown in Table 1, where the Validation Accuracy value obtained and the values of Learning Rate and number of neurons for the Bi-LSTM and Fully Connected Layers are shown for each model.

**Table 1.** Training results of the best three models with Bi-LSTM layers hyperparameters combinations for the three selected datasets

| Dataset | Model | Validation Accuracy | Learning Rate | Bi-LSTM 1 | Bi-LSTM 2 | F.C.L 1 | F.C.L 2 |
|---|---|---|---|---|---|---|---|
| Hockey Fights | 1 | 0.97 | 0.01 | 64 | 64 | 64 | 512 |
| | **2** | 0.97 | 0.01 | 512 | 64 | 1024 | 512 |
| | **3** | 0.97 | 0.01 | 64 | 256 | 64 | 256 |
| **RWF-2000** | **1** | 0.87 | 0.01 | 512 | 64 | 256 | 64 |
| | **2** | 0.87 | 0.001 | 512 | 256 | 64 | 256 |
| | **3** | 0.87 | 0.001 | 128 | 128 | 64 | 128 |
| **Violent Flow** | **1** | 0.96 | 0.001 | 1024 | 512 | 64 | 256 |
| | **2** | 0.96 | 0.01 | 256 | 256 | 256 | 128 |
| | **3** | 0.96 | 0.01 | 128 | 512 | 128 | 256 |

The three models stored for each dataset obtain the same validation accuracy during the training process. In the case of the Hockey Fights and Violent Flow dataset, the results are excellent, obtaining 96% validation accuracy, unlike the RWF-2000 dataset, which obtains 87% validation accuracy. It is understood that this is due to the greater complexity and variety contained in the RWF-2000 dataset videos, as well as greater complexity in the identification of the scene due to poorer lighting conditions, distant violent scenes, etc.

### 4.2 Violence Detection Model Testing

This Section presents the results obtained during the testing of the top 3 models for the architecture that, based on the spatial features extracted by the pre-trained VGG-19 network, utilizes Bi-LSTM layers for spatial feature extraction and densely connected layers for classification between violence and non-violence. Since Accuracy is the most commonly used metric in the state of the art [18], it will be the basis for choosing the best model among the three best stored by Keras Tuner for each of the three selected datasets.

Table 2 contains the test results of the architecture employing Bi-LSTM layers. In case there are two models with equal testing accuracy, the model with the highest selection by Keras Tuner is chosen. For the Hockey Fights dataset, the best model is the second one, with a testing accuracy of 96%. In the case of the RWF-2000 dataset, the best model is the second one, with a testing accuracy of 72%. Lastly, for the Violent Flow Dataset, the best model is the first model, with a testing accuracy of 86%.

**Table 2.** Testing results of the best three models for each of the three selected datasets

| Dataset | Model | Test Accuracy |
|---|---|---|
| Hockey Fights | 1 | 0.93 |
| | 2 | 0.95 |
| | 3 | 0.97 |
| RWF-2000 | 1 | 0.73 |
| | 2 | 0.73 |
| | 3 | 0.68 |
| Violent Flow | 1 | 0.90 |
| | 2 | 0.82 |
| | 3 | 0.88 |

### 4.3 Comparison with State-of-the-Art

This section presents a comparison of results with other state-of-the-art work on video violence detection. In the table, recent state-of-the-art papers that have used a combination of VGG-16 and VGG-19 in conjunction with some form of RNN are highlighted in blue in the *Cite* column. In orange are two articles, one of them using a CNN in combination with an Bi-LSTM and another article using only a pre-entry CNN (without combining it with an RNN). It can be seen how articles using VGG-16 pre-workout get better results compared to our model. Regarding the study using pre-trained VGG-19 together with more CNN and Bi-LSTM layers, our model obtains better results in the Hockey Fights Dataset, but not in the other two datasets used. With respect to the other state-of-the-art studies, our algorithm obtains better results in the Hockey Fights Dataset and similar results in the RWF-2000 dataset (Table 3).

**Table 3.** Comparison of results with other state of the art articles on violence detection

| **Cite** | **CNN** | **LSTM** | **Hockey Fight Acc.** | **Violent Flow Acc.** | **RWF-2000 Acc.** |
|---|---|---|---|---|---|
| [1] | PT VGG-16 (ImageNet) | LSTM | 99.1 | X | X |
| [16] | PT VGG-16 (ImageNet) | Bi-Conv-LSTM | 99.1 | 98.4 | 92.4 |
| [8] | PT VGG-16 (ImageNet) | LSTM/<br>Bi-LSTM | 97.6/<br>98.8 | 92.2/<br>95.1 | X |
| [22] | PT VGG16 (INRA) | Bi-GRU | 98 | 95.5 | X |
| [17] | PT VGG-19 (ImageNet) + extra CNN layers | Bi-LSTM | 91.29 | 90.47 | 90.47 |
| [21] | CNN | LSTM | 94.9 | 92.2 | 77.31 |
| [6] | CNN PT | N | 94.1 | X | 72 |
| **Proposed model** | PT VGG-19 (ImageNet) | Bi-LSTM | 97 | 90 | 73 |

All in all, the use of pre-trained VGG-19 with Bi-LSTM layers, testing a wide range of hyperparameters of learning rate and number of neurons of the Bi-LSTM and Fully connected layers has given good results, especially in the Hockey Fights Dataset, although not better than the results obtained by other works using pre-trained VGG-16. Still, it is competitive with other state-of-the-art papers using other architectures.

## 5 Conclusions and Future Work

This work addresses the challenge of violence detection in videos, a crucial problem with significant implications for public safety and social well-being [13,15]. In the state of the art, there are multiple works that contribute different kinds of algorithms for violence detection in videos using artificial intelligence [12,20], of which the most commonly used is the combination of Convolutional Neural Networks along with Long Short-Term Memory [1,16].

The aim of this work has been the development of a violence detection algorithm using pre-trained VGG19 in combination with Bi-LSTM layers, since the use of VGG-16 has yielded excellent results in violence detection, although VGG-19 has been considerably less utilized. Additionally, in theory, a higher number of convolutional layers could imply a better understanding of the scene, but a clear improvement in the use of certain hyperparameter values has not been found in the literature. Therefore, in this work, a wide range of combinations for the hyperparameters of learning rate and number of neurons in the Bi-LSTM and Fully Connected Layers has been used.

The results obtained have been positive, with accuracies of 96%, 86%, and 72% in the Hockey Fights dataset, Violent Flow Dataset, and RWF-2000, respectively. Although these are positive results, especially in the use of the Hockey Fights Dataset, the model did not achieve better results than other state-of-the-art works using pre-trained VGG-16 [1,8,16,22], even after training our model over a wide range of learning rate values and Bi-LSTM and Fully connected layers neuron numbers. Nonetheless, compared to other proposals from the state of the art, the results are positive [6,21].

As future work, the study of how the training of some layers of pre-trained VGG16 CNN combined with Bi-LSTM layers may improve the results of the state of the art works. Additionally, the application of XAI techniques for understanding why the algorithm is incorrectly predicting certain scenes and whether the mispredicted scenes follow any pattern is proposed.

**Acknowledgements.** This research has been supported by the project "Smart, Sustainable and coheSive Digitalization conceived as a Digital Innovation Hub (DIGIS3)", Grant Agreement 101083898. Co-financed by the European Union (EU) and the Ministry of Industry, Trade and Tourism and the EOI Foundation of the Government of Spain, in the framework of the Recovery and Resilience Mechanism financed by the Next Generation funds of the EU. The views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the EU, Ministry of Industry, Trade and Tourism or the EOI Foundation. Neither the EU nor the granting authority can be held responsible for them.

## References

1. Aarthy, K., Nithya, A.A.: Crowd violence detection in videos using deep learning architecture. In: 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), pp. 1–6. IEEE (2022)
2. Ahmed, M., et al.: Real-time violent action recognition using key frames extraction and deep learning (2021)
3. Aktı, Ş., Ofli, F., Imran, M., Ekenel, H.K.: Fight detection from still images in the wild. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 550–559 (2022)
4. 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). https://doi.org/10.1007/978-3-642-23678-5_39

5. Cheng, M., Cai, K., Li, M.: RWF-2000: an open large scale video database for violence detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4183–4190 (2021)
6. Cheng, S.-T., Hsu, C.-W., Horng, G.-J., Jiang, C.-R.: Video reasoning for conflict events through feature extraction. J. Supercomput. **77**(6), 6435–6455 (2021). https://doi.org/10.1007/s11227-020-03514-5
7. Enaifoghe, A., Dlelana, M., Durokifa, A.A., Dlamini, N.P.: The prevalence of gender-based violence against women in South Africa: a call for action. Afr. J. Gend. Soc. Develop. **10**(1), 117 (2021)
8. Gupta, H., Ali, S.T.: Violence detection using deep learning techniques. In: 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), pp. 121–124. IEEE (2022)
9. Hassner, T., Itcher, Y., Kliper-Gross, O.: 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. IEEE (2012)
10. Hillis, S., Mercy, J., Amobi, A., Kress, H.: Global prevalence of past-year violence against children: a systematic review and minimum estimates. Pediatrics **137**(3) (2016)
11. Jayasimhan, A., Pabitha, P.: A hybrid model using 2D and 3D convolutional neural networks for violence detection in a video dataset. In: 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), pp. 1–5. IEEE (2022)
12. Lohithashva, B.H., Aradhya, V.N.M.: Violent video event detection: a local optimal oriented pattern based approach. In: Mahmud, M., Kaiser, M.S., Kasabov, N., Iftekharuddin, K., Zhong, N. (eds.) AII 2021. CCIS, vol. 1435, pp. 268–280. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82269-9_21
13. Long, D., Liu, L., Xu, M., Feng, J., Chen, J., He, L.: Ambient population and surveillance cameras: the guardianship role in street robbers' crime location choice. Cities **115**, 103223 (2021)
14. Mahalle, M.D., Rojatkar, D.V.: Audio based violent scene detection using extreme learning machine algorithm. In: 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1–8. IEEE (2021)
15. Muarifah, A., Mashar, R., Hashim, I.H.M., Rofiah, N.H., Oktaviani, F.: Aggression in adolescents: the role of mother-child attachment and self-esteem. Behav. Sci. **12**(5) (2022)
16. Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A.H., Nawaz, H.S.: A frame-based feature model for violence detection from surveillance cameras using convLSTM network. In: 2021 6th International Conference on Image, Vision and Computing (ICIVC), pp. 55–60. IEEE (2021)
17. Mumtaz, N., Ejaz, N., Aladhadh, S., Habib, S., Lee, M.Y.: Deep multi-scale features fusion for effective violence detection and control charts visualization. Sensors **22**(23), 9383 (2022)
18. Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M.: State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. PeerJ Comput. Sci. **8**, e920 (2022)
19. de la Salud, O.M.: Violence against women (2021)
20. Su, Y., Lin, G., Zhu, J., Wu, Q.: Human interaction learning on 3D skeleton point clouds for video violence recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 74–90. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_5

21. Talha, K.R., Bandapadya, K., Khan, M.M.: Violence detection using computer vision approaches. In: 2022 IEEE World AI IoT Congress (AIIoT), pp. 544–550. IEEE (2022)
22. Traoré, A., Akhloufi, M.A.: 2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos. In: International Conference on Image Analysis and Recognition, pp. 152–160. Springer (2020)
23. Ullah, F.U.M., et al.: AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. IEEE Trans. Ind. Inf. **18**(8), 5359–5370 (2021)
24. Vomfell, L., Härdle, W.K., Lessmann, S.: Improving crime count forecasts using twitter and taxi data. Decis. Support Syst. **113**, 73–85 (2018)
25. Wilkinson, S.: Meet the heroic campaigners making cities safe for women. In: Action Aid, Global Street Harassment-Making Street Safer (2016)