ISSN: 2586-7652 (Print) Vol. 06, No. 01, March 2023
ISSN: 2635-7607 (Online)
International Journal of Advanced Engineering
Source: http://ictaes.org
Manuscript received: March 1, 2023; Revised: March 17, 2023; Accepted: March 20, 2023

Real-Time Detection of School Violence using Machine Learning and Human Skeleton Tracking

Trong Nguyen[1], Phi Ta[2], Thai Yen[3], Phung Tang[4], Tai Lam[5], Hau Le[6], and Nha Tran[7]

[1, 2, 3, 4, 5, 7]University of Education, Ho Chi Minh city, Viet Nam
6Bình Duong University, Binh Duong, Viet Nam

4601104200@student.hcmue.edu.vn[1], 4701104156@student.hcmue.edu.vn[2], 4701104250@student.hcmue.edu.vn[3], 4601103055@student.hcmue.edu.vn[4], lamtai2105@gmail.com[5], hault.cm@gmail.com[6], nhatt@hcmue.edu.vn[7]

Abstract

Detecting violence is an important task with many potential applications in various fields, such as security monitoring, school violence detection, and more. In this study, we propose a model capable of real-time violence situation recognition. Specifically, we apply a pre-trained YoloV8 model to detect each person in the input video, and use MediaPipe to extract the skeleton of each person. Then, we train a Long Short Term Memory (LSTM) model on the skeleton data to learn how to recognize violent situations. To evaluate the performance of our method, we test it on two datasets, Hockey Fight and RWF-2000, containing violent and non-violent interactions in videos. Our experimental results show that our method achieves the highest performance on these two datasets. Our approach can be widely applied to other human pose analysis tasks, including more complex situations involving multiple individuals and interactions.

Keywords : *Computer Vision, Violence Detection, Video classification, Action Recognition, Behavior Detection.*

1. Introduction

Violence has long been a serious social issue, stemming from many different causes including social vices, economic problems, personal conflicts, and more. However, regardless of its underlying causes, violence remains an unacceptable behavior. In recent years, the increase in global-level violent incidents has highlighted the urgency of detecting and preventing violence. With the widespread use of surveillance camera systems, developing a model capable of identifying violent situations is a practical task. Traditional techniques for

Corresponding Author : Nha Tran
Author’s affiliation : Ho Chi Minh University of Education
Email: nhatt@hcmue.edu.vn

detecting violence, such as manual monitoring, are costly, time-consuming, and prone to errors. However, the development of public and private surveillance camera systems, along with the emergence of artificial intelligence (AI), has provided many promising solutions for this challenge.

Detecting violence is an important task in the fields of computer vision and security. It involves identifying violent activities or behaviors in various environments, such as public places, schools, and prisons. In recent years, many deep learning techniques have been proposed that have shown promising results for this task [1-3]. These methods require the use of deep learning algorithms to recognize specific body positions or movements related to violence, such as punching, kicking, or shoving. Additionally, features such as facial expressions and tone of voice can also be exploited to identify violent behavior. However, detecting violent interactions in videos is a very challenging task due to variations in human posture, camera angles, lighting conditions, and the presence of obstacles. Recently, many deep learning models have been proposed to address this challenge, with some approaches showing promising potential, such as models based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Although CNNs excel in image classification tasks, they may struggle to capture the temporal dependency between frames in videos, thus leading to the loss of contextual information and decreased accuracy in identifying violent interactions related to subtle changes in body movements. In contrast, RNNs such as Long Short-Term Memory (LSTM) are specifically designed to model time-series data and have been successfully used in various video analysis tasks.

In recent years, human skeleton models have been applied in various fields such as human behavior recognition [4-5], gesture recognition of students in the classroom [6-7], sign language recognition [8-9], and have achieved impressive results. Many algorithms for estimating human poses have been developed, such as Alpha Pose [10], OpenPose [11], MediaPipe [12]. Detecting violence using the human skeleton is a relatively new research field that uses computer vision and machine learning techniques to analyze video scenes of human actions and movements to detect violent behavior. The basic idea behind this method is to use the bone structure of a person to identify motion patterns that represent violent behavior. The process involves using a camera to record video scenes of human actions, and then using an algorithm to analyze the motion of the human skeleton in each frame of the video. A challenge in the mission is estimating the situation of many people. To solve this challenge, two approaches have been proposed: the top-down approach and the down-top approach. The top-down approach involves first detecting human bodies in an image or video, and then estimating the pose of each individual body. This approach typically involves using object detection techniques to detect the human bodies, and then using a separate pose estimation model to estimate the pose of each body. In contrast, the bottom-up or down-top approach involves first detecting individual body parts, such as joints or keypoints, and then grouping them together to form complete poses. This approach is typically performed using a single model that simultaneously detects and localizes all body parts in an image or video, and then groups them together to form complete poses. Both approaches have their advantages and disadvantages. The top-down approach is generally more accurate and robust, but can be computationally expensive and may struggle with crowded scenes or overlapping body parts. The bottom-up approach is typically faster and more efficient, but may struggle with occlusions or missing body parts.

In this study, we introduce a top-down approach for real-time violence situation detection using human skeleton extracted from input videos as input for an LSTM model. To achieve this, we start by using a pre-trained YoloV8 model to detect each person present in the input video, followed by MediaPipe to extract the skeleton of each person. We then train this LSTM model on this data to learn how to represent distinguishing features for each skeleton. Our model was evaluated on two widely used benchmark datasets, Hockey Fight and RWF-2000, which include violent and non-violent interactions in videos. Notably, our proposed method demonstrates superior performance on these standard datasets. Our approach can be widely applied to other human pose analysis tasks, including more complex situations involving multiple individuals and interactions.

In the next section, the article will present the following sections: Section 2 will provide an overview of relevant works on violence detection, Section 3 will present the proposed methodology in detail, Section 4 will discuss the results and implications, and finally, Section 5 will conclude our work and discuss feasible future research.

2. Related Works

Detecting violent interactions in videos has been a topic of interest in recent years. Various approaches have been proposed for detecting violent behavior in videos, from manual methods to machine learning methods, and more recently, deep learning-based methods.

Many deep learning techniques have been applied in the recognition of violent behavior in videos and have shown promising results. One method for detecting multi-person violence based on deep 3D convolutional neural networks (3D CNNs) has been proposed by Li et al. [13], which extracts spatiotemporal features of multi-person violence. This method directly detects violence in the input video from start to finish. Experimental results show that this method has higher accuracy compared to artificial feature extraction methods in violence detection. Soliman et al. [14] proposed an end-to-end deep neural network model for the purpose of recognizing violence in videos. The proposed model uses the pre-trained VGG-16 on ImageNet as the spatial feature extractor, followed by Long Short-Term Memory (LSTM) as the temporal feature extractor, and a sequence of fully connected layers for classification. The achieved accuracy is nearly state-of-the-art. Traoré et al. [15] also proposed a deep learning architecture for violence detection, combining both Recurrent Neural Network (RNN) and 2 dimensional convolutional neural networks (2D CNN). In addition to video frames, they also use optical flow computed from captured sequences. The CNN extracts spatial features in each frame, while the RNN extracts temporal features. In addition to video frames, they also use optical flow to encode motion in the scene. The proposed method achieves comparable performance to modern techniques.

A popular new method for recognizing human behavior in videos is based on the skeletal features of humans in action. A skeleton-based method for detecting hostile behavior was proposed by Narynov et al. [16]. This method does not require much hardware, but performs very quickly. The approach of this method involves two stages: extracting features from video frames to estimate the posture of a person, and then classifying actions using a neural network to determine whether the frames include bullying situations or not. Liu et al. [17] designed a frame-based surveillance system for violence detection and abuser tracking with relatively low overall complexity. To detect violence, they proposed a new Skeletal Context Attention Network (SCAN) that is lightweight yet effective in exploiting the spatial and temporal representations of skeletal data. To track abusers, they introduced a Skeleton-Guided Correlation Filter (SGCF) that can continuously track the perpetrator even in extreme cases such as drastic changes in speed or color. Experiments on two benchmark datasets showed that the proposed system not only outperforms state-of-the-art methods for violence action recognition and abuser tracking, but also performs both tasks with less computation.

However, relying on skeletal frames also has some weaknesses when used for violence detection, as it can be influenced by external factors such as video resolution, low lighting. These factors can lead to data loss and reduced accuracy in violence detection. Using multiple frames from the video will improve the accuracy of violence detection. By taking multiple frames from the video, images are synthesized to create a comprehensive picture of violent behavior. LSTM can learn information from previous frames and use this information to help improve the accuracy of violence detection. Therefore, we propose to use the LSTM method combined with human skeletal features for the problem of violence detection.

3. Datasets

To evaluate the performance of the model, we conducted experiments on two benchmark datasets, RWF 2000 [21] and Hockey Fight [22].

The Hockey Fight dataset is a pioneering effort in the field of action recognition, specifically designed to detect instances of violent behavior. The dataset comprises 1000 video clips, each with a resolution of 360x288. The videos are divided into two categories: fighting and non-fighting, with each category containing 500 clips. The dataset is sourced from real events that took place during matches of the National Hockey League (NHL).

The RWF-2000 dataset is provided to the research community as an open resource with the aim of providing access to high-quality data describing real-life instances of violent behavior. The motivation behind creating this dataset is to address the current challenge of collecting reliable video data from environments in the real world, especially from surveillance footage. The dataset contains 2000 videos consisting of 1000 violent and 1000 non-violent interactions. Each video lasts an average of 5 seconds and is extracted from surveillance videos. These video clips include various violent activities such as attacks, fights, and robberies, recorded in many different public spaces such as streets, parking lots, and stores.

4. Method

Detecting and recognizing violent behaviors using features from human skeleton models is a new and promising field. Methods based on skeleton frameworks are less affected by external factors such as clothing (dresses, shirts, shoes) and scene locations (indoor or outdoor) because they rely on spatial relationships between joints rather than external appearances. The use of skeleton data also reduces the amount of data needed to process compared to traditional video analysis, which typically requires significant computational efforts. In addition, the simplicity and effectiveness of skeleton-based methods make them highly suitable for real-time applications, such as surveillance camera systems. Many algorithms for estimating human poses have been developed, such as Alpha Pose, OpenPose, and MediaPipe… In recent years, estimating human pose has been a challenging topic. To address this challenge, two approaches have been proposed, namely the top-down approach and the bottom-up approach.

In this study, we use the top-down approach by combining YoloV8 model as a tool for detecting individuals in the scene and MediaPipe for extracting the skeleton data of those individuals. YoloV8 is the latest version of Yolo, one of the most popular Object Detection models. Since its release, YoloV8 has shown significant improvements (especially in terms of speed) compared to the best models currently available. MediaPipe is a pose estimation model optimized for real-time detection, with fast speed and the ability to detect people at distant positions as well as in low-resolution data conditions. The top-down approach demonstrates computational efficiency, as it focuses on analyzing specific behavior patterns rather than processing a large amount of raw data. For the Hockey fight dataset, due to the fast-paced movements, we used all the frames in each video to ensure both the quality and quantity of the data. As for the RWF-2000 dataset, which has a stable speed and a large dataset size, we extracted one frame every 5 frames to generate data for the next step. Each keypoint consists of coordinate data (x, y, z). For each person, we extract only 12 keypoints at what we consider to be the most important positions (see Figure 1). This process generates a time series of skeleton data for each person appearing in the video and serves as input data for the LSTM model. The architecture of the model is illustrated in Figure 2.

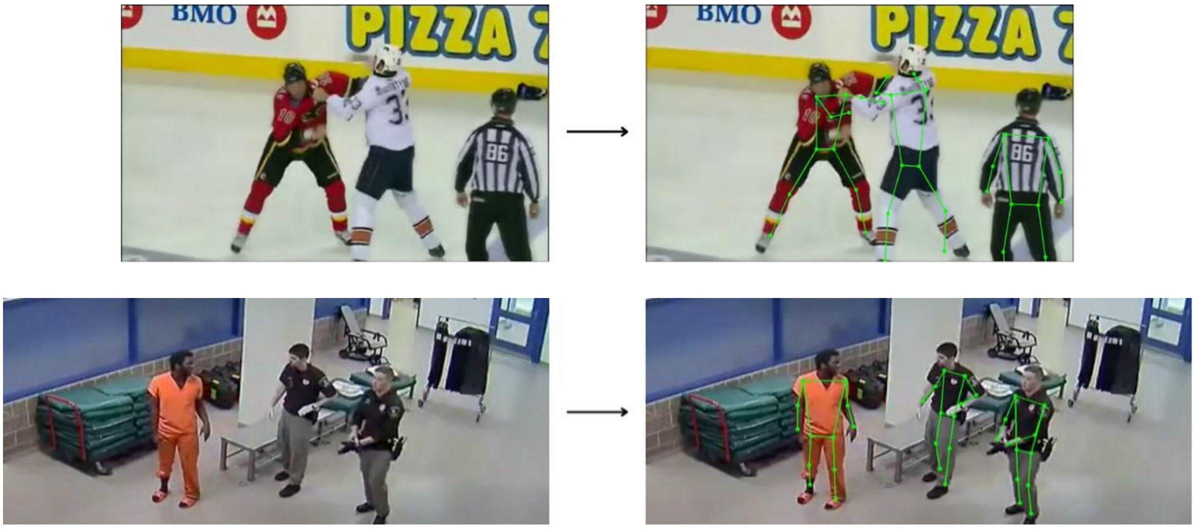

Figure 1. Illustration of the extracted skeletal model.

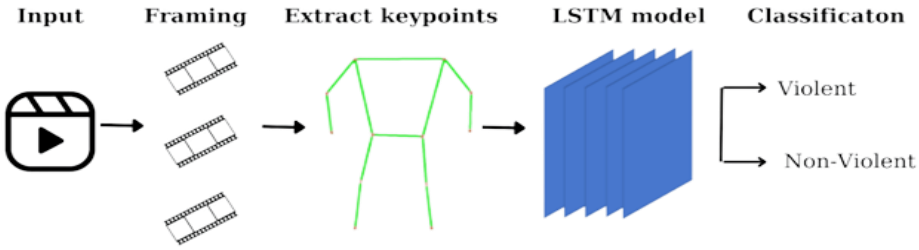

Figure 2. Architecture of a skeletal-based violence detection model using LSTM.

5. Results and Analysis

Our model is evaluated based on two benchmark datasets, Hockey Fight and RWF-2000. The results in Table 1 show that the model achieves promising prediction performance on both datasets.

For the Hockey dataset, the accuracy achieved was 99.53%, precision was 99.89%, and recall was 99.35%, and 99,62% F1-score which was higher than the corresponding prediction results on the RWF-2000 dataset, which achieved 99.58% accuracy, 98.02% precision, 99.16% recall, and 98.59% F1-score. Fig. 3, 4 demonstrates the results of the proposed model testing. The results indicate that the performance of the model on the Hockey fight dataset is better than on the RWF-2000 dataset.

Table 1. The performance of the model on two standard datasets, Hockey Fight and RWF-2000.

Dataset	Accuracy	Precision	Recall	F1-score
Hockey fight	99,53%	99,89%	99,35%	99,62%
RWF-2000	98,58%	98,02%	99,16%	98,59%

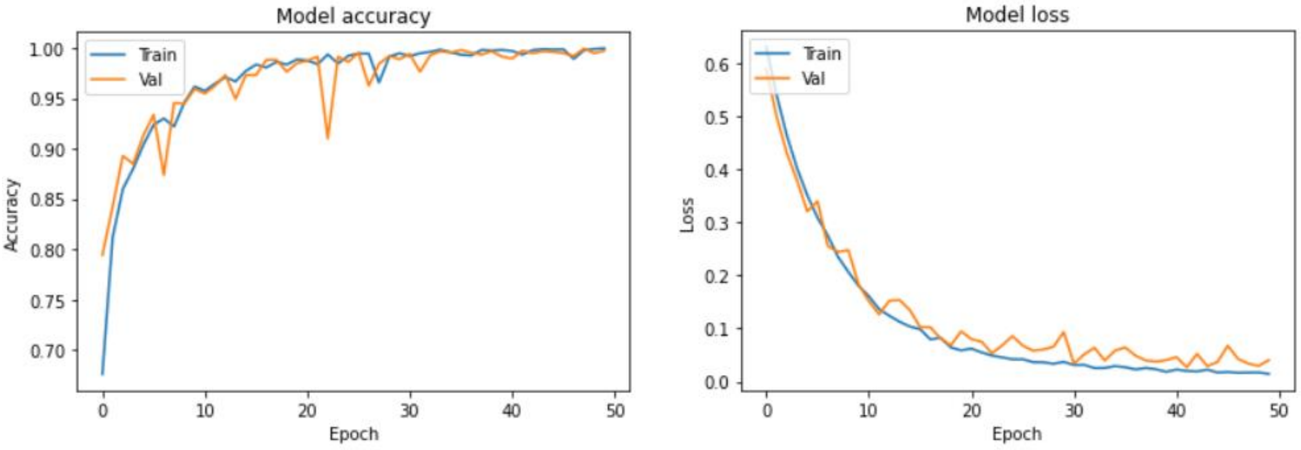

Figure 3. Our method's accuracy and loss curves in the Hockey Fight dataset.

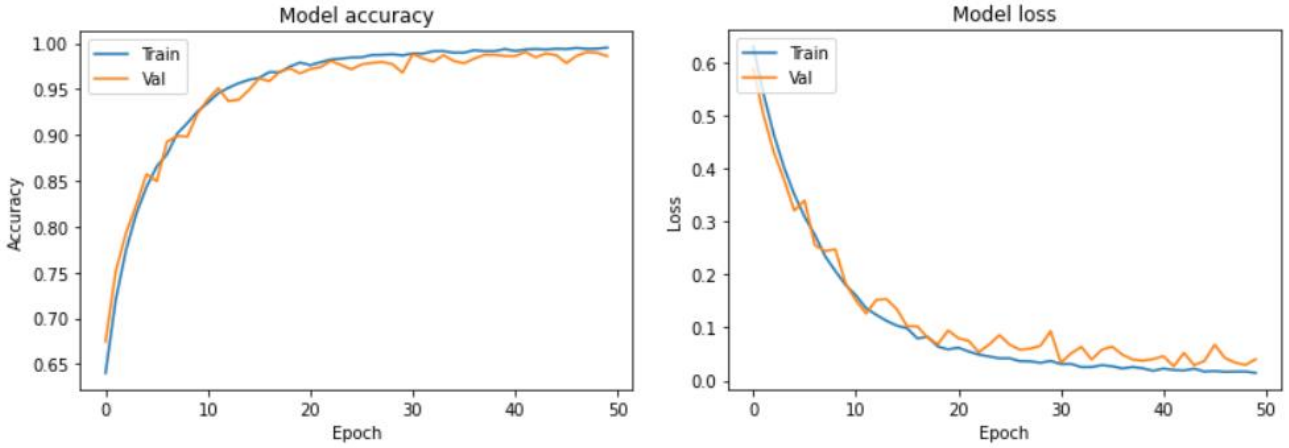

Figure 4. Our method's accuracy and loss curves in the RWF-2000 dataset.

To make a comprehensive comparison, we compared our results with previous works that achieved good performance on these two datasets in Table 2, and Table 3. The comparison results show that our model achieves impressive performance on these two datasets.

Table 2. Accuracy of a several compared to other methods on Hockey fight dataset.

Method	Accuracy
Mumtaz et al [18]	99,28%
Traoré et al [15]	99,00%
Abdali et al [19]	98,00%
Islam et al [20]	99.50%
Ours	99.53%

Table 3. Accuracy of a several compared to other methods on RWF-2000 dataset.

Method	Accuracy
Cheng el al [21]	87.25%
Su el al [23]	89.30%
Islam et al [20]	89.75%
Ours	98.58%

6. Conclusion

Detecting and identifying violent behaviors is an important task in the fields of security and computer vision. In this article, we present a method for detecting and identifying violent behaviors by exploiting information from the human skeleton extracted from a surveillance camera system. To achieve this goal, we built a deep learning model, specifically using the pre-trained YoloV8 model to detect each person in the input video, followed by MediaPipe to extract the skeleton of each person. The results obtained from this process will be input data for the LSTM model used to learn violent interactions from the time series of human skeleton data. Our proposed method focuses on analyzing the specific time-motion of the human skeleton and is capable of detecting subtle motion features that may be signs of violent behavior. Experimental results show that our method is effective and accurate.

References

[1] Ramzan, M., Abid, A., Khan, H.U., Awan, S.M., Ismail, A., Ahmed, M., Ilyas, M. and Mahmood, A.((2019). “A Review on State-of-the-Art Violence Detection Techniques”. IEEE Access, 7, 107560-107575.

[2] Biswas, M., Jibon, A.H., Kabir, M., Mohima, K., Sinthy, R., Islam, M.S. and Siddique, M. ((2022). “State-of-the-Art Violence Detection Techniques: A review. Asian Journal of Research in Computer Science”, 29-42.

[3] Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A. and Khassanova, M. (2022). “State-of-the-art violence detection techniques in video surveillance security systems: a systematic review”. PeerJ Computer Science, 8.

[4] Setiawan, F., Yahya, B.N., Chun, S.-J. and Lee, S.-L. (2022). “Sequential inter-hop graph convolution neural network (SIhGCN) for skeleton-based human action recognition”. Expert Systems with Applications, 195, 116566.

[5] Peng, W., Hong, X., Chen, H. and Zhao, G. (2020). “Learning Graph Convolutional Network for Skeleton-Based Human Action Recognition by Neural Searching”. Proceedings of the AAAI Conference on Artificial Intelligence, 34(3), 2669-2676.

[6] Yan, Q., Hu, Y., Huang, G. and Chen, Z. (2021). “A domain adaptive and continual learning method for skeleton behavior recognition in classroom environment”. 2021 IEEE International Conference on Engineering, Technology & Education (TALE), 138-144.

[7] Lin, F.-C., Ngo, H.-H., Dow, C.-R., Lam, K.-H. and Le, H.L. (2021). “Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection”. Sensors, 21(16), 5314.

[8] Shi, B., Brentari, D., Shakhnarovich, G. and Livescu, K. (2021). “Fingerspelling Detection in American Sign Language”. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4166-4175

[9] Duy Khuat, B., Thai Phung, D., Thi Thu Pham, H., Ngoc Bui, A. and Tung Ngo, S. (2021). “Vietnamese sign language detection using Mediapipe”. 2021 10th International Conference on Software and Computer Applications, 162-165.

[10] Fang, H.-S., Xie, S., Tai, Y.-W. and Lu, C. (2017). “RMPE: Regional Multi-person Pose Estimation”. 2017 IEEE International Conference on Computer Vision (ICCV), 2334-2343.

[11] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y. (2021). “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 172-186.

[12] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). “Mediapipe: A framework for building perception pipelines”. arXiv preprint arXiv:1906.08172.

[13] Li, C., Zhu, L., Zhu, D., Chen, J., Pan, Z., Li, X. and Wang, B. (2018). "End-to-end Multiplayer Violence Detection based on Deep 3D CNN". Proceedings of the 2018 VII International Conference on Network, Communication and Computing, 227-230.

[14] Soliman, M.M., Kamal, M.H., El-Massih Nashed, M.A., Mostafa, Y.M., Chawky, B.S. and Khattab, D. (2019). "Violence Recognition from Videos using Deep Learning Techniques". 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 80-85.

[15] Traore, A. and Akhloufi, M.A. (2020). "Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks". 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 154-159.

[16] Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M. and Omarov, B. (2021). "Physical Violence Detection in Video Streaming Using Partitioned Skeleton Analysis". 2021 21st International Conference on Control, Automation and Systems (ICCAS), 225-230.

[17] Liu, J., Wang, G., Duan, L.-Y., Abdiyeva, K. and Kot, A.C. (2018). Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. IEEE Transactions on Image Processing, 27(4), 1586-1599.

[18] A. Mumtaz, A. B. Sargano and Z. Habib. (2018). "Violence Detection in Surveillance Videos with Deep Network Using Transfer Learning" 2018 2nd European Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, 558-563, doi: 10.1109/EECS.2018.00109.

[19] A. -M. R. Abdali and R. F. Al-Tuma. (2019). "Robust Real-Time Violence Detection in Video Using CNN And LSTM". 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, 104-108, doi: 10.1109/SCCS.2019.8852616.

[20] Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M.H. and Farazi, M. (2021). "Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM". 2021 International Joint Conference on Neural Networks (IJCNN), 1-8.

[21] Cheng, M., Cai, K. and Li, M., (2021). "RWF-2000: an open large scale video database for violence detection". In 2020 25th International Conference on Pattern Recognition (ICPR), 4183-4190.

[22] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar. (2011). ''Hockey fight detection dataset,'' in Computer Analysis of Images and Patterns". Seville, Spain: Springer, 332–339.

[23] Su, Y., Lin, G., Zhu, J. and Wu, Q. (2020). "Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition". Computer Vision – ECCV, 74-90.