Computers, Materials & Continua
DOI: 10.32604/cmc.2022.024566
Article

Tech Science Press

A Skeleton-based Approach for Campus Violence Detection

Batyrkhan Omarov[1,2,3,4,*], Sergazy Narynov[1], Zhandos Zhumanov[1,2], Aidana Gumar[1,5] and Mariyam Khassanova[1,5]

[1]Alem Research, Almaty, Kazakhstan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[3]International University of Tourism and Hospitality, Turkistan, Kazakhstan
[4]Suleiman Demirel University, Almaty, Kazakhstan
[5]Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan
*Corresponding Author: Batyrkhan Omarov. Email: batyahan@gmail.com

Received: 22 October 2021; Accepted: 07 December 2021

Abstract: In this paper, we propose a skeleton-based method to identify violence and aggressive behavior. The approach does not necessitate high-processing equipment and it can be quickly implemented. Our approach consists of two phases: feature extraction from image sequences to assess a human posture, followed by activity classification applying a neural network to identify whether the frames include aggressive situations and violence. A video violence dataset of 400 min comprising a single person's activities and 20 h of video data including physical violence and aggressive acts, and 13 classifications for distinguishing aggressor and victim behavior were generated. Finally, the proposed method was trained and tested using the collected dataset. The results indicate the accuracy of 97% was achieved in identifying aggressive conduct in video sequences. Furthermore, the obtained results show that the proposed method can detect aggressive behavior and violence in a short period of time and is accessible for real-world applications.

Keywords: PoseNET; skeleton; violence; bullying; artificial intelligence; machine learning

1 Introduction

The issue of preventing violent situations in the education system is very relevant, as it is worldwide. According to the United Nations, globally, every second child aged 2–17 years faces violence in one form or another every year [1]. In addition, every tenth student in the world is exposed to violence at school, and this figure is growing every year. About a fifth of all violence cases against adolescents and young people are committed in education [2]. Bullying generates numerous destructive phenomena and consequences: it increases the risk of suicidal and auto–aggressive tendencies among adolescents, leads to increased aggression and violence in the group and at school, reduced academic performance, emotional problems-an increased risk of anxiety and depression disorders [3].

Working on this topic for many years, in 1993, the Norwegian psychologist Olweus published a generally accepted definition of bullying among children and adolescents: bullying is a deliberate, systematically repeated aggressive behavior that includes inequality of social power or physical strength [4]. Therefore, bullying is always a challenge to the school as an educational institution. Therefore, measures to prevent bullying are required to be equally systematic, planned, and long-lasting.

One of the possible solutions to the problem of bullying is the automated detection of a scene of violent actions in video surveillance cameras. However, even though violence detection systems in video surveillance are developing, the performance problem remains open [5–7]. Usually, video processing requires high-performance computers, and in many cases, it is impossible to get a quick response [8]. This refers to preventing the use of video violence detection methods for real-world applications. Human skeletal data can now be retrieved from images, and violence detection based on the skeleton is better suited to the systems that require fast processing [9,10].

The reminder of this paper is organized as following: Next section reviews state-of-the-art violence detection systems, and the problem statement is defined. The third section explains the aim and objectives of the study. Forth section describes the human skeleton-based video violence detection method. The fifth section explains the data collection process and presents the results of the investigation. The sixth section discusses results and describes current challenges in violence detection in videos. Finally, the last section concludes the paper and explains plans and challenges in violence detection in video. The relevance of the study is the development of an automated, fast fight detection system in video surveillance cameras based on human skeleton points. The proposed approach allows detecting violent actions in the video without requiring high-processed hardware.

2 Related Works

Violence detection from Surveillance Cameras is an ongoing image processing and computer vision research field that recognizes human behavior and categorizes it into normal and abnormal categories. Abnormal activities are uncommon, like strange human actions in public areas, such as fighting, kicking somebody, running, boxing, fleeing crowds, disputes and assaults, vandalism, and crossing boundaries [11]. The usage of video surveillance to track human behavior is on the rise these days, which helps to avoid suspicious human behavior.

The Lagrangian theory offers comprehensive tools for evaluating non-local, long-term motion information in computer vision. Authors propose a specialized Lagrangian method for automatically identifying violent situations in video footage based on this theory [12]. The authors propose a new feature based on a spatio-temporal model that utilizes appearance, background motion correction, and long-term motion information and leverages Lagrangian direction fields. They use an expanded bag-of-words method in a late-fusion way as a classification strategy on a per-video basis to guarantee suitable spatial and temporal feature sizes. Experiments were conducted in four datasets as “Hockey Fight”, “Violence in Movies”, “Violent Crowd”, and “London Metropolitan Police (London Riots 2011)” datasets. Multiple public benchmarks and non-public, real-world data from the London Metropolitan Police verify the proposed system. Experimental results demonstrated that the implementation of Lagrangian theory is a valuable feature in aggressive action detection and the classification efficiency rose over the state-of-the-art techniques like two-stream convolutional neural network (CNN, ConvNet), Violent Flow (ViF), Space Time Interest Point (STIP), histogram of oriented gradients (HoG), Histogram of optical flow (HOF), Bag of Words (BoW), HoF+BoW with STIP, HOG+BoW with STIP, etc. in terms of accuracy and the Area Under the Curve Receiver Operator Characteristic (AUC-ROC) measure.

Surveillance systems are grappling to identify violence. However, it has not received nearly as much attention as action recognition. Existing vision-based techniques focus primarily on detecting violence and make little attempt to pinpoint its location. A new approach in [13] presented a quick and robust method for identifying and localizing violence in surveillance situations to address this issue. Firstly, a Gaussian Model of Optical Flow (GMOF) is suggested for this purpose to extract potential violent areas, which are adaptively modeled as a departure from the usual crowd behavior seen in the picture. Following that, each video volume is subjected to violence detection by intensively sampling the potential violent areas. The authors also propose a new descriptor called the Orientation Histogram of Optical Flow (OHOF), which is an input into a linear SVM to differentiate violent events from peaceful ones. Experimental results on violent video datasets like "Hockey", "BEHAVE", "CAVIAR" have shown the superiority of the proposed methodology over the state-of-the-art descriptors like The scale-invariant feature transform (SIFT) and Motion SIFT (MoSIFT), HOG, HOF, and Combination of HOG and HOF (HNF), in terms of detection accuracy, AUC-ROC, and processing performance, even in crowded scenes.

Shallow modeling approaches cannot learn features independently and instead rely on manual methods to extract features that must be fed into a shallow network for classification [14]. Shallow models are best suited for supervised learning, which requires marked data. The major disadvantage of this modeling approach is that it does not automatically adjust to dynamic changes. The labeling procedure may also be labor-intensive. Some state-of-the-art studies present a video violence detection descriptor that simulates crowd behavior for violence detection by embedding variations in audience texture, applying temporal representations of gray-level co-occurrence matrix data, and using a random forest classifier with k-fold cross-validation [15]. In UCF Violent Flows (ViF), and UMN datasets, their approach surpasses the state-of-the-art findings. Likewise, [16] utilized an improved Fisher vectors (IFVs) extension to describe films utilizing local characteristics and spatio-temporal locations for aggressive behavior identification and analysis. In four publicly accessible datasets, their findings showed substantial improvement.

Unlike shallow machine learning models, most of the deep learning models do not require a special feature extractor since they use the feature learning method to learn their features from the supplied data and categorize them [17]. Nevertheless, unlike edge learning, the collected features may also be fed into support vector machines (SVM) and other shallow model classifiers as input. Utilizing features through handcrafted feature descriptors and providing them to a deep classifier is another method to build deep models [18].

Sharma et al. applied deep learning techniques for video-based violence detection problem [19]. Authors uses pre-trained ResNet-50 architecture to extract necessary features from videos, and send them to ConvLSTM block. The proposed approach was tested by using three different datasets as KTH dataset, Hockey fight dataset, and Violent-Flows dataset. The results shown accuracy between 59% and 90% accuracy depending on dataset and hyperparameters types. Some models can be used with both supervised and unsupervised machine learning techniques, although they are more suited to the latter. After all, they operate with unmarked data that necessitate large amounts of data and processing power. To train the regular patterns in the training movies for physical violence detection, a convolutional spatio-temporal autoencoder is proposed [20]. Even while the model can detect anomalous occurrences and is noise-resistant, there may be more false alarms depending on the complexity of the behavior. Another approach to this model is a convolutional long short-term memory (CLSTM) used to train a violence detection model [21]. When compared to other state-of-the-art methods, their suggested method showed promise on the data sets they used. Tab. 1 demonstrates comparison of state-of-the-art researches in video-based violence detection problem.

Table 1: Video based violence detection methods

Study	Approach	Method	Features	Scene type	Accuracy
Fenil et al., 2019, [11]	Framework for stadium comprising of big data analysis through bidirectional LSTM	Bidirectional LSTM	HOG, SVM	Crowded	94.5
Senst et al., 2017, [12]	Lagrangian fields of direction and bags of word framework to recognize the violence in videos	Lagrangian theory and STIP method for extract motion features	Late fusion for classification	Crowded	91% to 94%
Zhang et al., 2016, [13]	GMOF framework with tracking and detection module	Support vector machine, Gaussian mixture model	OHFO for optical flow extraction	Crowded	82%–89%
Yao et al., 2021, [18]	Multiview fight detection method	Random forest	Optical flow	Crowded, uncrowded	97.66%
Sharma et al. [19]	Deep learning for violence detection	ResNet-50 with ConvLSTM block	Data augmentation, CNN retrain	Crowded, uncrowded	87.5%

Many scholars have used the following broad procedures to build intelligent surveillance applications for detecting aberrant human behaviors. First step is detection of foreground objects. Background subtraction is a robust technique for detecting and extracting foreground items from a series of frames. Next stage is object detection. Object identification in video frames may be accomplished using either non-tracking or tracking-based methods. A tracking-based method is used to determine an object's trajectory over time by identifying its location in each video frame. After identification of objects feature extraction is applied. For object identification, different methods extract shape and motion-based characteristics of the item, and its feature vector is occasionally given as input to the classifier. Final stage is classification. Object categorization is a method of distinguishing the various items in a movie. This mechanism aids in determining between multiple things such as humans, vehicles, and so on. Support Vector Machine, Haar-classifier, Bayesian, K-Nearest Neighbor, Skin color detection, and Face recognition are some methods used to classify things.

3 The Aim and Objectives of the Study

The study aims to develop a system for the fast detection of violent scenes in video surveillance cameras. The scientific novelty of this work is the development of a quick violence detection method by training the neural network based on human skeleton points extracted by PoseNET. To achieve this aim, the following objectives are accomplished: a) Collected video dataset of one-person human violent actions; b) Human skeleton points were extracted from the video frames for further transmission to the neural network; c) Implemented a video violence detection system by training the neural network using the extracted human skeleton points and testing the implemented system using various neural network performance test measures such as detection accuracy, confusion matrix, and results visualization.

4 Data

The issue of detecting violence and aggressive behavior is divided into a number of sub-tasks. The flowchart of the research is shown in Fig. 1. The research flowchart is divided into main parts: data characteristics, Data collection, and Classification. The data characteristics section defines the aggressor's pattern parameters. The data collection section ensures the supply of necessary video data, marks up videos by classes, saves them in .json format, and cuts the marked video scenes containing violent scenes to create a dataset. Finally, the classification section provides a classification of the videos into violence and non-violence. This section consists of subsections as data preparation and preprocessing, feature extraction, model training, and testing.

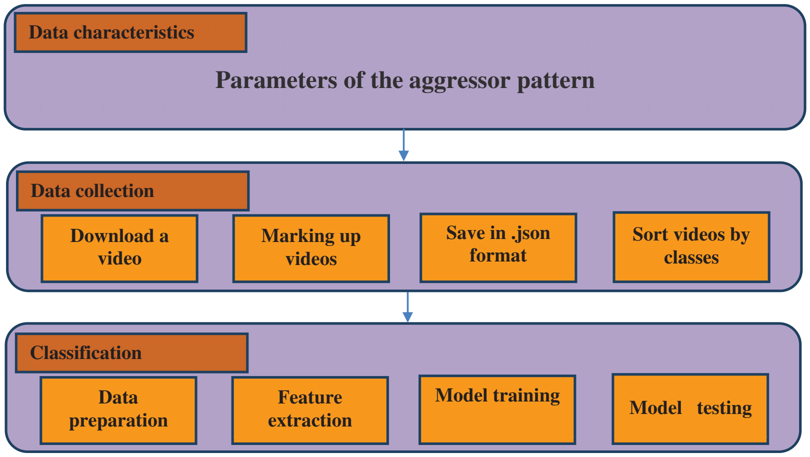


Figure 1: Flowchart of the research

4.1 Data Collection Criteria

The primary stage is to identify the types of data that should be gathered. We came up with four different kinds of characteristics to characterize a victim and a bully. At first, we identified factors of a victim and a bully based on the predefined classifications that should be evaluated throughout the data gathering procedure. Then, we created 13 categories to characterize the features of the victim and bully conduct.

4.2 Data Collection Process

We used terms like "aggression," "physical aggression," "violence," "bullying," "fight," "group fight," and others to search for films accessible in free access on the Internet and on social networks. After collecting them, we classified spatiotemporal portions inside movies with suitable classes, and the labeling information was stored in *.json format. VGG Image Annotator was used to do this. After the tagging was completed, all of the films were clipped and organized into courses.

4.3 Dataset

In the first part of our research, we collected videos of one-person violent actions. Violent actions are classified into thirteen classes. In the total research, there were identified 80 classes that belongs to an aggressor and a victim. In order to realize training the model we have to identify actions of one person. For this purpose, we divided 13 classes that can be realized by one person. Tab. 2 illustrates the thirteen classes that used to train the model. In our study, we created own dataset that consists of the predefined thirteen classes. Our proposed model was trained and tested by using the own dataset. After that, it was tested by using the open video violence datasets.

Table 2: Explanation of each keypoint that PoseNET can retrieve

Class id	Class type
0	Large amplitude
1	Head raised
2	Body facing the victim
3	Shoulders straight, arms back
4	Hands on hips
5	Takes off his outer clothing
6	Kick
7	Punch
8	Covers the face
9	Foots pointing in different directions
10	Bouncing in place during a series of punches
11	Bent over
12	Finger pointing

Fig. 2 demonstrates information about the collected videos. Videos are collected in three formats like .mp4, .mov, and .wmv. Fig. 2a demonstrates statistics regarding the video data file types that were gathered. Fig. 2b depicts the distribution of video data. A total of 2,093 video clips depicting instances of bullying and violent conduct were gathered. The video data was gathered for about 20 h. The following are the file formats in which the video data was collected:

- video in .mp4 format: 2017 files;
- video in .mov format: 44 files;
- video in .wmv format: 32 files.

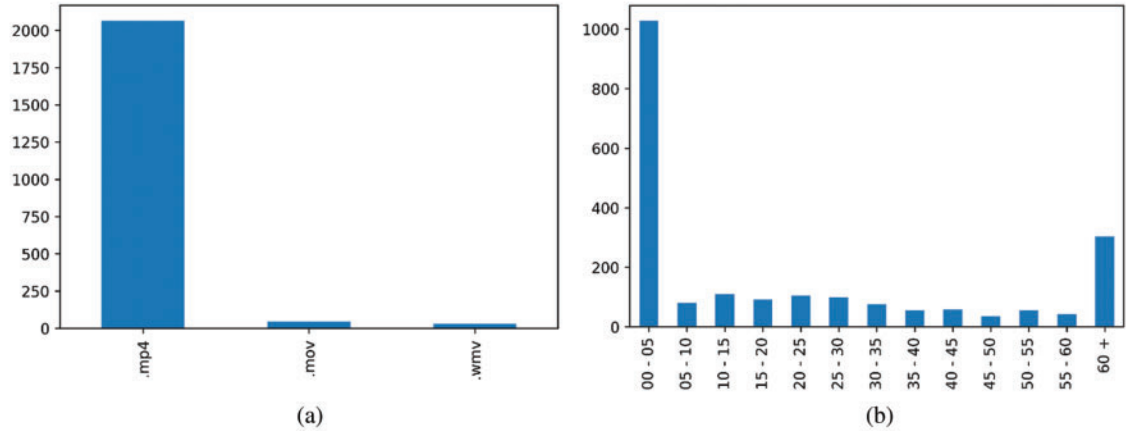


Figure 2: Collected video data. (a) Collected video data by file formats (b) Distribution of the gathered video by types

In addition to the data gathered, we recorded films of one individual imitating violent behaviors. For the first phases of training machine learning models, more videos are needed. The new video material lasts approximately 400 min in total. There are two types of violent videos depending on the goal. One type of video is crowd violence that participated four or more people in fighting, the second one is uncrowded violence that can be done between two people where one of the participants is an aggressor and the second one is a victim.

5 Materials and Methods

5.1 The Proposed Approach

In this section, we describe our approach that is the skeleton-based violence detection. Fig. 3 demonstrates the overall architecture of the proposed system. The system consists of three subtasks. Firstly, we apply the PoseNET model to input video frames to estimate human pose on each video frame. In the second stage, we extract key points as vectors from each video frame. PoseNET returns 17 key points per frame. Consequently, we get vectors that contain 34 elements. In the next stage, we concatenate each k vector into one vector and send it to the feature learning and activity recognition step. Finally, in the third stage, we train a convolutional neural network to violence detection problems. Top-down and bottom-up algorithms are two types for identifying a human body position based on RGB pictures. The first ones activate a human detector and assess bodily joints in bounding boxes that have been identified. PoseNET [22], HourglassNet [23], and Hornet [24] are examples of top-down approaches. Open space [25] and PifPaf [26] are some of the bottom-up algorithms.

We utilized training based on a skeleton methodology. The provided method has the potential to lower computing expenses. To produce an accurate evaluation of the aggressor's or victim's figure, PoseNET based neural network is used. A function extractor may transfer gained knowledge from the source domain to the destination domain using a pre-trained PoseNET. The PoseNET output depicts the human body with 17 main body points and their locations and confidences. The nose, eyes, ears, shoulders, elbows, wrists, thighs, knees, and ankles are 17 essential points. Fig. 4 shows an example of 17 critical points that PoseNET could retrieve that is applied to feed the neural network. The key points are represented as x and y coordinates in the two-dimensional coordinate space. Tab. 3 demonstrates an explanation of each key point that can be extracted by the PoseNET model.

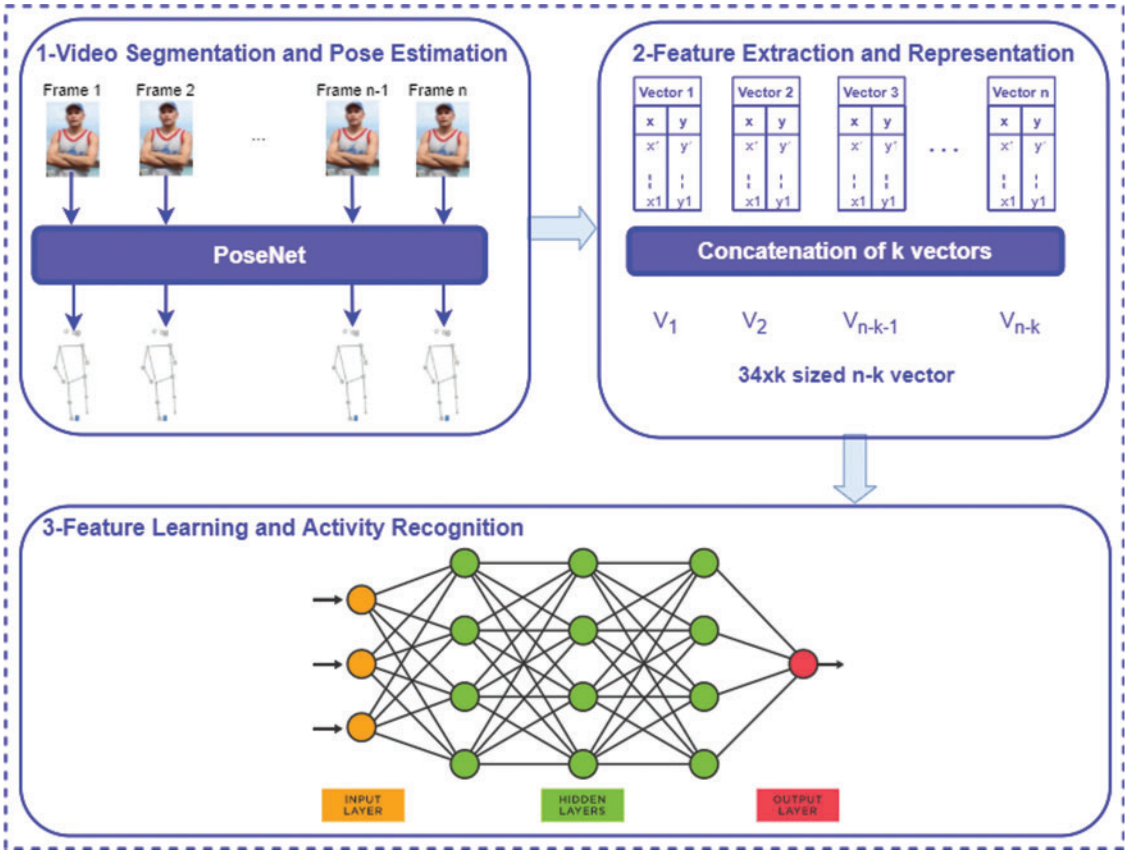


Figure 3: Flowchart of violent action detection framework

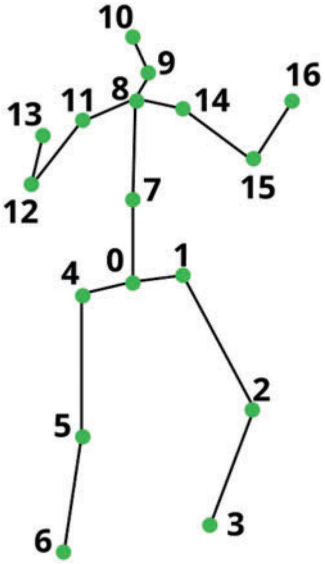


Figure 4: 17 key points extracted by PoseNET

Table 3: Explanation of each keypoint that retrieved by PoseNET

Keypoint if	Explanation
0	Bottom torso
1	Left hip
2	Left knee
3	Left foot
4	Right hip
5	Right knee
6	Right foot
7	Center torso
8	Upper torso
9	Neck base
10	Center head
11	Right shoulder
12	Right elbow
13	Right hand
14	Left shoulder
15	Left elbow
16	Left hand

The human body may be represented in the following way:

$$r_b(x_i;\theta), \tag{1}$$

where θ is neural network parameters, and x_i is training samples of the data set. A fully connected neural network layer is deployed to classify the representation of the human body $rb(x_i;\ \theta)$. Before being normalized by the "Softmax" layer, the additional neural network can be trained by reducing category cross-entropy loss. The architecture of the PoseNET ANN is shown in Fig. 5. First, human action frames are sent to PoseNET to extract key points. Then, skeleton points are represented in feature space by indicating coordinates. After that, the neural network is trained using the human skeleton key points.

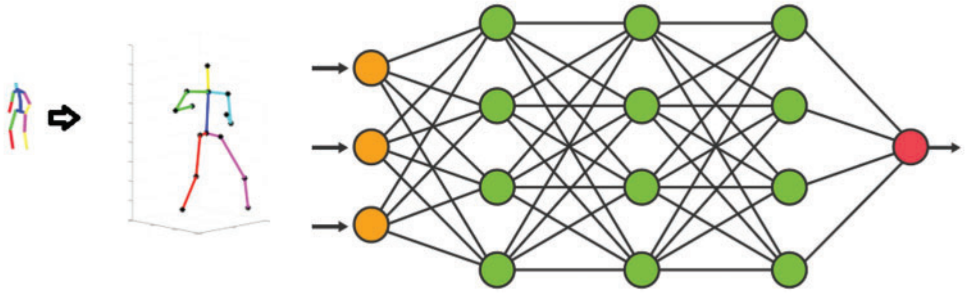


Figure 5: ANN for PoseNET based violence detection architecture

Thus, we collect the necessary data, preprocessed and divided by classes the collected data, prepared a dataset to feed the neural network in the first part of the study. The second part of the research is the extraction of human skeleton points applying PoseNET. Human skeleton points are used to train a neural network in order to recognize human actions. The final component of the proposed approach is developing a neural network for violent action detection that follows by training and testing the results in order to identify weather the proposed approach acceptable of practical use or not.

5.2 Evaluation

A confusion matrix is used to display the results of a prediction model. Actual classes are represented by columns, whereas rows of the matrix represent predicted classes [27]. For each class c, the matrix shows the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) values. The confusion matrix is used to calculate many efficiency metrics, including accuracy, precision, recall, and F1-score [28]. Eqs. (2)–(5) illustrates formulas as precision, recall, F2-score, and accuracy that are applied to evaluate the results of the proposed approach.

$$precision = \frac{TP}{TP + FP}, \tag{2}$$

$$recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}, \tag{4}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{5}$$

We utilized the weight-averaging method to integrate metrics computed for each class into a single variable that weights values in results according to class percentage. To verify prediction models, we utilized a conventional train/test split [29]. The dataset divided as 80% to 20%. Eighty percent of the data was used during training, while twenty percent was used to test the model.

6 Experiment Results

In this section, we demonstrate the results of data collection, feature extraction, and violence detection problems. First subsection represents human skeleton points' extraction results, next subsection demonstrates violent actions detection results. In the end of the second subsection, we compare the obtained results with the state-of-the-art research results. The obtained results are presented by using the evaluation parameters as confusion matrix, model accuracy, precision, recall, and F1-score.

6.1 Human Skeleton Points Extraction

In this section, we extracted human skeleton points in the video stream. PoseNET model was applied to extract 17 key points. Fig. 6 demonstrates extracted human skeleton points in a video stream frame. Human key points extraction was provided in a period of every one second of the video by shot one frame. As a video stream changes quickly, in the case of fights, the position of fight participants can be changed promptly. Consequently, there can be several sequenced classes of the each fight participant. Thus, in video violence detection, fast decision-making is critical.

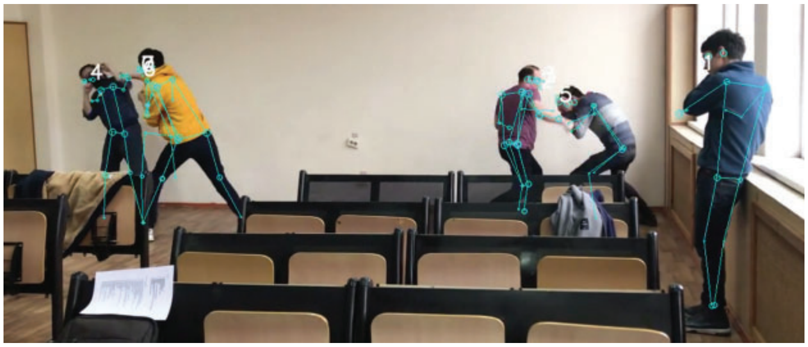
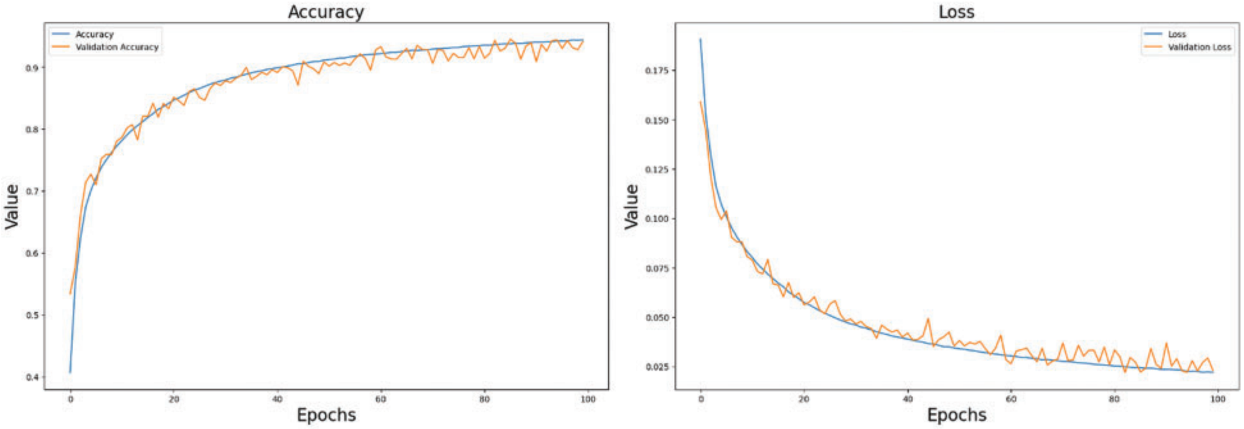


Figure 6: Testing the proposed model by indicating human skeleton points. model

6.2 Detection of Violent Actions

We built and evaluated machine learning models for violence detection throughout the trials. The PoseNET structure was used to train machine learning software models based on neural networks. We chose 13 classes from the labelled video data for which we recorded additional video data. The action recognition model developed using that data performed very well in identifying aggressive behavior.

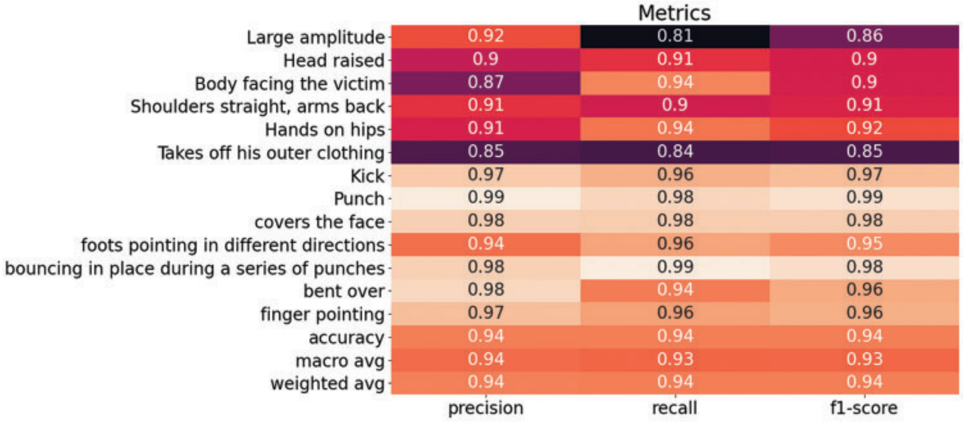
Fig. 7 demonstrates the results of the proposed model testing. Fig. 7a shows neural networks' validation and test accuracy for physical bullying detection during eight epochs of learning. The findings indicate that after eight epochs of training, accuracy approaches 98%. Fig. 7b shows the values of the neural network loss function throughout eight training epochs. The findings indicate that even during the initial period of training, validation loss is extremely minimal.



(a) (b)

Figure 7: Model testing. (a) Model accuracy (b) Test and validation loss

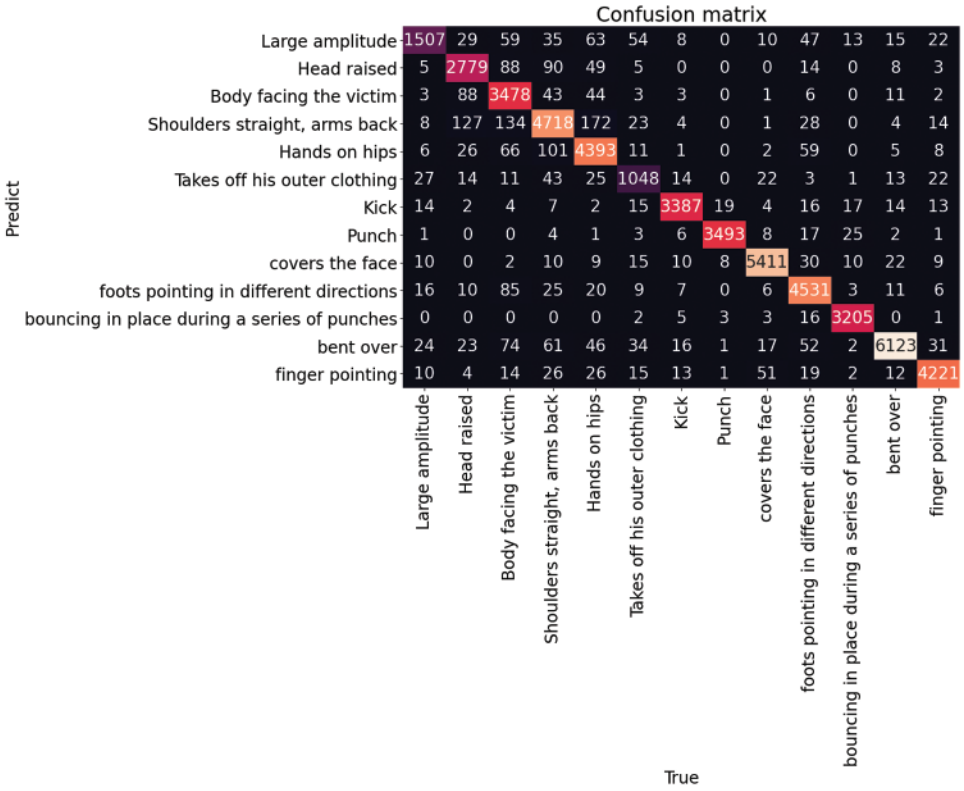
The assessment of categorization results for 13 classes is shown in Fig. 8. All of the assessment parameters, as can be seen, are of excellent quality. For example, the precision ranges from 0.92 to 0.98, the recall ranges from 0.89 to 1.0, and the F1-score ranges from 0.92 to 0.99. Fig. 9 depicts a confusion matrix for 13 different types of physical violence. The confusion matrix shows a very high classification rate and minimum confusion between classes.



Metrics

	precision	recall	f1-score
Large amplitude	0.92	0.81	0.86
Head raised	0.9	0.91	0.9
Body facing the victim	0.87	0.94	0.9
Shoulders straight, arms back	0.91	0.9	0.91
Hands on hips	0.91	0.94	0.92
Takes off his outer clothing	0.85	0.84	0.85
Kick	0.97	0.96	0.97
Punch	0.99	0.98	0.99
covers the face	0.98	0.98	0.98
foots pointing in different directions	0.94	0.96	0.95
bouncing in place during a series of punches	0.98	0.99	0.98
bent over	0.98	0.94	0.96
finger pointing	0.97	0.96	0.96
accuracy	0.94	0.94	0.94
macro avg	0.94	0.93	0.93
weighted avg	0.94	0.94	0.94

Figure 8: Evaluation of classification results



Confusion matrix

Predict / True

	Large amplitude	Head raised	Body facing the victim	Shoulders straight, arms back	Hands on hips	Takes off his outer clothing	Kick	Punch	covers the face	foots pointing in different directions	bouncing in place during a series of punches	bent over	finger pointing
Large amplitude	1507	29	59	35	63	54	8	0	10	47	13	15	22
Head raised	5	2779	88	90	49	5	0	0	0	14	0	8	3
Body facing the victim	3	88	3478	43	44	3	3	0	1	6	0	11	2
Shoulders straight, arms back	8	127	134	4718	172	23	4	0	1	28	0	4	14
Hands on hips	6	26	66	101	4393	11	1	0	2	59	0	5	8
Takes off his outer clothing	27	14	11	43	25	1048	14	0	22	3	1	13	22
Kick	14	2	4	7	2	15	3387	19	4	16	17	14	13
Punch	1	0	0	4	1	3	6	3493	8	17	25	2	1
covers the face	10	0	2	10	9	15	10	8	5411	30	10	22	9
foots pointing in different directions	16	10	85	25	20	9	7	0	6	4531	3	11	6
bouncing in place during a series of punches	0	0	0	0	0	2	5	3	3	16	3205	0	1
bent over	24	23	74	61	46	34	16	1	17	52	2	6123	31
finger pointing	10	4	14	26	26	15	13	1	51	19	2	12	4221

Figure 9: Confusion matrix for multi-classification

Fig. 10 shows how the pre-trained neural network may be used in a group combat scenario. Finally, we detect the activity of each person, classify them, and determine their position, type of action, aggressor or victim in real-time. This type of demonstration of results can be convenient for video operators to detect the fighting of physical violence in real time and quickly detect the aggressor and victim in crowded and uncrowded scenes of violence.

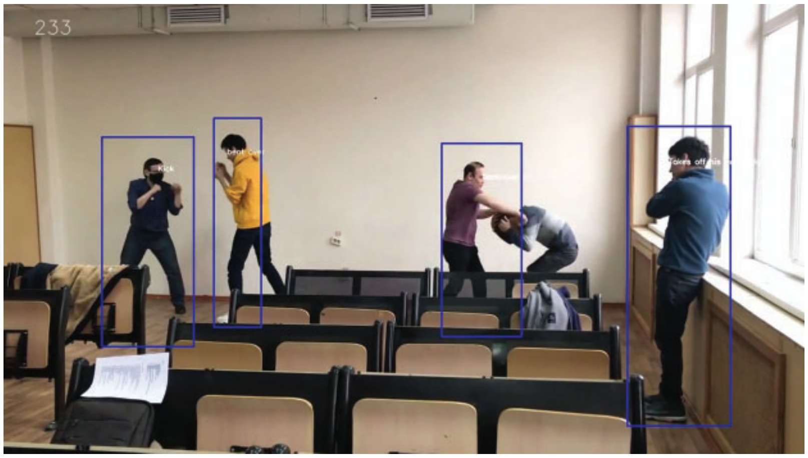


Figure 10: Testing the proposed model

Tab. 4 compares the obtained results with the state-of-the-art research results. We compared the violence detection researches by three main evaluation parameters as precision, recall, and f-score. However, many studies do not use recall and F-score evaluation parameters. In such cases, Accuracy is the main evaluation parameter to compare the performances of the proposed approaches. In addition, most studies do not show the processing time of their approaches, as it is inconvenient because of differences in datasets and the performance of computing equipment.

Table 4: Comparison of the received results

Study	Approach	Precision	Recall	F-score
The proposed approach	**Skeleton based violence detection**	**0.94**	**0.93**	**0.93**
Fenil et al., 2019, [11]	Bidirectional LSTM	0.94	-	-
Senst et al., 2017, [12]	Scale-sensitive video-level representation	0.91–0.94	-	-
Zhang et al., 2016, [13]	Linear SVM	0.82–0.89	-	-
Sharma et al., 2020 [19]	ResNet-50 and ConvLSTM	0.924	-	-
Cheng et al., 2020 [30]	Flow gated network	0.8725	-	-
Carneiro et al., 2019 [31]	Multi-stream CNN	0.8910	-	-
AlDahoul et al., 2021 [32]	CNN-LSTM based model	0.7335	0.7690	0.7401
Deepak et al., 2020 [33]	Autocorrelation of gradients based violence detection	0.91	0.88	0.88

The results show that the proposed system can be applied to real-time practical applications for violence detection using surveillance video. Applying skeleton points to train and test the neural network makes the proposed approach faster than the model that uses only images. Moreover, the proposed system will be helpful in different places, for example, in educational institutions such as schools, universities, kindergartens, shopping malls, and other areas equipped with video surveillance cameras.

7 Discussion

This research developed a skeleton-based violence detection in videos that allows the proposed model to be used in real-time and does not require high-processed hardware. The main advantage of the proposed system is as follows. Firstly, there is no need to feed the system with large video streams and images. Through the use of PoseNET based key points of the human skeleton, our system can operate faster than existing systems. In its case, this feature makes the proposed system capable for real-time real-world applications.

One of the limitations of the proposed system is the probability of confusion between people's identities during a long video. In addition, there is a possibility of changing identities between people in crowded video scenes. In further research, we will solve this problem by applying the Deepsort technique [34]. This tracking-by-detection algorithm takes into account both the parameters of the bounding box of the detection results and information about the appearance of the tracked objects in order to associate detections in a new frame with previously tracked objects [35].

We may see the reduction in frames per second as a drawback of research, which affects the speed of recognition of machine learning algorithms. Therefore, parallel computing using graphics processing units' performance is considered for further study to process a high volume of videos and enhance algorithm performance and speed.

8 Conclusion

The proposed research is aimed at rapid detection of violent actions by video surveillance cameras in real-time. To achieve this goal, we introduce three proposals as follows: Classify violent actions in a video stream that contains two types of violent actions. The first part of the dataset contains violent actions of a single person that last more than 400 h. The single person violent actions were divided into 13 classes, and the videos in the dataset were filmed from different angles and collected using different devices. The second part of the dataset contains crowd violent actions. Single person violent actions are applied for neural network training, crowd violent actions are used to test the proposed system.

To fetch an artificial neural network with skeleton key points instead of high-volume video, we extracted skeleton points by applying the PoseNET model. Skeleton points were extracted from the video frames with a period of 1 s. Because of using human key points, there is no need to load a vast amount of video frames or images. Instead, it is enough to send the coordinates of the key points as input parameters of the neural network. Finally, we created an artificial neural network for violence detection in video. The extracted key points are utilized as input parameters of the neural network for violence detection. In own case, it allows detecting violent actions in real-time without requiring high-performance computers or servers. The developed system is capable of detecting 13 classes of violent actions, and it can be used in video surveillance cameras to ensure the safety of people. The experiment results show an accuracy of 95%–99% in video-based violence detection that proves the proposed system's applicability for practical use.

In further research, we are going to apply tracking-by-detection algorithms as Deepsort and MoveNET in order to prevent confusion of people's identities in the crowd violent scenes. In addition, we are going to apply the performance of graphics processing units to handle high volume video with high quality video in real time.

Funding Statement: This work was supported by the grant "Development of artificial intelligence-enabled software solution prototype for automatic detection of potential facts of physical bullying in educational institutions" funded by the Ministry of Education of the Republic of Kazakhstan. Grant No. IRN AP08855520.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Philpot, L. Liebst, K. Møller, M. Lindegaard and M. Levine, "Capturing violence in the night-time economy: A review of established and emerging methodologies," *Aggression and Violent Behavior*, vol. 46, no. 1, pp. 56–65, 2019.

[2] A. Ross, S. Banerjee and A. Chowdhury, "Security in smart cities: A brief review of digital forensic schemes for biometric data," *Pattern Recognition Letters*, vol. 138, no. 1, pp. 346–354, 2020.

[3] I. Rodríguez-Moreno, J. Martínez-Otzeta, B. Sierra, I. Rodriguez and E. Jauregi, "Video activity recognition: State-of-the-art," *Sensors*, vol. 19, no. 14, pp. 3160, 2020.

[4] G. Sreenu and M. Durai, "Intelligent video surveillance: A review through deep learning techniques for crowd analysis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–27, 2019.

[5] P. Vennam, T. Pramod, B. Thippeswamy, Y. Kim and P. Kumar, "Attacks and preventive measures on video surveillance systems: A review," *Applied Sciences*, vol. 11, no. 12, pp. 5571, 2021.

[6] R. Nawaratne, D. Alahakoon, D. De Silva and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2019.

[7] A. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018.

[8] B. Omarov, B. Omarov, S. Shekerbekova, F. Gusmanova, N. Oshanova *et al.,* "Applying face recognition in video surveillance security systems," in *Int. Conf. on Objects, Components, Models and Patterns*, TOOLS 2019, Innopolis, Russia, pp. 271–280, 2019.

[9] H. Pham, L. Khoudour, A. Crouzil, P. Zegers and S. Velastin, "Exploiting deep residual networks for human action recognition from skeletal data," *Computer Vision and Image Understanding*, vol. 170, no. 1, pp. 51–66, 2018.

[10] Z. Shao, J. Cai and Z. Wang, "Smart monitoring cameras driven intelligent processing to big surveillance video data," *IEEE Transactions on Big Data*, vol. 4, no. 1, pp. 105–116, 2017.

[11] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva *et al.,* "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Computer Networks*, vol. 151, pp. 191–200, 2019.

[12] T. Senst, V. Eiselein, A. Kuhn and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2945–2956, 2017.

[13] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang *et al.,* "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications*, vol. 75, no. 12, pp. 7327–7349, 2016.

[14] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang *et al.,* "Weakly-supervised crowd counting learns from sorting rather than locations," in *Computer Vision–ECCV 2020: 16th European Conf.*, Glasgow, UK, pp. 1–17, 2020.

[15] K. Lloyd, P. Rosin, D. Marshall and S. Moore, “Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures,” *Machine Vision and Applications*, vol. 28, no. 1, pp. 361–371, 2017.

[16] P. Bilinski and F. Bremond, “Human violence recognition and detection in surveillance videos,” in *2016 13th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Colorado Springs, CO, pp. 30–36, 2016.

[17] M. Karim, M. Razin, N. Ahmed, M. Shopon and T. Alam, “An automatic violence detection technique using 3D convolutional neural network,” *Sustainable Communication Networks and Application*, vol. 55, no. 1, pp. 17–28, 2021.

[18] A. Naik and M. Gopalakrishna, “Deep-violence: Individual person violent activity detection in video,” *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18365–18380, 2021.

[19] M. Sharma and R. Baghel, “Video surveillance for violence detection using deep learning,” in *Advances in Data Science and Management*, ICDSM 2019, Singapore, pp. 411–420, 2020.

[20] M. Asad, J. Yang, J. He, P. Shamsolmoali and X. He, “Multi-frame feature-fusion-based model for violence detection,” *The Visual Computer*, vol. 37, no. 6, pp. 1415–1431, 2021.

[21] Y. Chong and Y. Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” *Lecture Notes in Computer Science*, vol. 10262, no. 1, pp. 189196, 2017.

[22] B. Schmidt and L. Wang, “Automatic work objects calibration via a global–local camera system robot,” *Computer-Integrated Manufacturing*, vol. 30, no. 1, pp. 678–683, 2014.

[23] A. Newell, K. Yang and J. Deng, “Stacked hourglass networks for human pose estimation,” in *European Conf. on Computer Vision*, ECCV 2016, Amsterdam, The Netherlands, pp. 483–499, 2016.

[24] K. Shrikhande, I., White, D. Wonglumsom, S. Gemelos, M. Rogge *et al.,* “HORNET: A packet-over-WDM multiple access metropolitan area ring network,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, pp. 2004–2016, 2000.

[25] A. Zanchettin, N. Ceriani, P. Rocco, H. Ding and B. Matthias, “Safety in human-robot collaborative manufacturing environments: Metrics and control,” *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 882–893, 2016.

[26] A. Hornung, K. Wurm, M. Bennewitz, C. Stachniss and W. Burgard, “OctoMap: An efficient probabilistic 3D mapping framework based on octrees,” *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[27] B. Omarov, N. Saparkhojayev, S. Shekerbekova, O. Akhmetova, M. Sakypbekova *et al.,* “Artificial intelligence in medicine: Real time electronic stethoscope for heart diseases detection,” *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2815–2833, 2022.

[28] M. Murzamadieva, A. Ivashov, B. Omarov, B. Omarov, B., Kendzhayeva *et al.,* “Development of a system for ensuring humidity in sport complexes,” in *2021 11th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, Uttar Pradesh, India, pp. 530–535, 2021.

[29] J. Cai, J. Luo, S. Wang and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2017.

[30] M. Cheng, K. Cai and M. Li, “Rwf-2000: An open large scale video database for violence detection,” in *2020 25th Int. Conf. on Pattern Recognition (ICPR)*, Milan, Italy, pp. 4183–4190, 2021.

[31] S. Carneiro, G. da Silva, S. Guimaraes and H. Pedrini, “Fight detection in video sequences based on multi-stream convolutional neural networks,” in *IEEE SIBGRAPI Conf. on Graphics, Patterns and Images (SIBGRAPI)*, Rio Grande do Sul, Brazil, pp. 8–15, 2019.

[32] N. AlDahoul, H. Karim, R. Datta, S. Gupta, K. Agrawal *et al.,* “Convolutional neural network-long short term memory based IOT node for violence detection,” in *2021 IEEE Int. Conf. on Artificial Intelligence in Engineering and Technology (IICAIET)*, Kota Kinabalu, Malaysia, pp. 1–6, 2021.

[33] K. Deepak, L. Vignesh and S. Chandrakala, “Autocorrelation of gradients based violence detection in surveillance videos,” *ICT Express*, vol. 6, no. 3, pp. 155–159, 2020.

[34] C. Duan and X. Li, "Multi-target tracking based on deep sort in traffic scene," *Journal of Physics: Conference Series*, vol. 1952, no. 2, pp. 022074, 2021.

[35] A. Pramanik, S. Pal, J. Maiti and P. Mitra, "Granulated RCNN and multi-class deep sort for multi-object detection and tracking," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 1, no. 1, pp. 1–11, 2021.