Article

Not peer-reviewed version

Efficient Human Violence Recognition for Surveillance in Real-Time

Herwin Alayn Huillcen Baca *, Flor de Luz Palomino Valdivia , Juan Carlos Gutierrez Caceres

Posted Date: 2 November 2023

doi: 10.20944/preprints202311.0110.v1

Keywords: Human violence recognition; video surveillance; real-time; spatial attention; spatial motion extractor; short temporal extractor; global temporal extractor; VioPeru

Article

Efficient Human Violence Recognition for Surveillance in Real-Time

Herwin Alayn Huillcen Baca [1,*], Flor de Luz Palomino Valdivia [1] and Juan Carlos Gutierrez Caceres [2]

[1] Jose Maria Arguedas National University, Andahuaylas 03701, Peru; fpalomino@unajma.edu.pe
[2] San Agustin of Arequipa National University, Arequipa 04001, Peru; jgutierrezca@unsa.edu.pe
* Correspondence: hhuillcen@unajma.edu.pe; Tel.: +51978483090

Abstract: Human violence recognition is an area of great interest in the scientific community, given its broad spectrum of applications, especially in video surveillance systems, since detecting violence in real-time could prevent criminal acts and save lives. Despite the number of existing proposals and research, most focus on the precision of results, leaving aside efficiency and its practical implementation. Thus, this work proposes a model that is effective and efficient in recognizing human violence in real-time. The proposed model consists of three modules: a first module called Spatial Motion Extractor (SME), in charge of extracting regions of interest from a frame; a second module called Short Temporal Extractor (STC), whose function is to extract temporal characteristics of rapid movements, finally the Global Temporal Extractor (GET) module, responsible for identifying long-lasting temporal features and fine-tuning the model. The proposal was evaluated regarding efficiency, effectiveness, and ability to operate in real-time. The results obtained on Hockey, Movies, and RWF-2000 datasets demonstrated that this approach is highly efficient compared to other alternatives. A VioPeru dataset was created to validate real-time applicability with violent and non-violent videos captured by real video surveillance cameras in Peru. The effectiveness results in this dataset outperformed the best existing proposal. Therefore, our proposal has contributions in efficiency, effectiveness, and real-time.

Keywords: human violence recognition; video surveillance; real-time; spatial attention; spatial motion extractor; short temporal extractor; global temporal extractor; VioPeru

1. Introduction

In recent years, with the development of real-time video platforms and video capture cameras, visual data availability has increased rapidly. Due to this constant growth, computing needs to make sense of it to present users with information in an organized manner and, above all, provide valuable services for users. To achieve this, video processing and analysis attempt to identify patterns in the data to enhance aspects widely used in modern society [1].

Video processing and analysis has been a topic of interest in machine learning and pattern recognition for years, focusing on many different problems and tasks, such as action recognition [2], action localization [3], anomaly detection [4], scene recognition [5], among others.

One of the main difficulties when processing videos is their high spatiotemporal nature. Each frame, in principle, can be viewed as a static image containing visual (spatial) information. This simple fact makes the video processing task computationally expensive, even when processing short video clips since it can include many images; furthermore, because there is a dynamic between the spatial content of consecutive frames, a temporal dimension is created.

How to describe spatial and temporal information to understand the content of a video continues to be a research question and, at the same time, a challenge when trying to provide proposals efficient enough to be applied to real problems.

On the specific topic of the recognition of violent human actions, it is observed that it is an area of great interest for the scientific community due to its various applications, such as robotics, medicine,

 doi:10.20944/preprints202311.0110.v1

psychology, human-computer interaction, and mainly video surveillance. An automatic system for detecting human actions and violence could alert about an occurrence or crime and allow measures to be taken to mitigate said occurrence. Therefore, it is essential to detect violent activity in real-time. Although violence recognition in video surveillance has achieved many improvements, most works aim to improve accuracy on known datasets, but few aim at a real scenario.

There are many techniques to detect violence in videos. Typical methods include optical flow [6–10]. Combining optical flow with other methods such as RGB frames as input, Two Stream, CNN variants, and 3D CNN variants achieves good results [11–14]. Therefore, optical flow is a motion representation for video action recognition tasks. However, extracting optical flow requires computational cost and is inefficient for real-time violent human action recognition tasks.

The most promising techniques are based on deep learning [13,15–19], which, unlike optical flow, uses neural networks as feature extractors, encoding, and classification. These techniques achieve better performance, reducing the computational cost of optical flow. However, it is still heavy regarding parameters and FLOPS, so applying them in a real scenario remains a challenge.

We focus on recognizing violent human actions in video surveillance, which can be applied in a real scenario. Classification models must identify human violence at the precise moment of its occurrence, that is, in real-time.

To the best of our knowledge, there are no datasets aimed at the video surveillance domain; the current reference datasets contain a mixture of videos taken from mobile devices, movies, and hockey, where the camera adopts characteristics and positions oriented to the best shot, in a real scenario, this does not happen. The violent human actions captured by video surveillance cameras are more complex since some factors degrade the ability to recognize them, such as occlusion between the people in the scene, the time of day, excessive artificial light from poles and vehicles, type of cameras, camera resolution, and the proportion of the size of the violent scene concerning the size of the frames. In fact, in a recent review, these aspects have been proposed as challenges and problems that still need to be resolved by the current proposals [20] because they are mainly oriented towards effectiveness in sets of known data.

Thus, a model based on deep learning is proposed for the recognition of violent human actions in real-time video surveillance, which can be used in a real scenario, is efficient and at the same time effective.

1.1. Problem

According to the recent study by Ullah et al. [20], it has been observed that in the field of recognition of human actions, there are still problems and challenges that have not been addressed by current proposals. These challenges include occlusion, differentiation between indoor and outdoor cameras, lighting variation in different scenarios, scenes involving crowds, real-time processing, and the complexity and efficiency of existing approaches.

In the specific context of violence detection in video surveillance, additional problems can be identified, such as the proportion of the violent action in relation to the size of the video frame, the direction of focus of the violent action, and the differentiation between day and night. To our knowledge, no proposal has comprehensively addressed these issues; instead, the primary focus has been on improving the effectiveness of the models on specific datasets. It is important to note that current datasets are generic and heterogeneous, which underlines the need to consider the abovementioned issues and even develop datasets more oriented towards specific domains.

1.2. Motivation

The main motivation is that although there are different proposals for the recognition of violent human actions in video, there is no defined proposal aimed at solving the identified problems so that it can be used in a real scenario that is efficient but at the same time effective.

 doi:10.20944/preprints202311.0110.v1

Most have focused on effectiveness, and few on efficiency. Thus, there are very accurate models but complex ones with high computational costs, which could not be used in real-time.

The inefficiency of the proposals in real applications, the limited applicability to a specific domain, and its underlying challenges constitute the motivation of this work.

1.3. Objetives

The general objective of this work is to propose a model based on deep learning for recognizing violent human actions in real-time video surveillance.

Among the specific objectives:

- Generate a dataset from real surveillance cameras, which integrates characteristics of real human violence.
- Develop an effective model in terms of accuracy based on attention and temporal fusion mechanisms.
- Develop an efficient model in terms of the number of parameters and FLOPS based on temporal changes and 2D CNN.
- Develop a compact model for recognizing violent human actions in video surveillance, with minimal latency times close to real-time.

1.4. Contributions

- A model for recognizing violent human actions, which enables its use in a real scenario and in real-time.
- An efficient model in recognizing violent human actions, in terms of the number of parameters and FLOPS. In turn, the model is effective in recognition, in terms of accuracy, whose results contribute to the state-of-the-art.
- Likewise, our proposal contributes to the scientific community by generating and publishing a dataset oriented to the domain of video surveillance.

1.5. Work organization

Section Related Wowrk describes the techniques and results of the best proposals related to the objectives of this work. Section Proposal first describes the work related to the proposal and subsequently details and explains the operation of the proposed architecture and its respective modules. Section Results presents the results according to the objectives. Finally, section Conclussions show the conclusions of this work.

2. Related work

Human action recognition is an area of active research lately, and there are many approaches; in this section, we present some of them, starting with more straightforward approaches and ending with the most novel contributions. First, the work related to the proposal modules is explained, and then the 2D CNN-based models related to the backbone of the proposal are addressed. Finally, a review is made of the cutting-edge techniques for recognizing violent human actions in video surveillance, considering Movies [21], Hockey [21], and RWF-2000 dataset [22].

2.1. Related work to the proposal

2.1.1. Extraction of regions of interest

Video surveillance cameras are mostly fixed-position cameras; violent scenes do not occupy the entire size (HxW) of video. On the contrary, it usually occupies a small portion concerning the size of the video; in this way, the remaining area becomes the background of the violent scene, whose spatio-temporal characteristics do not contribute to correct efficiency in recognition, but rather the

 doi:10.20944/preprints202311.0110.v1

degrades and confuses. On the other hand, extracting these redundant features makes the models less efficient. Therefore, extracting the regions of interest from each frame is essential for effectiveness and efficiency.

Extracting regions of interest have been addressed using attention mechanisms; Ulutan et al. [23] propose to extract regions of moving actors; for this, it uses object detectors with an I3D architecture as a backbone, it also uses amplification and attenuation of the actors. Amplification and attenuation are essential in the extraction of regions of interest. Our proposal carries out the same process without detectors but with morphological deformation processes in each frame, which benefits the model's efficiency by no longer using detectors based on pre-trained backbones.

Zhang et al. [24], concerned about the inefficiency of optical flow in identifying movement limits, present their proposal based on the Euclidean distance of two consecutive frames to later use a convolution backbone. Their proposal is efficient and significantly reduces the number of FLOPs in the process; However, the use of the backbone is still heavy in terms of efficiency. Our proposal takes this proposal as a reference, and we use the Euclidean distance of two consecutive frames, but we do not consider the convolution backbone. Instead, we use morphological deformations to represent the regions of interest.

2.1.2. Short-duration spatiotemporal feature extraction

Proposals based on 3D CNN networks can simultaneously extract spatiotemporal features, such as those proposed by Tran et al. [13], and Carreira et al. [25]; However, it has a high computational cost in terms of efficiency. For this problem, several proposals arise to replace 3D CNN networks with 2D CNN networks without compromising effectiveness, but the replacement improves efficiency.

Lee et al. [26] propose to extract spatiotemporal features from motion filters in a 2D CNN network, and Xie et al. [27] propose to mix modules based on 3D CNN and 2D CNN networks. Both studies achieve adequate effectiveness results in human action recognition datasets rather than violence-oriented datasets. These proposals are considered in our work since using a 2D CNN network generates better efficiency conditions, especially if the objective is to get detection in real-time.

In this way, Lin et al. [28] propose to use 2D CNN networks but with a substantial improvement, the temporal change module, in which consecutive frames replace the dimension of the channels of the frames and generate extraction of spatiotemporal features with several convolutions. Our proposal considers replacing channel information with temporal information from consecutive frames, but we only use a single 2D CNN network (backbone). Finally, it is possible to extract spatiotemporal characteristics of short duration, in this case from three consecutive frames, significantly reducing the number of FLOPs of the Lin et al. proposal [28]

2.1.3. Global spatiotemporal feature extraction

Extracting spatiotemporal features from three consecutive frames can finally recognize human actions, as proposed by Huillcen et al. [29], which has better efficiency results but still fails to surpass the state-of-the-art proposals in terms of effectiveness. Our proposal adds a module to extract temporal characteristics from a larger set of frames, for example, thirty, in such a way as to look for characteristics of movements that cover more significant numbers of frames, and not just three, without losing the objective of efficiency.

We return to the proposal of Zhang et al. [24], referring to its feature reduction module in the time dimension through max-pooling layers, to merge it with features that can enable the recognition of human actions. The proposal takes this module, but we use only two average pooling layers to make the model compact, merge it with the input, and recognize violent actions through fully connected layers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

2.2. Reference backbones

According to the previous section, our work considers using a pre-trained 2D CNN backbone. A comparison of the different alternatives is made to choose the backbone of our proposal.

In the literature review, different proposals and techniques stand out. For the choice of alternatives, efficient models are considered without neglecting their effectiveness. Among them are ResNet50 [30], InceptionV3 [31], DenseNet121 [32], SqueezeNet [33], MobileNet V2 [34], MobileNet V3 [35], EfficientNet B0 [36], MnasNet [37], GhostNet V2 [38], and Vision Transformers [39].

The comparative analysis is done based on the effectiveness results on the ImageNet dataset [40]. This original dataset has 1280000 training images and 50,000 validation images with 1000 classes.

The Table 1 shows the comparative analysis of these models. The possible candidates to be used have been marked in bold.

Table 1. Summary of effectiveness and efficiency results of backbone models, tested on the ImageNet dataset [40]

Model	Accuracy (%)	Number of parameters (M)	FLOPS (G)
Resnet50	76	25,6	3,8
InceptionV3	78,8	23,2	5,0
DenseNet121	74	8,0	2,8
Squeezenet	57,5	1,25	0,83
GhostNetV2	75,3	12,3	0,39
EffcientNetB0	**78**	**5,3**	**1,8**
MobileNetV2	**72,6**	**3,4**	**0,3**
MobileNetV3 L	**76,6**	**7,5**	**0,36**
MasNet	**75,2**	**3,9**	**0,315**
Vision Transformers (VIT - Huge)	88,55	632	-

Vision transformers in sequence-based problems have shown significant performance [41], especially in natural language processing tasks, in image detection and recognition tasks [39]. Similarly, there are applications in recognizing violent actions [42]. However, to our knowledge, they have yet to be evaluated on the RWF-2000 reference dataset [22], nor are there any results on the efficiency associated with these proposals. Therefore, we do not take this model as a candidate. However, the main reason is that, in general, proposals based on transformers aim to improve the effectiveness of their results but at a higher computational cost than proposals based on 2D CNNs.

2.3. Benchmark

2.3.1. Benchmark on classic datasets

The methods for recognizing violent human actions are divided into two groups: handcraft methods and deep learning methods. Handcraft methods do not achieve good results, especially in efficiency, so we have the most representative works, such as Gao et al. [6] with Oriented Violence Flows (OViF), Deniz et al. [7] with Randon Transform, Bilinski et al. [8] with Fisher Vectors, Zhang et al. [9] with Weber local descriptor (MoI-WLD), and Deb et al. [10] outlier-resistant VLAD (OR-VLAD).

Deep learning methods use deep neural networks as feature extractors. Among the most important, we have Dong et al. [15] with multiple streams based on stream model [12], Zhou et al. [16] with time slice networks (TSN) and FightNet, Serrano et al. [17], whit Hough Forests. Estas propuestas aún usan flujo óptico combinado con deep learning; por tanto, aun presentan problemas de eficiencia y dependencia de Handcraft methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

As a solution to these problems, we have other proposals such as Sudhakaran et al. [18] with 2D ConvNets and ConvLSTM, Hanson et al. [19] with ConvLSTM (Bi-ConvLSTM) architecture.

Recently, 3D CNN-based approaches improved the effectiveness of previous proposals but at a high computational cost, typical of 3D CNN models. Thus we have the proposal of Tran et al. [13], and Li et al. [43]

An improvement to the previous approach in terms of efficiency is presented by Huillcen et al. [44], which uses a DenseNet architecture but with different configurations of dense layers and dense blocks to make the model more compact. Later, Huillcen et al. [29] present a new proposal based on extracting spatiotemporal features using a 2D CNN and extracting regions of interest to make the model more compact.

Table 2 summarizes the results of all these proposals in classic datasets: Hockey Fights Dataset [21], Movies Dataset[21] and Violent Flows [45].

Table 2. Summary of the methods for recognizing violent human actions in video surveillance in classic datasets.

Method	Hockey Fight dataset	Movies dataset	Violent Flow dataset
ViF + OViF [6]	87.5 ± 1.7%	-	88 ± 2.45%
Random Transform [7]	90.1 ± 0%	98.9 ± 0.22%	-
STIFV [8]	93.4%	99%	96.4%
MoIWLD [9]	96.8 ± 1.04%	-	93.19 ± 0.12%
OR-VLAD [10]	98.2 ± 0.76%	**100 ± 0%**	93.09 ± 1.14%
Three streams + LSTM [15]	93.9%	-	-
FightNet [16]	97.0%	100%	-
Hough Forests + CNN [7]	94.6±0.6%	99±0.5%	-
ConvLSTM [18]	97.1±0.55%	**100±0%**	94.57±2.34%
Bi-ConvLSTM [19]	98.1±0.58%	**100±0%**	93.87±2.58%
3D CNN end to end [43]	98.3±0.81%	100±0%	97.17±0.95%
3D-DenseNet(2,6,12,8) [44]	97.0%	100%	90%
SA+TA [29]	97.2%	100%	-

2.3.2. Benchmark on RWF-2000 dataset

The methods described in the previous subsection show an analysis on datasets that are not specific for video surveillance, nor is an analysis of efficiency done; their approaches are oriented towards effectiveness but not efficiency.

An RWF-2000 dataset recently emerged, proposed by M. Cheng et al. [22], which consists of 2000 videos extracted from Youtube, are videos of different resolutions, sources, and camera positions, which makes it a reference dataset and at the same time a challenge for the methods tested in classic datasets. Below is an analysis of the recent proposals tested on this dataset, with results in effectiveness and efficiency.

According to a recent study by Mumtaz et al. [46], the main methods for recognizing violent human actions have little information about their results in terms of efficiency, that is, information about the complexity and number of parameters, confirming the fact that the objective is to find high results in terms of accuracy, but increasing the complexity of the model. It is shown that the proposals based on optical flow, 3D CNN, LSTM, Two Stream networks, and 3D skeletons have high computational costs and cannot be used in real-time scenarios.

However, there are proposals whose objectives were to find models with good results of effectiveness and at the same time of efficiency, see Table 3 and Figure 1. Carreira et al. [25], with I3D's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

feature-based mechanism , and Cheng et al. [22], where the authors achieve an accuracy of 87.25% with only 0.27 million parameters, but in the complexity calculation, they do not consider preprocessing with optical flow.

Table 3. Summary of the methods for recognizing violent human actions in video surveillance in RWF-2000 dataset.

Model	Accuracy(%)	#Params (M)	FLOPs(G)
C3D (Tran et al.)[13]	82,75	94,8	40,04
I3D + RGB (Carreira et al.)[25]	85,57	12,3	55,7
I3D + Two Stream (Carreira et al.)[25]	81,75	24,6	-
I3D + Optical Flow (Carreira et al.)[25]	75,5	12,3	-
ConvLSTM (Sudhakaran et al.)[18]	77	94,8	14,4
Flow Gated Network (Cheng et al.)[22]	87,25	0,27	-
SA+TA (Huillcen et al.)[29]	87,75	5,29	4,17
SepConvLSTM (Islam et al.) [47]	**89,75**	**0,33**	1.93

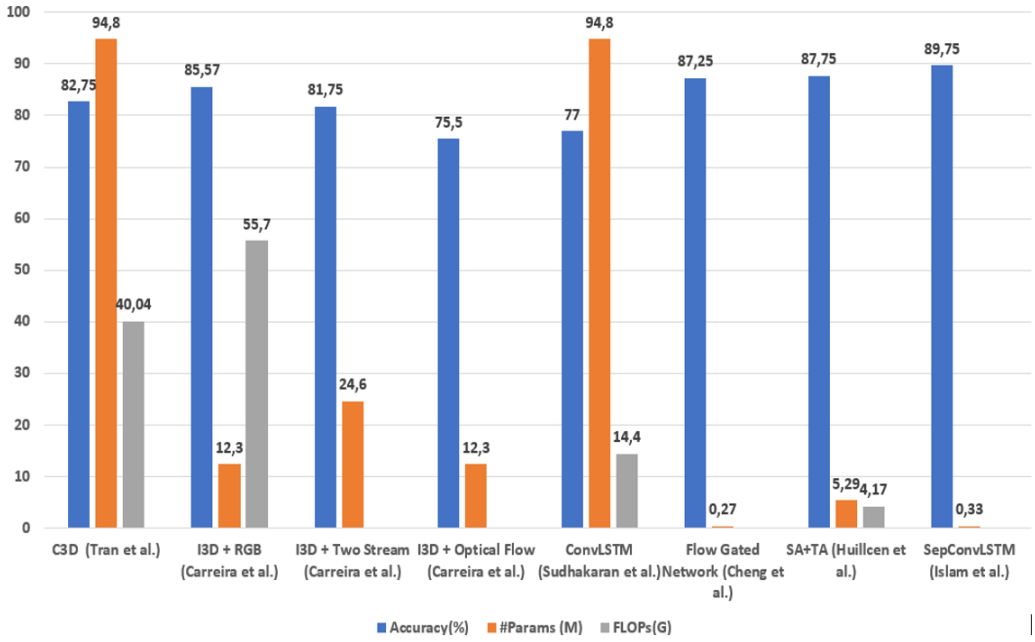

Figure 1. Summary of the methods for recognizing violent human actions in video surveillance in RWF-2000 dataset.

Sudhakaran et al. [18] applied 2D ConvNets to extract spatial feature maps, followed by ConvLSTM to encode the spatiotemporal information, improving efficiency to 77%, but at a high cost (94.8 million parameters). Something similar happened with Tran et al. [13] with his proposal based on 3D CNN.

There are techniques based on 3D skeletons, the most representative of which was presented by Su et al. [48], who achieve an efficiency of 89.3%, but the fact of using the extraction of key points in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

recognition of skeletons has several associated problems, first that it has a high computational cost, it is not suitable for the domain in video surveillance, since in a real scenario there is no camera focus with the complete bodies and with the direction towards the violent scene.

Outstanding proposals also emerged based on a sequence of spatial and temporal feature extraction, which managed to reduce the complexity of 3D CNN networks by taking advantage of the efficiencies of 2D CNNs. Huillcen et al. [29] improves its proposal in terms of efficiency but still needs to surpass the effectiveness of some proposals. For this, it was based on replacing the 3D CNN approach with a 2D CNN and also used preprocessing to identify regions of interest.

To the best of our knowledge, the best proposal in terms of efficiency and effectiveness on the RWF-2000 dataset [22] is presented by Islam et al. [47], who was based on a separable convolutional network (SepConvLSTM) and MobileNet, reached an efficiency of 89.75% and an efficiency of 0.333 million parameters. However, the efficiency results could be more questionable since Mobilenet V2 [34] alone has 3.4 million parameters.

3. Proposal

According to the review of state-of-the-art and having to respond to the challenges of proposing an efficient, effective model whose results are a contribution to state-of-the-art, and in turn can be used in real-time, We propose an architecture of three modules, a first module called Spatial Motion Extractor (SME), in charge of extracting regions of interest from a frame, a second module called Short Temporal Extractor (STC), whose function is to extract temporal characteristics of fast movements and of short duration, finally the Global Temporal Extractor (GET) module, in charge of identifying long-term temporal features and fine-tuning the model for better accuracy.

3.1. Proposal architecture

The general objective of the proposal has an inversely proportional nature: be efficient and, at the same time, effective in recognizing human actions in real-time. When reviewing the state-of-the-art, we have that proposals with high efficiency often have low efficiency, and vice versa. In this way, the proposal contains modules that allow high effectiveness with efficient methods. Thus, a hybrid architecture is proposed, composed of three techniques: Spatial motion extraction, 2D CNN with frame averaging, and temporal feature extractors. Figure 2 shows the proposed architecture:

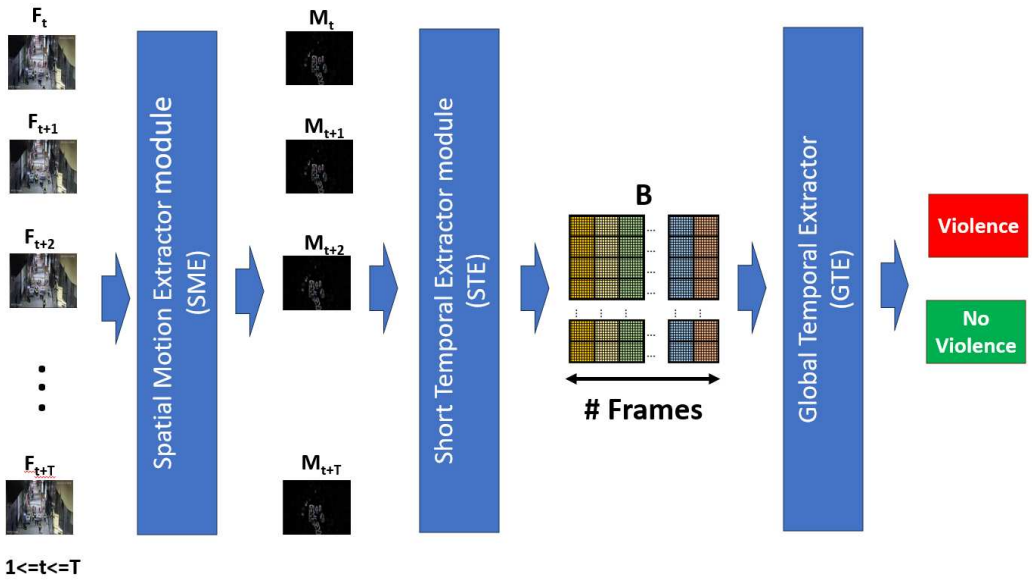

Figure 2. Summary of the Proposed Architecture.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

The input is the sequence of video frames $F_t, F_{t+1}, F_{t+2}, ..., F_{t+T}$, for $1 \leq t \leq T$, and $T = 30$, which are resized to a resolution of 224 x 224 pixels. The details of each module are detailed in the subsequent subsections:

3.2. Spatial Motion Extractor module(SME)

This module is based on the natural process of a human being when observing a scene. When the scene is static, sensory attention covers the entire scene; however, when some movement occurs, sensory attention is oriented to the specific movement area, making visual perception and possible identification of movement more successful. We take this natural process into account to do frame-by-frame preprocessing of a video in order to extract one or several regions of interest involved in the movement since all violent human action lies mainly in movement, especially rapid movements.

Extracting a region of the frame where movement occurs increases recognition efficiency since it will make the model extract characteristics only from the movement, leaving aside areas that do not contribute and degrade the effectiveness. This module is essential in our proposal, and the details are shown in Figure 3.

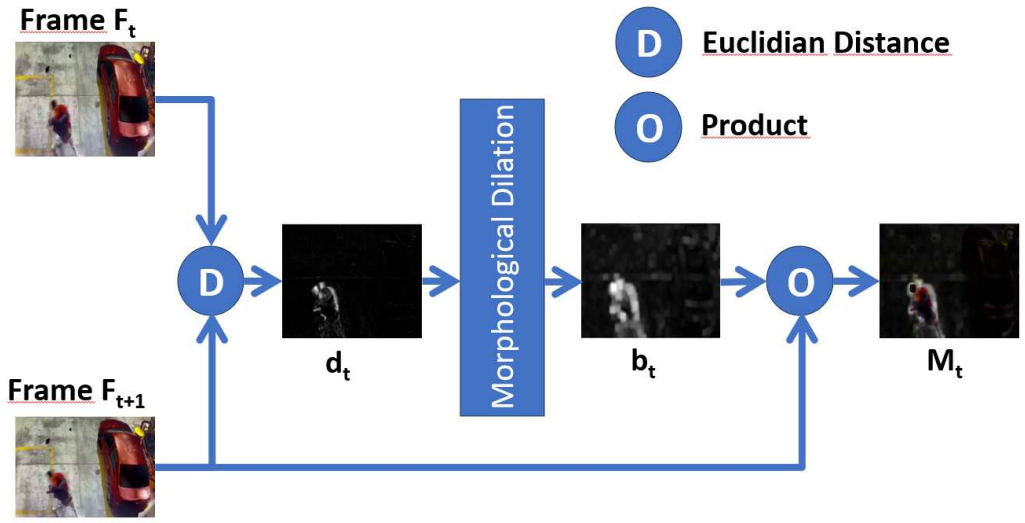

Figure 3. Spatial Motion Extractor module (EME).

The Spatial Motion Extractor (SME) module takes two consecutive RGB frames $F_t, F_{t+1} \in \mathbb{R}^{3\times W\times H}$, for $1 \leq t \leq T$ and calculates the Euclidean distance D, for each pixel and each channel, according to:

$$d_t = \sqrt{\sum_{i=1}^{3}(F_{t+1}^i - F_t^i)^2} \tag{1}$$

Where: T represents the number of frames to be processed, i represents the RGB channels, F a specific frame, and $d_t \in \mathbb{R}^{3\times W\times H}$. When a pixel remains with the same value and the same position, the difference will be zero; therefore, the pixels without movement will be black, causing the resulting frame to extract the background of the initial frame, but if the pixel of the same position changes value, there is some movement, and the difference will be one grayscale pixel. Figure 4 shows an example of the administrative distance of two consecutive frames:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

Figure 4. Euclidean distance of two consecutive frames F_t, F_{t+1}.

It is observed that d_t represents grayscale motion limits of the frames F_t, F_{t+1} and with the respective removal of the background; However d_t does not yet represent a region of interest, morphological deformations are applied so that the limits of movement act as perimeters of the region of interest, in such a way as to convert limits into regions. For this, dilation was used with a 3 x 3 kernel and 12 iterations. Finally, Figure 5 shows the result b_t of the previous example.

Figure 5. Morphological deformations in d_t.

With b_t the region is identified, but not the actual pixels of the movement, so with a dot product procedure between the frame F_{t+1} and b_t we obtain $M_t \in \mathbb{R}^{3\times W\times H}$, see Figure 6, which represents the region of interest where movement occurs, with the elimination of the background.

Figure 6. M_t: Spatial extraction of motion.

3.3. Short Temporal Extractor module (STE)

This module has the function of achieving effectiveness and, at the same time, efficiency in recognizing violent human actions. To do this, we consider that violent actions, such as punching, kicking, throwing, and others, are rapid movements, which could be reflected in the variation of pixels in consecutive frames. Thus, it is proposed to extract spatiotemporal characteristics of short duration, specifically from three consecutive frames.

Extracting spatiotemporal features is the fundamental task of recognizing human actions in general; it is a task that has many techniques and proposals in recent years. As discussed in Section 2, 3D CNN-based architectures are the most appropriate to extract these spatiotemporal features. However, given its high computational complexity in the number of parameters and FLOPS, it is an unviable technique for real-time video surveillance. Although there are many other techniques to alleviate the associated computational cost, to the best of our knowledge, they have yet to propose a model for recognizing human actions in a real scenario.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

According to the above, the Short Temporal Extractor module (STE) takes three consecutive RGB frames from the Spatial Motion Extractor module (SME) M_t, M_{t+1} and M_{t+2}, transforms them into a single frame $P_{t,t+1,t+2}$, to finally extract spatiotemporal features through a 2D CNN network, which constitutes the backbone of this module, it was chosen the MobileNet V2 network [34], Figure 7 shows the details of this module:

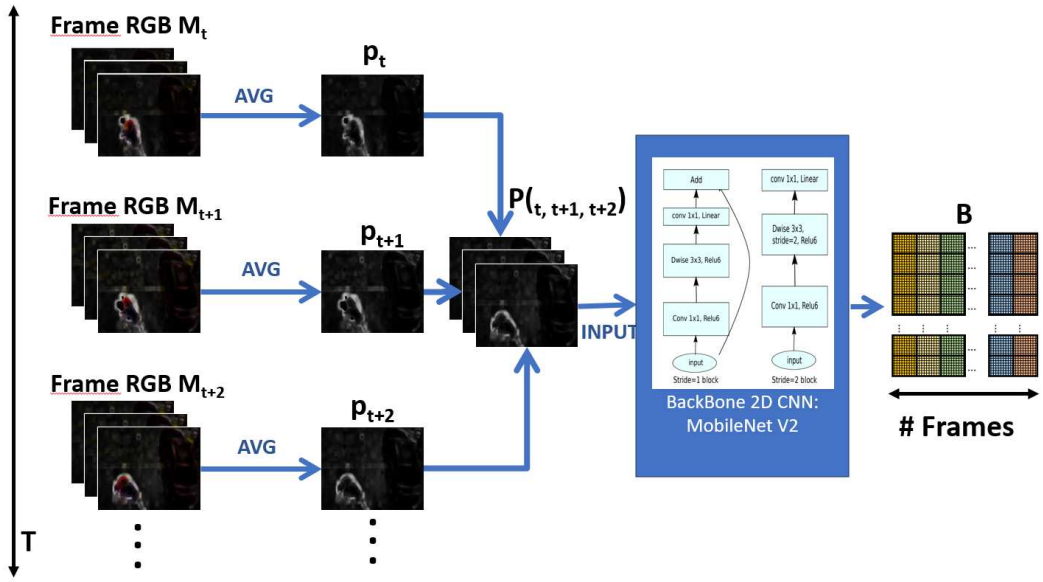

Figure 7. Short Temporary Extractor module (STE).

p_t, represents the average of the three channels c of M_t, according to:

$$p_t = \sum_{c=1}^{3} (M_t^c) \tag{2}$$

In this way, when p_{t+1} and p_{t+2} are obtained, they are considered as channels and are assembled into a simple frame $P_{t,t+1,t+2}$, as if it were an image. In this way, the color information of the boxes M_t, M_{t+1} and M_{t+2} is lost, which does not contribute to the recognition of violent human actions but the extraction of short-term temporal characteristics is possible.

On the other hand, the Short Temporal Extractor module (STE) provides efficiency to the model since, as can be seen, the number of processed frames T is reduced to a third: $\frac{T}{3}$, making the number of FLOPs is substantially reduced during processing in the 2D CNN network.

The frame $P_{t,t+1,t+2}$ containing temporal information enters a 2D CNN network, as if it were an image, to extract spatiotemporal features. The 2D CNN network chosen was MobileNet V2 [34], because it is the most efficient and has the best results. The output is a feature map $B \in \mathbb{R}^{\frac{T}{3} \times C \times W \times H}$, where C is the number of channels and H, W is the size of a feature matrix.

3.4. Global Temporal Extractor (GTE)

Short Temporal Extractor Module (STE) can capture spatiotemporal features B and, without any additional module recognize violent human actions, as proposed by Huillcen et al [29]; However, it still does not take into account temporal characteristics between all frames $\frac{T}{3}$, in this way, it is possible to improve efficiency with some modifications that do not compromise efficiency too much.

The fact is taken into account that the feature map B contains information from the frames $\frac{T}{3}$, and each channel frame C. In this way, by processing the relationship between the channels of each frame, global spatiotemporal characteristics are achieved that improve the effectiveness of the model, based on the proposal of Zhang et al. [24], a reduction process is done, where we reduce the dimensions of B spatially and temporally, and then merge the features and with a final fully connected layer, obtain the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

outputs "violence", and "no violence", Figure 8 shows the Global Temporal Extractor Module (ETG) in a general way.

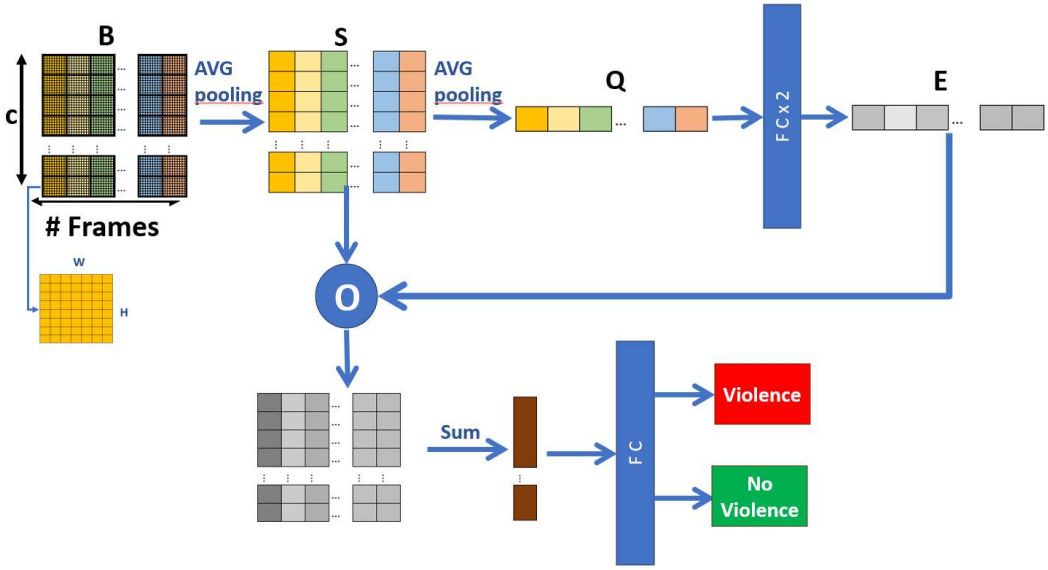

Figure 8. Global Temporal Extractor module (GTE).

Taking the feature map B as input, we do a spatial compression or reduction process to obtain S, of the form:

$$S^c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} B^c(i,j) \tag{3}$$

Where $H \times W$ is the size of a channel, S^c represents the average of the elements of the channel c, corresponding to applying a Global Average Pooling layer. the result S is a set of vectors of the form $S^c = [b_1^c, b_2^c, ..., b_{\frac{T}{3}}^c]$.

Once again, a temporal compression process is carried out on the set of vectors S for each channel to obtain Q, in the form:

$$q = \frac{1}{C} \sum_{c=1}^{C} S^c \tag{4}$$

q is a vector $q = [q_1, q_2, ..., q_{\frac{T}{3}}]$, which connects two fully connected layers followed by a sigmoid function and obtains the temporal characteristics E, which is a vector $[e_1, e_2, ..., e_{\frac{T}{3}}]$.

The next step is to recalibrate the weights (excitation) and obtain characteristics resulting from the channel relationships. To do this, a point-to-point multiplication is performed between E and S. The result is added over time, and A new vector is obtained, representing the final temporal characteristics. Therefore, it is connected to the fully connected layer, and the final output is the recognition of "violence" and "non-violence".

4. Results

We show the results of the proposal according to the objectives of efficiency, effectiveness, and real-time. But first, the dataset and the model configuration are described.

4.1. Datasets

According to the Related Works section, there are several freely distributed datasets; However, only some are taken as a reference to evaluate the performance of the different proposals. For our tests,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

we take the classic datasets such as Hockey Fights [21], Movies [21], and the reference datasets current: RWF-2000 [22].

As a contribution, we present the VioPeru dataset, created with real sources from Peruvian video surveillance cameras.

4.1.1. Hockey Fights dataset

Hockey Fights [21] contains 1000 clips extracted from hockey games. Figure 9 shows an example of a clip classified as Fight (Fi).

Figure 9. Frame sequence of a video from the Hockey Fights dataset.

This dataset only presents sequences of hockey fights, usually two-person fights, almost always wearing the same sports clothing; the violent action occupies almost the entire size of the frames, and the background is almost the same in all videos. These characteristics differ from a violent scene taken by video surveillance cameras; the lighting conditions are the same. Therefore, there are better candidates than this dataset to train a human action recognition model in video surveillance. However, we can take it as a reference to compare effectiveness results with other models.

4.1.2. Movies dataset

Movies [21] has 200 clips extracted from action movies. Figure 10 shows an example of a clip classified as Fight (Fi).

 doi:10.20944/preprints202311.0110.v1

Figure 10. Frame sequence of a video from the Movies dataset.

Movies dataset is a set of videos with characteristics similar to the Hockey dataset as they respond to a prepared and planned scene. However, it is more heterogeneous in showing different scenes, although the majority are boxing, still presents biases and cannot be an adequate dataset to be used in a violence recognition model in real video surveillance cameras.

4.1.3. RWF-2000 dataset

RWF-2000 [22] is the most significant violence detection dataset; it contains 2000 videos extracted from YouTube, lasting 5 seconds. Figure 11 shows clips of videos labeled as violent.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

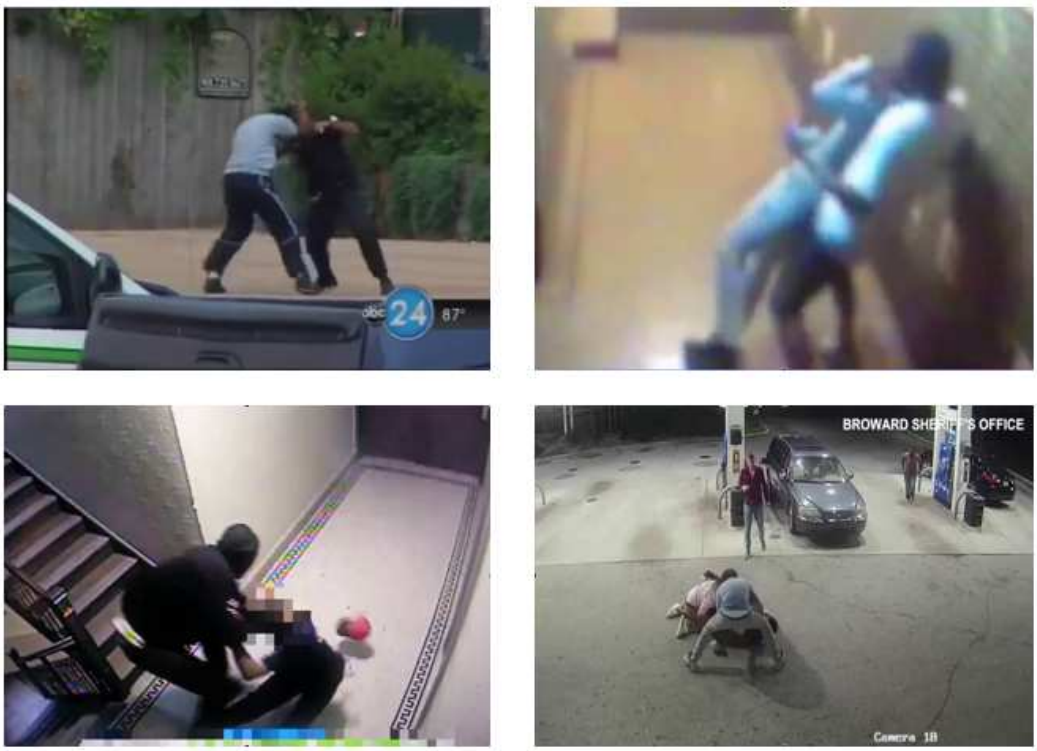

Figure 11. Examples of violent clips from the RWF-2000 dataset.

This dataset has greater diversity and heterogeneity in the videos, as it presents violent scenes from video surveillance cameras. However, the videos have been previously prepared, cut, corrected, and edited by having YouTube as a collection source. It also presents videos taken from smartphones indoor type scenes, with almost no night scenes. These characteristics do not necessarily represent video surveillance scenes; however, it is a reference for comparing results with the state-of-the-art.

4.1.4. VioPeru dataset

The various biases presented by the Hockey Fights [21], Movies [21], and RWF-2000 [22] datasets were extensively discussed. Certainly, when analyzing real videos of violence from video surveillance cameras in our environment, certain characteristics of video surveillance are extracted:

- The violent scene involves two people, several people, or crowds.
- Cameras have different resolutions.
- There is a different proportion of the violent scene concerning the size of the frames; that is, the violent action is large or so small that it can go unnoticed by the human eye.
- Violent human actions occur primarily at night when lighting can negatively influence detection.
- Violence in video surveillance is not only made up of fights; there is also looting, vandalism, violent protests, attacks on property, and confrontations between groups of people.
- Occlusion is typical in video surveillance; that is, the same people, trees, and vehicles, among others, cover violent scenes.

According to the above, as part of this work, we contribute and make available a new dataset called VioPeru, which consists of 280 videos collected from records of real video surveillance cameras currently used as part of the surveillance system. The videos were collected from Talavera, San Jerónimo, and Andahuaylas districts in the Apurímac region, Peru.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

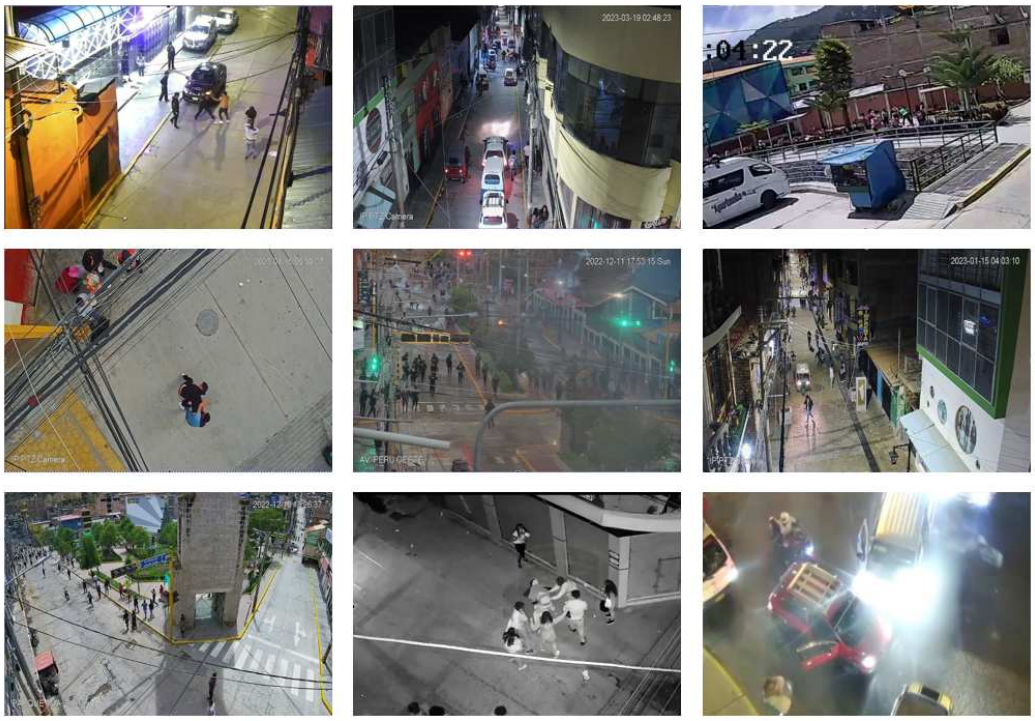

Figure 12. Examples of violent frames from the VioPeru dataset.

VioPeru presents videos that include all the features detailed above; see Figure **??**. It was collected and accessed thanks to agreements between the José María Arguedas National University of Andahuaylas and the municipalities of the respective districts.

VioPeru is the dataset that serves as the basis for generating the model applied to recognize violent human actions for real scenarios in such a way as to become a model oriented to video surveillance.

The dataset is available at: https://github.com/hhuillcen/VioPeru

4.2. Model configuration

The proposal used Python version 3.8 and the PyTorch version 1.7.1 library as a base. The Hardware was a workstation with NVidia GeForce RTX 3080 Ti GPU, 32 GB RAM, and a 32-core Intel Core i9 processor.

The datasets were divided into training and test subsets, with 80% and 20%, respectively.

The following configuration was used in the training phase:

- Learning rate: 10^{-3}, for all datasets.
- Batchsize: 2.
- Number of epochs: 100.
- Optimizer *Adam*, with textitEpsilon: 10^{-9}, *weight decay*: 10^{-2}, and to calculate the loss function: *Cross Entropy*.
- One-Cycle Learning Rate Scheduler, with *min-lr*: 10^{-8}, *patience*: 2, and *factor*: 0.5.

4.3. Results evaluation

4.3.1. Results evaluation on classical datasets

The results were extracted from the *accuracy* metric in the classic datasets. The Table 4 shows the results. A comparison with the most significant proposals is also shown.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

Table 4. Comparison of results obtained on classical dataset.

Method	Hockey Fight dataset	Movies dataset	Violent Flow dataset
ViF + OViF [6]	87.5 ± 1.7%	-	88 ± 2.45%
Random Transform [7]	90.1 ± 0%	98.9 ± 0.22%	-
STIFV [8]	93.4%	99%	96.4%
MoIWLD [9]	96.8 ± 1.04%	-	93.19 ± 0.12%
OR-VLAD [10]	98.2 ± 0.76%	**100 ± 0%**	93.09 ± 1.14%
Three streams + LSTM [15]	93.9%	-	-
FightNet [16]	97.0%	100%	-
Hough Forests + CNN [7]	94.6±0.6%	99±0.5%	-
ConvLSTM [18]	97.1±0.55%	**100±0%**	94.57±2.34%
Bi-ConvLSTM [19]	98.1±0.58%	**100±0%**	93.87±2.58%
3D CNN end to end [43]	98.3±0.81%	100±0%	97.17±0.95%
3D-DenseNet(2,6,12,8) [44]	97.0%	100%	90%
SA+TA [29]	97.2%	100%	-
Proposal	**98.2%**	**100%**	-

The results show the comparison with other state-of-the-art proposals in terms of the effectiveness of the models. Performing the analysis for the Hockey dataset, cutting-edge results and contributions to the state-of-the- art are observed. Our proposal reaches 98.2% and is only surpassed by the 3D CNN End to End model [43] with 98.3%, with a 0.1% difference; However, it must be noted that this proposal is based on a 3D convolutional network CNN, and given its high number of d and FLOPS parameters, it is not feasible to be used in real-time.

For the case of the Movies dataset, the result of our proposal reached the maximum *accuracy*, like other proposals, verifying state-of-the-art results. It is necessary to indicate that this dataset has simple characteristics compared to other datasets, such as RWF-2000.

4.3.2. Results evaluation on RWF-2000 datasets

Unlike classical datasets, RWF-2000 [22] is the reference dataset for proposals aimed at effectiveness and efficiency. In this way, there are proposals with current techniques and better general results; thus, it becomes a good criterion for comparing our results.

As in the previous case, the *accuracy* metric was used to analyze the effectiveness of the models. For efficiency results, the FLOPS and the number of parameters calculated by each model were used, and the results were compared with the most representative proposals of the state-of-the-art.

Table 5 shows the results and their respective comparison in terms of effectiveness and efficiency.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

Table 5. comparison of results obtained in the RWF-2000 dataset, taking efficiency and effectiveness as reference.

Model	Accuracy(%)	#Params (M)	FLOPs(G)
C3D (Tran et al.)[13]	82,75	94,8	40,04
I3D + RGB (Carreira et al.)[25]	85,57	12,3	55,7
I3D + Two Stream (Carreira et al.)[25]	81,75	24,6	-
I3D + Optical Flow (Carreira et al.)[25]	75,5	12,3	-
ConvLSTM (Sudhakaran et al.)[18]	77	94,8	14,4
Flow Gated Network (Cheng et al.)[22]	87,25	0,27	-
SA+TA (Huillcen et al.)[29]	87,75	5,29	4,17
SepConvLSTM (Islam et al.) [47]	**89,75**	**0,33**	1.93
Proposal	**88,5**	**3,51**	**3,15**

It is observed that the proposal has the lowest amount of FLOPS, with a value of 3.15, after SepConvLSTM [47] with a value of 1.93. The other proposals are much higher. This result suggests that recognizing violent human actions has a very short latency and allows its use in devices limited in computational power. That is, the proposal can be used in real video surveillance scenarios. On the other hand, this result contributes to the state-of-the-art regarding the efficiency of models for recognizing violent human actions.

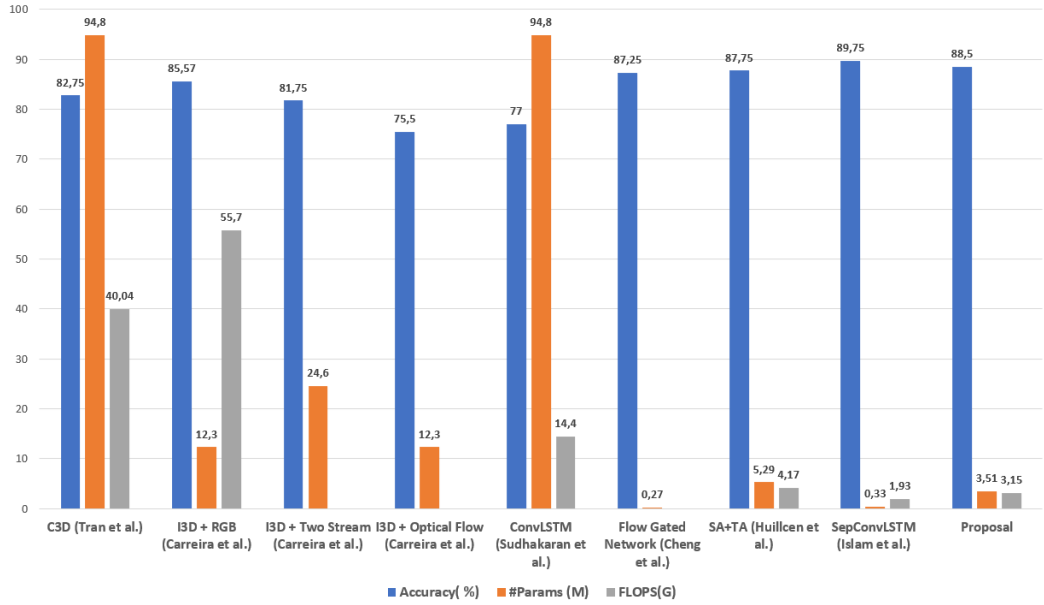

Figure 13. Graphic comparison of results obtained in the RWF-2000 dataset, taking efficiency and effectiveness as reference. .

For effectiveness results, in terms of *accuracy*, it is also observed that our proposal has a value of 88.5% and is only below SepConvLSTM [47] with a value of 89.75%, the Other proposals have close results, but inferior to our proposal. This result demonstrates that our proposal has cutting-edge

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

effectiveness, contributes to the state-of-the-art, and can be used in real situations of violence identification in video surveillance cameras.

Finally, for efficiency results in the number of parameters, our proposal achieves better results than SA+TA [29], with a value of 5.29 million parameters, and both are below SepConvLSTM [47] with a value of 0.33 million parameters. However, when analyzing the complexity of the modules of our proposal, it is noted that the Short Temporal Extractor module (STE) uses the 2D CNN network MobileNetV2 [34], which has 3.4 million parameters, the rest of the modules only occupy 0.11 million parameters, that is, almost all the complexity of our proposal is the result of using the MovileNetV2 network [34]

The same analysis is done for the SepConvLSTM proposal [47]. It is observed that its architecture is a Two Stream: a first stream with the *background suppression* technique followed by a 2D CNN MobileNetV2 network [34] and then with *separate convolutional LSTM* layers. A second stream with frame difference followed by another 2D CNN MobileNetV2 network [34], and then with *separate convolutional LSTM* layers, to finally join the flows with the final classifier. It is known that the number of parameters of MobileNetV2 is 3.4 million parameters; when using this network for each stream, there are 6.8 million parameters only in the 2D CNN networks. In this way, it is not easy to assume that the entire model presented by SepConvLSTM [47] only has 0.33 million parameters.

According to what was analyzed, we demonstrate that our proposal with 3.51 million parameters contributes to the state-of-the-art as one of the most compact proposals in terms of efficiency. It also demonstrates that it is suitable for use in real devices and scenarios. of video surveillance systems.

4.3.3. Results evaluation on VioPeru dataset

Since the SepConvLSTM proposal [47] is still above our proposal in terms of effectiveness, that is, *accuracy*; is a good candidate to test with the dataset presented in this research: VioPeru. In this way, both proposals' *accuracy* was calculated. Table 6 shows the results:

Table 6. Comparison of results on VioPeru dataset, taking efficacy as a reference.

Model	Accuracy(%)
SepConvLSTM (Islam et al.) [47]	**73,21**
Proposal	**89,29**

It is observed that in the case of the VioPeru dataset, our proposal is much superior to SepConvLSTMN [47] in terms of effectiveness, reaching an *accuracy* of 89.29% compared to 73.21% of the other proposal.

This result should not be analyzed in numbers alone, although our proposal is superior. However, it is necessary to clarify that the VioPeru dataset consists of scenes of real violence and non-violence extracted from real video surveillance cameras in the province of Andahuaylas in Peru. Therefore, it is for us a reference dataset to test our proposal from the point of view of its validation in a real-time scenario, which is ultimately the objective of our research.

This result also demonstrates that our model is not aimed at surpassing results in datasets with mixed and varied videos but rather at becoming a general-use model oriented to the domain of violence detection in video surveillance cameras in real-time since it does not only is it an efficient model, but it is also effective in real scenarios and also on state-of-the-art datasets.

4.3.4. Results evaluation in Real-Time

To the best of our knowledge, there is no formal method to measure whether a model can be used in real-time. However, we are based on similar works [29,44], and we take into account that evaluating results in real-time is done by measuring the processing time for every 30 frames, assuming that video surveillance videos have this default setting, that is, a speed of 30 frames per second.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0110.v1

The processing time of our proposal was measured for every 30 frames; for this, a laptop with a 2.7 GHz 13-core Intel Core i7 processor, 16 GB RAM, and NVIDIA Quadro P620 graphics card with 2 GB GPU memory was used.

The result was 0.922 seconds on average, which is the latency time of the model when processing 30 frames. In other words, our proposal only needs 0.092 seconds to process a 1-second video. If we consider that real-time has a latency of 0 seconds, the result is very close to real-time. However, from a human's point of view, the perception time of a latency of 0.922 milliseconds is considered real-time.

This result confirms that the proposal is lightweight to work on devices with low computational power, with cutting-edge results in effectively detecting violence in real-time.

5. Ablation study

According to the proposal presented in Section 3, the Short Temporal Extractor (STE) module uses a pre-trained 2D CNN network as a backbone. The proposal uses the MobileNetV2 architecture [34], mainly due to the cutting-edge results in ImageNet [**?**] with fewer parameters and FLOPS. However, according to Table 1, it is observed that there are other alternatives to be used as backbone since there are good candidates such as EffcientNet B0 [36], MobileNetV3 L[35] and MnasNet [37].

This section evaluates the proposal with the abovementioned models regarding effectiveness and efficiency on the RWF 2000 [22] and VioPeru dataset. Table 7 shows the results.

Table 7. Proposal results with different backbones.

Proposal Variations	RWF-2000 Accuracy(%)	VioPeru Accuracy(%)	Parameteres (M)	FLOPS (G)
Proposal with EfficientNet B0 backbone	88.25	87.5	5.29	4.17
Proposal with MobileNet V2 backbone	88.5	89.29	3.51	3.15
Proposal with MobileNet V3 backbone	88.25	89	7.62	4.1
Proposal with MNasNet backbone	75.25	62.5	2.22	1.13

Taking into account the VioPeru dataset, it is observed that the proposal has the best efficiency with the MobileNetV2 backbone, with an *accuracy* of 89.29%, followed by MobileNetv3 [35] with 89% and EfficientNet B0 [36] with 87.5%0. On the other hand, the proposal has better efficiency with the MobileNetV2 backbone [34], with a value of 3.51 million parameters and 3.15 GFLOPS.

Now, with RWF-2000 dataset [22], the proposal also has the best efficiency with the MobileNetV2 backbone, with an *accuracy* of 85.5%, followed by MobileNetv3 [35] and EfficientNet B0 [36] with the same value of 88.25%.

Although the proposal with the MNasNet [37] backbone has the best efficiency, with a value of 2.22 million parameters and 1.13 GFLOPS, it has a very low efficiency compared to the other backbones.

In this way, the backbone chosen is MobileNetV2, for having the best results in effectiveness and at the same time in efficiency.

6. Conclusions

In this work, a model based on deep learning is proposed for the recognition of violent human actions in real-time video surveillance. We propose an architecture of three modules. The first module,

Spatial Motion Extractor (SME), extracts regions of interest from a frame using frame difference and morphological dilation. A second module called Short Temporal Extractor (STE), whose function is to extract temporal features from fast and short-duration movements through temporal fusion and the use of the MobileNet V2 backbone. Finally, the Global Temporal Extractor (GTE) module identifies long-term temporal characteristics and fine-tuns the model for better precision, using Global Average Pooling and dot product. The tests were initially carried out on the RWF-2000, Movies, and Hockey datasets, reaching cutting-edge results in both effectiveness and efficiency; the results are only below the best proposal. Demonstrating that the current datasets are not oriented to video surveillance, a dataset called VioPeru was generated with real videos from video surveillance cameras in Peru; the results show that our proposal is the best in efficiency and effectiveness in VioPeru. The proposal has a recognition latency 0.922 seconds for every 30 frames; the latency is close to real-time. Our proposal has high efficiency and effectiveness in real-time video surveillance systems and can be used in devices with low computational power.

Author Contributions: Conceptualization, H.H. and F.P.; methodology, H.H.; software, H.H.; validation, F.P., H.H. and J.P.; formal analysis, J.C.; investigation, H.H.; resources, F.P.; data curation, F.P.; writing, H.H.; writing—review and editing, H.H.; visualization, F.P.; supervision, J.C.; project administration, F.P.; funding acquisition, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jose Maria Arguedas National University, Peru, as part of the 2022 research project competition, approved with Resolution No. 0456-2022-CO-UNAJMA.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets VioPeru used in this work can be accessed on the following link: https://github.com/hhuillcen/VioPeru.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
2. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
3. Shou, Z.; Wang, D.; Chang, S.F. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1049–1058.
4. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding* **2017**, *156*, 117–127.
5. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.
6. Gao, Y.; Liu, H.; Sun, X.; Wang, C.; Liu, Y. Violence detection using oriented violent flows. *Image and vision computing* **2016**, *48*, 37–41.
7. Deniz, O.; Serrano, I.; Bueno, G.; Kim, T.K. Fast violence detection in video. In Proceedings of the 2014 international conference on computer vision theory and applications (VISAPP). IEEE, 2014, Vol. 2, pp. 478–485.
8. Bilinski, P.; Bremond, F. Human violence recognition and detection in surveillance videos. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2016, pp. 30–36.
9. Zhang, T.; Jia, W.; He, X.; Yang, J. Discriminative dictionary learning with motion weber local descriptor for violence detection. *IEEE transactions on circuits and systems for video technology* **2016**, *27*, 696–709.
10. Deb, T.; Arman, A.; Firoze, A. Machine cognition of violence in videos using novel outlier-resistant vlad. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018, pp. 989–994.

11. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* **2012**, *35*, 221–231.
12. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* **2014**, *27*.
13. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
14. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
15. Dong, Z.; Qin, J.; Wang, Y. Multi-stream deep networks for person to person violence detection in videos. In Proceedings of the Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part I 7. Springer, 2016, pp. 517–531.
16. Zhou, P.; Ding, Q.; Luo, H.; Hou, X. Violent interaction detection in video based on deep learning. In Proceedings of the Journal of physics: conference series. IOP Publishing, 2017, Vol. 844, p. 012044.
17. Serrano, I.; Deniz, O.; Espinosa-Aranda, J.L.; Bueno, G. Fight recognition in video using hough forests and 2D convolutional neural network. *IEEE Transactions on Image Processing* **2018**, *27*, 4787–4797.
18. Sudhakaran, S.; Lanz, O. Learning to detect violent videos using convolutional long short-term memory. In Proceedings of the 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2017, pp. 1–6.
19. Hanson, A.; Pnvr, K.; Krishnagopal, S.; Davis, L. Bidirectional convolutional lstm for the detection of violence in videos. In Proceedings of the Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.
20. Ullah, F.U.M.; Obaidat, M.S.; Ullah, A.; Muhammad, K.; Hijji, M.; Baik, S.W. A comprehensive review on vision-based violence detection in surveillance videos. *ACM Computing Surveys* **2023**, *55*, 1–44.
21. Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R. Violence detection in video using computer vision techniques. In Proceedings of the Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14. Springer, 2011, pp. 332–339.
22. Cheng, M.; Cai, K.; Li, M. RWF-2000: an open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 4183–4190.
23. Ulutan, O.; Rallapalli, S.; Torres, C.; Srivatsa, M.; Manjunath, B. Actor Conditioned Attention Maps for Video Action Detection. In the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.
24. Zhang, C.; Zou, Y.; Chen, G.; Gan, L. Pan: Towards fast action recognition via learning persistence of appearance. *arXiv preprint arXiv:2008.03462* **2020**.
25. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
26. Lee, M.; Lee, S.; Son, S.; Park, G.; Kwak, N. Motion feature network: Fixed motion filter for action recognition. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 387–403.
27. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 305–321.
28. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7083–7093.
29. Huillcen Baca, H.A.; de Luz Palomino Valdivia, F.; Solis, I.S.; Cruz, M.A.; Caceres, J.C.G. Human Violence Recognition in Video Surveillance in Real-Time. In Proceedings of the Future of Information and Communication Conference. Springer, 2023, pp. 783–795.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

31. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
32. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
33. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360* **2016**.
34. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
35. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.
36. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 6105–6114.
37. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. Mnasnet: Platform-aware neural architecture search for mobile. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2820–2828.
38. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: enhance cheap operation with long-range attention. *Advances in Neural Information Processing Systems* **2022**, *35*, 9969–9982.
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
40. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
42. Singh, S.; Dewangan, S.; Krishna, G.S.; Tyagi, V.; Reddy, S.; Medi, P.R. Video vision transformers for violence detection. *arXiv preprint arXiv:2209.03561* **2022**.
43. Li, J.; Jiang, X.; Sun, T.; Xu, K. Efficient violence detection using 3d convolutional neural networks. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.
44. Huillcen Baca, H.A.; Gutierrez Caceres, J.C.; de Luz Palomino Valdivia, F. Efficiency in human actions recognition in video surveillance using 3D CNN and DenseNet. In Proceedings of the Future of Information and Communication Conference. Springer, 2022, pp. 342–355.
45. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, 2012, pp. 1–6.
46. Mumtaz, N.; Ejaz, N.; Habib, S.; Mohsin, S.M.; Tiwari, P.; Band, S.S.; Kumar, N. An overview of violence detection techniques: current challenges and future directions. *Artificial intelligence review* **2023**, *56*, 4641–4666.
47. Islam, Z.; Rukonuzzaman, M.; Ahmed, R.; Kabir, M.H.; Farazi, M. Efficient two-stream network for violence detection using separable convolutional lstm. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1–8.
48. Su, Y.; Lin, G.; Zhu, J.; Wu, Q. Human interaction learning on 3d skeleton point clouds for video violence recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer, 2020, pp. 74–90.