INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 04-08-2023

Accepted: 30-08-2023

Published: 30-09-2023

Citation: Jaiswal SG, Mohod SW, Sharma D, Hinge A (2023) Violent Video Classification with Transfer Learning Approach Using Inception-V3 and Support Vector Machine. Indian Journal of Science and Technology 16(37): 3018-3026. https://doi.org/10.17485/IJST/v16i37.1972

* **Corresponding author**.

mr.snehil.jaiswal@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Violent Video Classification with Transfer Learning Approach Using Inception-V3 and Support Vector Machine

Snehil G Jaiswal[1]*, Sharad W Mohod[2], Dinesh Sharma[3], Ankita Hinge[4]

1 Research Scholar, SantGadge Baba Amravati University Amravati and Registrar, G H Raisoni University, Amravati, Maharashtra, India
2 Professor and Head of Department, Department of Electronics and Telecommunication Engineering, Prof. Ram Meghe Institute of Technology & Research, Amravati, Maharashtra, India
3 Assistant Professor, Department of Electronics and Telecommunication Engineering, Chandigarh College of Engineering and Technology, (Government Institute Under Chandigarh UT Administration), Chandigarh, India
4 M.Tech Student, Department of Electronics and Telecommunication Engineering,G H Raisoni University, Amravati, Maharashtra, India

Abstract

Objectives: Research in surveillance systems is growing, with cameras in public places capturing actions for the live surveillance, goal-driven investigation, event forecasting, and intrusion detection. Violent video classification system plays a critical role in development of violence detection system for public security and safety. Such system is useful in identification of violent behaviors, such as fighting or assault. **Methods:** The Inception-V3 architecture using Convolutional neural networks extracts the informative features from the input video frames. Support Vector Machine is used to select features for classification once the remaining layers of the Inception-V3 model have been frozen. **Findings:** The datasets used in many contemporary and current innovative techniques, including the Hockey battle dataset and the Movies dataset, are used to train and assess the proposed hybrid model. The experiment findings show that the suggested violence detection algorithm performs well in terms of average metrics, with accuracy, precision, recall, and F-Score being 96 $\pm$ 2%, 98 $\pm$ 2%, 96 $\pm$ 1%, 0.95 respectively. **Novelty:** Transfer Learning approach is applied which involves lightly retraining pre-trained models on different datasets, resulting in improved performance in terms of computational resources and accuracy.

Keywords: Deep Learning; Convolutional Neural Network; Action Recognition; Violence Detection; Action Recognition; InceptionV3; Support Vector Machine (SVM)

1 Introduction

Violence detection is an action recognition area that analyses video footage in public environments to identify aberrant human activities. Researchers are increasingly interested in violence detection and prevention due to its accessibility and accessibility. Violence detection and classification are critical in spotting anomalies in video situations, particularly ones with minimal or no violence[1]. Public protests and riots can escalate into violent incidents, necessitating the establishment of intelligent systems that are efficient and precise. These systems analyze crowded areas to detect and prevent suspicious events, enhancing security measures. Deep learning models have proven to be effective in the development of surveillance systems for law enforcement agencies[2,3].

Detecting and categorizing violence in videos is challenging due to various factors such as source and nature of videos. In movies, techniques like extreme camera shakes and agitated background music helps to identify cinematic tropes and thus becomes a well-defined problem statement for video action classification[4]. However, the problem become challenging while analyzing surveillance films.

Identifying and classifying violence in streaming services is crucial for ensuring rules and regulations, removing inappropriate content, and recommending appropriate content for children[5]. Due to issues in crowded environments and massive datasets, human-controlled monitoring systems for theft, aggressiveness, and strange behaviors are out of date. Issues include tracking difficulties, non-stationary backgrounds, blur motion, and occlusion[6]. Advancements in artificial intelligence have led to strategies using deep neural networks and machine learning techniques. Trained convolutional neural networks can classify videos based on extracted spatial and temporal features.

Automatic violence detection in videos is primarily based on action recognition techniques. There are two classes: local features, which use Points of Interest (POIs) to represent actions between frames, and global features, which evaluate characteristics from multiple frames. These techniques use spatial, motion, and temporal information[7]. Deep learning-based approaches are emerging, and related works are divided into three groups: techniques based on deep learning, global feature analysis, and local feature analysis.

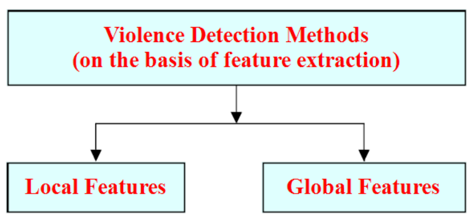

Fig 1. Types of Violence Detection Methods

1.1 Detecting Violence Using Local Features

Researchers in[8] used acceleration to detect potential aggression in speed variations, but its accuracy is limited due to partial information from adjacent frames, limiting the representation of all spatio-temporal information in videos. The approach proposed in[8] is extracting LHOG and LHOF variations of HOG and HOF, using motion regions instead of Points of Interest. The performance of this approach is quite good on benchmark datasets. For local spatio-temporal descriptors to be acceptable for classifiers like SVM, they must have encoding features for a more significant representation and a fixed dimension descriptor.

The method based on extraction of visual hand-crafted features, such as angles, velocity and contact between two human subjects and creating a feature vector with encoded temporal information was proposed in[9]. Further a binary classification SVM model is utilized to predict violent behavior. Recent deep learning-based methods have a fixed output dimension and traditional local feature-based method may produce useless points of interest in crowded situations where there are many moving subjects. This makes traditional local feature-based methods less accurate. However, deep learning-based approaches show better generalization capacity in violence detection, outperforming the most recent techniques.

1.2 Global Features based violence detection

Global features-based approaches often rely on acoustic or visual features to classify violence scenes. While traditional visual strategies define violent scenes with specific visual characteristics like blood, fire, blasts and weapons, audio-based methods define violence as occurrences like shots, explosions, battles, and screaming. However, such methods may give false positives and may generalize violence in videos, making audio-based methods crucial for accurate classification.

The authors in [10] introduced the Violent Flow descriptor (VIF) for identifying violence in crowded places on the basis of rate of change of optical flow magnitude. VIF has a faster response time and is designed for real-time detection. The extended version was proposed as Oriented VIF, which was based on the theory that VIF may lose important information due to consideration of rate of change of magnitude factor of optical flow alone [11]. The method proved the importance of addition of orientation of optical flow. OVIF works more effectively in detecting interpersonal violence but is less effective in crowded environments.

1.3 Deep Learning Techniques for Violence Detection

The approaches for detecting violence in the recommended architectures that use deep learning algorithms are outlined in detail here. In video violence detection, CNN is frequently utilized, with deep learning algorithms being the foundation [12]. Neural networks incorporated with additional convolutional layers lays a foundation to categorize violent recognition based on data and extracted features. Some of the methods applying deep learning algorithms to identify violence are described here. Deep learning-based methods for detecting violence in movies are outlined in Table 1, along with object identification, feature extraction, classification, application, and evaluation criteria.

A real-time violence surveillance system [13] uses a simulation of human intelligence to analyze streaming data and identify hostility. For the purpose of detecting violent scenes, the system divides frames from a large dataset of real-time video from diverse sources and bidirectional LSTM. The network trains on a 2314-movie dataset named as violent interaction dataset (VID) with 1077 fights and 1237 no-fights, including neutral and violence scenes. The system's effectiveness and strength are supported by its accuracy of 94.5% in identifying violent behavior.

An ensemble model using Mask RCNN and LSTM was proposed to identify violent behaviors in individuals [14]. Experiments on Weizmann, KTH, and own datasets showed the model outperformed individual models, achieving a 93.4% accuracy rate by extracting temporal features using key points-mask integration. This strategy is appropriate for industry and benefits society in terms of security.

A two-stream CNN architecture and SVM classifier are proposed for detecting violence with minimum processing time. Feature extraction, training, and label synthesis are steps in the technique, which make use of Image net VGG-f architectures that have already been trained [15]. While the second stream pulls motion data, the initial stream collects visual data from frame variations. However, this method has difficulty detecting close-range violent behaviors, making it difficult to detect violence in big crowds.

Table 1. Analysis of Methods using Deep Learning Reported in Literature

Method	Feature extraction, object detection and classification method	Types of scenes	Accuracy Obtained
Big data, long short-term memory networks, deep learning framework for football stadium. [13]	Bi-LSTM, Histogram of Oriented Gradients (HOG) image descriptor and a	Surge Environment	93 to 94 %
Deep violence detection framework using handcrafted features. [16]	An innovative differential motion energy image as an exclusive feature ConvNet	both densely populated and uncrowded	Around 97%
Technique for detecting violence that utilizes ConvNet and Support Vector Machine (SVM)in bi-channels. [17]	Bi-channels ConvNet and Linear Support Vector Machine (SVM)	Both densely populated and uncrowded	95 to 98%: Hockey fight 93 to 95%: Violence crowd
Detecting violent scenes using ConvNet and deep acoustic characteristics [18]	Multiple-Feature-Branch Convolutional Neural Network	Surge Environment	Around 90%

[19] demonstrate a method for identifying violent scenes from auditory information present in the video.They employ a deep acoustic feature extraction algorithm and CNN as a classifier, splitting MFB features into three maps. The violent scene detection process is applied to each frame, and detection is generated by pooling at the segment level. SVM classifiers for violent scene identification employ CNN-based features and apply maximum or minimum segment pooling. Studies using the MediaEval dataset reveal that this strategy beats basic methods in terms of average accuracy. The suggested approach outperforms basic strategies in experiments utilizing the MediaEval dataset in terms of average accuracy.

This study presents a method for detecting violence using the Inception-V3 network modules and the C3D network architecture. The integration of features extracted using most of the layers of Inception-V3 network and other classifier results into transfer learning approach. The research uses deep learning architecture and CNN models for feature extraction, driven by Support vector machine as classifier.

2 Methodology

The proposed violent video detection and classification system's layout is shown in Figure 2, which consists of a linear SVM classifier and an Inception-V3 model that has been customized. The step-by-step operations executed in the proposed algorithm are discussed in subsequent subsections.

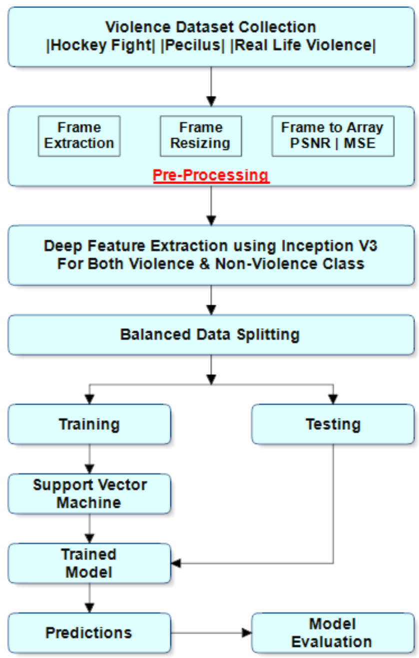

Fig 2. Proposed algorithm for violent video classification using Inception-V3and SVM

2.1 Violence Dataset Collection

Violence detection and classification datasets come from various sources, including YouTube, real-time CCTV footage, movies, and mobile phone recordings. Challenges include small data amounts, video quality, and size, impacting detection in both surveillance and non-surveillance domains. Hockey Fight and Movie Dataset are the two distinguished datasets collected and utilized for testing the proposed system.

Researchers in[16] developed a new video dataset for evaluating violence detection systems in dynamic settings. They collected 1000 National Hockey League clips, labeling them as "fight" or "non-fight," to measure the performance of various recognition approaches. This dataset named as Hockey Fight Dataset is widely used in various methods proposed earlier for violence detection task. In addition to this Movie dataset[16] consisting of 200 videos, 100 violence and 100 non-violence, with diverse backgrounds, contrasting hockey dataset is also considered for present study.

2.2 Preprocessing

Various necessary operations such as video to frame conversion, frame resizing etc are required to be done prior to feature extraction. Preprocessing is conversion from raw data to a suitable data for efficient feature extraction[17]. In a preliminary operation, all the frames of the input video are extracted and read. Each individual frame can now be treated as image. The input image size to the first layer of Inception-V3 network is of $299\times299\times3$. There is a chance that the database videos may be of different size and specialization and thus the frame resizing operation is required[18].

In a video clip, violent action may occur in specific parts, with motions being prominent in some frames. Human actions are not without motions. To detect frames involving major violent events, Mean Square Error (MSE) and PSNR estimation is used for detecting frames using algorithm in [20]. This frame selection approach reduces feature extraction frame count, reducing computational cost. Based on the selected frames, an array containing information major event frames is built for further processing.

2.3 Deep Feature Extraction

Deep feature extraction using Inception-V3 pre-trained deep learning model efficiently utilizes representational power, reducing computational power by modifying previous architectures [21]. Inception Networks (GoogLeNet/Inception v1) are more computationally efficient than VGGNet, reducing parameters and resource costs. However, adaptation to different use cases is challenging due to uncertainty in efficiency. For simpler adaptation, Inception-V3 models recommend optimizing methods including factorized convolutions, regularization, reduction of dimensions, and distributed calculations.

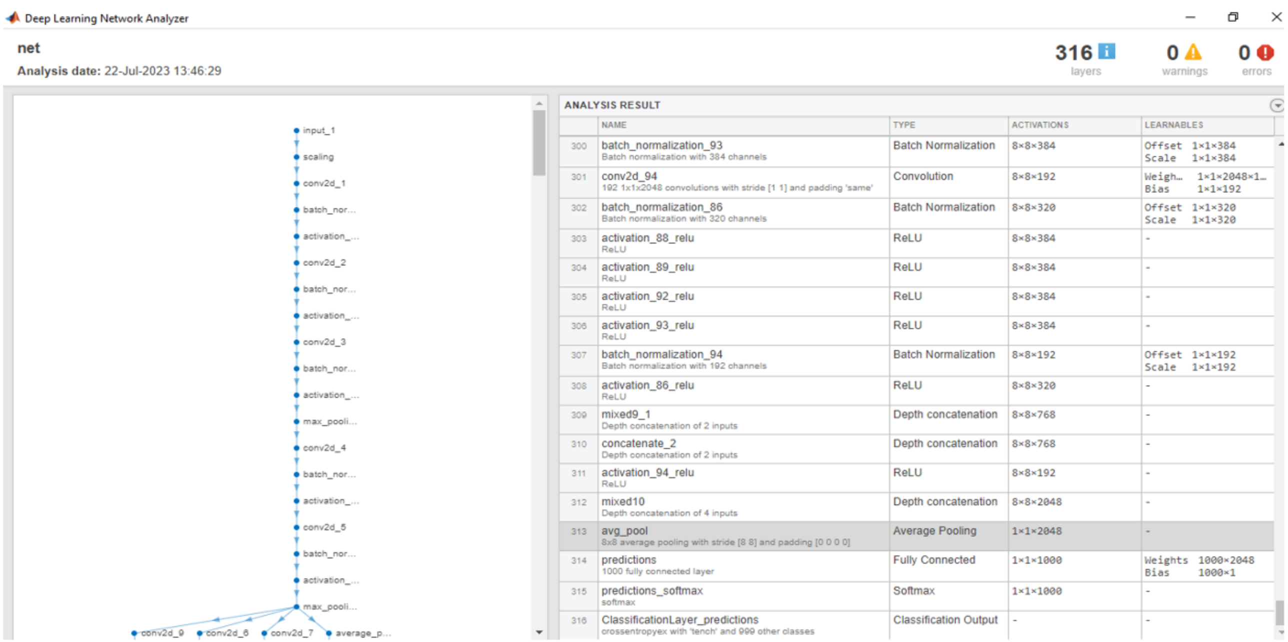

Fig 3. Layers of Inception-V3 model visualized Network Analyzer of Matlab

The pre-trained Inception-V3 model utilized in present research consists of total 316 different layers which includes Input layer, Scaling layer, Convolution layer, AvgPoollayer, MaxPool layer, Concat layer, Dropout layer, SoftMax layer, etc. The pictorial visualization using Deep learning network analyzer of Matlab is shown in Figure 3.

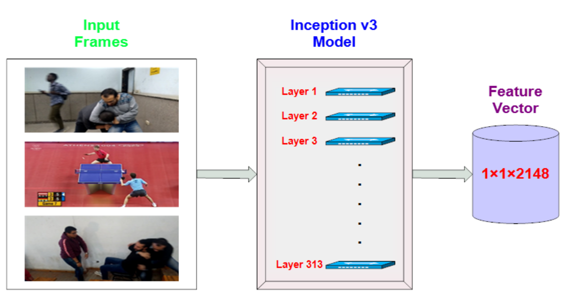

Fig 4. Inception-V3model setup for feature extraction

Transfer learning in Deep Learning utilizes existing models to train neural networks with limited data, making it crucial in data science scenarios requiring more labeled data. Transfer learning transfers trained machine learning model knowledge to related problems, enabling classifiers to predict 'A' and identify 'B' using their training knowledge if trained for any one of the classes. Transfer learning uses learned concepts in one task to understand others, utilizing weights automatically making it suitable for problems in computer vision area. Transfer learning offers shortened training period, better neural network efficiency, and minimal data, enabling the generation of good machine learning models with pre-trained models. Due to such advantages, transfer learning and representation learning is utilized in proposed work to extract the deep features from violence detection dataset. Deep learning aids in identifying optimal problem representations by identifying key features, resulting in better results than hand-designed representations (features) [22].

The individual extracted frames are feed as input to first layer of pretrained Inception-V3 model and output is derived from layer number 313 (avg_pool layer) freezing the subsequent last three layers of the network (predictions, predictions_softmax, classification layer) as shown in Figure 4.

2.4 Classification using Support Vector Machines

Guided learning models called support-vector machines (SVMs) are used in machine learning for regression and classification. They build models from training examples, assigning new examples to one or the other category, and maximizing the gap between categories using methods like Platt scaling.

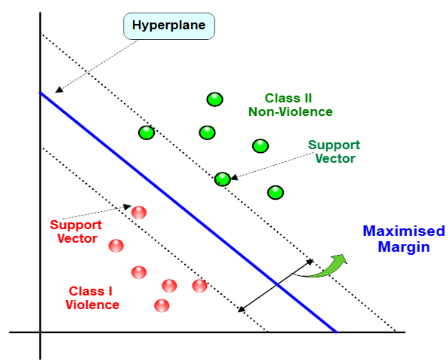

Fig 5. SVM Hyperplane for violence detection

It seeks to establish the optimal decision boundary, as hyperplane, to categorise n-dimensional space. SVM selects extreme points for hyperplane creation using support vectors as shown in Figure 5.

For data that can be divided into two groups using a straight line, a linear support-vector machine classifier, or linear SVM, is utilized. The hyperplane of SVM is the ideal decision threshold for categorizing data points in space with n dimensions. The characteristics of the dataset determine the hyperplane's dimensions, as in case with 2 features it will be a straight line and a 2-dimension plane for 3 features. Support vector are points which are nearest to the hyperplane, affect the position of the hyperplane.

SVM algorithm is used in proposed violence detection algorithm to classify pairs of coordinates in either violence or non-violence class. In a 2-d space, a straight line can separate these classes, but multiple lines can separate them. The algorithm finds the best decision boundary, called a hyperplane, by finding support vectors and maximizing the margin between them. The optimal hyperplane is the hyperplane with the maximum margin.

The extracted features from avg_pool layer of Inception-V3 network is given to support vector machine for training. The model is now trained to categorize the data provided from the testing portion when the training phase is complete. Multiple tests are performed to find out the evaluation parameters of the proposed system.

2.5 Model Evaluation

Multimedia analysis requires performance evaluation to improve methods like classification, regression, detection, and summarization. Computer vision evaluation metrics include mean square error, root mean square, and confusion matrix, with basic evaluation metric studies focusing on TP, TN, FP, and FN.

- True Positive (TP): The number of incidents that have been appropriately categorized as violent is represented by its value.

- False Negative (FN): This metric's value corresponds to the number of impartial videos that were incorrectly categorized as violent programming.
- False positive (FP): The number of nonviolent classes that were mistakenly categorized as violent is represented by this value.
- True negative (TN): The number of normal classes that have been accurately identified as normal is represented by the value.

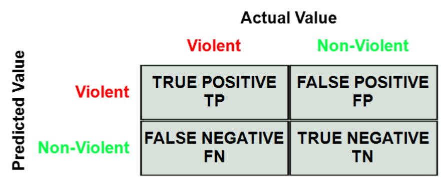

Fig 6. Confusion Matrix for Violent Video Classification Problem

The confusion metric indicates how well the model separates the violent from the non-violent classes as shown in Figure 6. Accuracy measures the number of correct predictions per test sample, indicating a model's training and general performance accuracy. Precision measures the percentage of positive predictions, while recall represents the proportion of correct positives identified correctly. Both metrics are crucial for accurate classification and are calculated as per the formulas in [23].

F1-score measures test accuracy, a Harmonic Mean between precision and recall, with a score between 0 and 1. In present study will focus on calculating the F1-score [9] to check the model behavior using following formula (1).

$$F_1 = (1+\alpha^2)\left[\frac{\text{Precision } \times \text{Re call}}{(\alpha^2 \times \text{ Precision }) + \text{Re call}}\right] \tag{1}$$

Where α is positive real factor, where is chosen such that recall is considered α times as important as precision

3 Result and Discussion

Table 2. Performance Evaluation with variation inSVM Kernel Function

Dataset	SVM Kernel Function	Accuracy	Precision	Recall	F-Score
Hockey Fight Dataset	Linear	94.33	92.30	96	0.935
	Polynomial	**96.33**	**96.02**	**96.66**	**0.961**
Movie Dataset	Linear	**98.33**	**100**	**96.66**	**0.991**
	Polynomial	98.33	96.77	100	0.974

The ROC curve and confusion matrix for Hockey Fight Dataset analyzed with 'Linear' kernel function and for Movie Fight Dataset analyzed with 'Polynomial' kernel function is represented in Figure 7 first row and second row respectively. Table 3 below compares the results of the validation accuracy for the proposed model 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 3. Accuracy Comparisonbetween the Proposed Models and the State-Of-The-Art Techniques

Method	Hockey fight	Movie
Jaiswal et al [23]	90%	91.66%
Soliman et al [24]	95.1%	97%
Xingyu et al [25]	95.40%	-
Karisma et al [26]	92%	-
Seydi Keceli et al [27]	92.90	98.7
Proposed	**96 ± 2%**	**98 ± 2%**

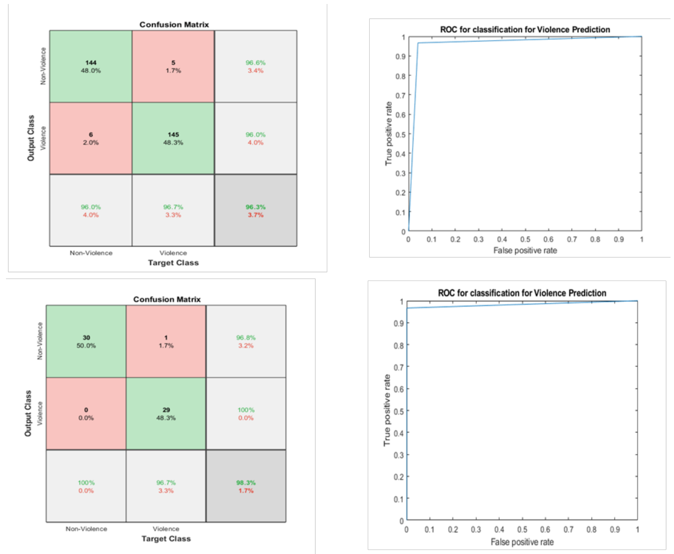

Fig 7. Confusion Matrix and ROC Curve for Hockey Fight Dataset with Linear Kernel Function (First Row) and Movies Dataset with Polynomial Kernel Function (Second Row).

4 Conclusion

A 3D Convolutional Neural Network ArchitectureInception-V3, is used in a novel technique for violence detection in videos to extract motion data. The transfer learning approach for feature extraction and SVM as classifier comes out as better combination in comparison with local features and machine learning model. The method outperforms competing algorithms in person-to-person and crowd battle datasets and achieves higher accuracy. However, mistakes might happen when friendly behaviors are misclassified. The proposed algorithm is having an average accuracy of 96 $\pm$ 2% over both the benchmarking datasets. Future development aims to enhance expertise in developing new datasets and models using video data from surroundings. These models will be trained on sophisticated CNN models like Resnet, LSTM, and reduced gesture recognition for aggression detection.

Acknowledgement

The authors thank the researchers who contributed their efforts into the field of action recognition and violent action detection using video processing that motivated us to work in this field. Authors also thank Sant Gadge Baba Amravati University and G H Raisoni University, Amravati for providing the required technical support for performing the experiments.

References

1) Singh S, Tyagi B. Computational Comparison of CNN Based Methods for Violence Detection. . Available from: https://doi.org/10.21203/rs.3.rs-3130914/v1.
2) 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 Computer Science*. 2022;8:e920. Available from: https://doi.org/10.7717/peerj-cs.920.
3) Min FU, Obaidat MS, Ullah A, Ullah K, Hijji M, Baik SW. A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos. *ACM Computing Survey*. 2023;55. Available from: https://doi.org/10.1145/3561971.
4) Peixoto B, Lavi B, Bestagini P, Dias Z, Rocha A. Multimodal Violence Detection in Videos. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2020;p. 2957–2961. Available from: https://doi.org/10.1109/ICASSP40776.2020.9054018.
5) Kaur G, Singh S. Violence Detection in Videos Using Deep Learning: A Survey. In: Advances in Information Communication Technology and Computing;vol. 392. Springer Nature Singapore. 2022;p. 165–173. Available from: https://doi.org/10.1007/978-981-19-0619-0_15.
6) Mumtaz N, Ejaz N, Habib S, Mohsin SM, Tiwari P, Band SS, et al. An overview of violence detection techniques: current challenges and future directions. *Artificial Intelligence Review*. 2023;56(5):4641–4666. Available from: https://doi.org/10.1007/s10462-022-10285-3.
7) Accattoli S, Sernani P, Falcionelli N, Mekuria DN, Dragoni AF. Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines. *Applied Artificial Intelligence*. 2020;34(4):329–344. Available from: https://doi.org/10.1080/08839514.2020.1723876.
8) Zhou P, Ding Q, Luo H, Hou X. Violence detection in surveillance video using low-level features. *PLOS ONE*. 2018;13(10):e0203668. Available from: https://doi.org/10.1371/journal.pone.0203668.

9) Nova D, Ferreira A, Cortez P. A Machine Learning Approach to Detect Violent Behaviour from Video. In: A, T, editors. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering;vol. 273. Springer International Publishing. 2019;p. 85–94. Available from: https://doi.org/10.1007/978-3-030-16447-8_9.
10) 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. IEEE. 2012;p. 1–6. Available from: https://doi.org/10.1109/CVPRW.2012.6239348.
11) Gao Y, Hong L, Xiaohu S, Wang, Liu Y. Violence detection using Oriented VIolent Flows. 2016. Available from: https://doi.org/10.1016/j.imavis.2016.01.006.
12) Zhang Q, Zhang M, Chen T, Sun Z, Ma Y, Yu B. Recent advances in convolutional neural network acceleration. *Neurocomputing*. 2019;323:37–51. Available from: https://doi.org/10.1016/j.neucom.2018.09.038.
13) Fenil E, Manogaran G, Vivekananda GN, Thanjaivadivel T, Jeeva S, Ahilan A. Real time violence detection framework for football stadium comprising of bigdata analysis and deep learning through bidirectional LSTM. *Computer*. 2019;151:191–200. Available from: https://doi.org/10.1016/j.comnet.2019.01.028.
14) Naik AJ, Gopalakrishna MT. Deep-violence: individual person violent activity detection in video. *Multimedia Tools and Applications*. 2021;80(12):18365–18380. Available from: https://doi.org/10.1007/s11042-021-10682-w.
15) Xia Q, Zhang P, Wang J, Tian M, Fei C. Real Time Violence Detection Based on Deep Spatio-Temporal Features. In: Zhou, J, et al Biometric Recognition CCBR, editors. Biometric Recognition;vol. 10996. Springer International Publishing. 2018;p. 157–165. Available from: https://doi.org/10.1007/978-3-319-97909-0_17.
16) Nievas EB, Suarez OD, García GB, Sukthankar R. Violence Detection in Video Using Computer Vision Techniques. In: Computer Analysis of Images and Patterns;vol. 6855. Springer Berlin Heidelberg. 2011;p. 332–339. Available from: https://doi.org/10.1007/978-3-642-23678-5_39.
17) Jaiswal SG, Mohod SW. 2021. Available from: https://link.springer.com/chapter/10.1007/978-981-15-8335-3_56.
18) Jaiswal SG, Mohod SW. Recapitulating the Violence Detection Systems. 2019. Available from: https://link.springer.com/chapter/10.1007/978-981-15-8335-3_56.
19) Mu G, Cao H, Jin Q. Violent Scene Detection Using Convolutional Neural Networks and Deep Audio Features. In: Cheng, H, editors. Communications in Computer and Information Science;vol. 663. Springer Singapore. 2016;p. 451–463. Available from: https://doi.org/10.1007/978-981-10-3005-5_37.
20) Jaiswal SG, Mohod DSW. Classification Of Violent Videos Using Ensemble Boosting Machine Learning Approach With Low Level Features. *Indian Journal of Computer Science and Engineering*. 2021;12(6):1789–1802. Available from: https://doi.org/10.21817/indjcse/2021/v12i6/211206165.
21) Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2016;p. 2818–2826. Available from: https://doi.org/10.1109/CVPR.2016.308.
22) Himeur Y, Al-Maadeed S, Kheddar H, Al-Maadeed N, Abualsaud K, Mohamed A, et al. Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. *Engineering Applications of Artificial Intelligence*. 2023;119:105698. Available from: https://doi.org/10.1016/j.engappai.2022.105698.
23) Jaiswal SG, Mohod SW, Sharma D. Machine Learning Approach for Violent Video Classification. *Indian Journal of Science & Technology*. 2023;16(34):2709–2718. Available from: https://doi.org/10.17485/IJST/v16i34.1777.
24) Soliman MM, Kamal MH, Nashed MAEM, Mostafa YM, Chawky BS, Khattab D. Violence Recognition from Videos using Deep Learning Techniques. In: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS). IEEE. 2019;p. 80–85. Available from: https://doi.org/10.1109/ICICIS46948.2019.9014714.
25) Xu X, Wu X, Wang G, Wang H. Violent Video Classification Based on Spatial-Temporal Cues Using Deep Learning. In: 2018 11th International Symposium on Computational Intelligence and Design (ISCID). IEEE. 2018;p. 319–322. Available from: https://doi.org/10.1109/ISCID.2018.00079.
26) Karisma, Imah EM, Wintarti A. Violence Classification Using Support Vector Machine and Deep Transfer Learning Feature Extraction. In: 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA). IEEE. 2021;p. 337–342. Available from: https://doi.org/10.1109/ISITIA52817.2021.9502253.
27) Keceli AS, Kaya A. Violent activity classification with transferred deep features and 3d-Cnn. *Signal, Image and Video Processing*. 2023;17(1):139–146. Available from: https://doi.org/10.1007/s11760-022-02213-3.