Machine Vision and Applications (2022) 33:15
https://doi.org/10.1007/s00138-021-01264-9

SPECIAL ISSUE PAPER

Inflated 3D ConvNet context analysis for violence detection

David Freire-Obregón[1] · Paola Barra[2] · Modesto Castrillón-Santana[1] · Maria De Marsico[2]

Received: 15 April 2021 / Revised: 16 July 2021 / Accepted: 24 October 2021 / Published online: 31 December 2021

Abstract
According to the Wall Street Journal, one billion surveillance cameras will be deployed around the world by 2021. This amount of information can be hardly managed by humans. Using a Inflated 3D ConvNet as backbone, this paper introduces a novel automatic violence detection approach that outperforms state-of-the-art existing proposals. Most of those proposals consider a pre-processing step to only focus on some regions of interest in the scene, i.e., those actually containing a human subject. In this regard, this paper also reports the results of an extensive analysis on whether and how the context can affect or not the adopted classifier performance. The experiments show that context-free footage yields substantial deterioration of the classifier performance (2% to 5%) on publicly available datasets. However, they also demonstrate that performance stabilizes in context-free settings, no matter the level of context restriction applied. Finally, a cross-dataset experiment investigates the generalizability of results obtained in a single-collection experiment (same dataset used for training and testing) to cross-collection settings (different datasets used for training and testing).

Keywords Violence detection · People tracking · I3D model · Context analysis · Transfer learning

1 Introduction

Continuous monitoring of visual streams for the timely detection of emergency/anomalous situations is critical for effective intervention whenever two or more persons can interact, especially in public spaces. A common example is represented by protest demonstrations, but also sport events or crowded environments can require this kind of regular activity for law enforcement. Violence detection stems in a sense from action recognition but aims solely at recognizing violent actions. From one side it is more general, since it relies on a pure binary classification, but on the other side just for the same reason it may result more complex. It requires to train a classifier on a whole class of actions. It could be worth clarifying the terms used in the following. The term "category" is borrowed from literature to indicate single types of actions, i.e., combinations of gestures that, though being naturally performed in different ways, have the same effect (e.g., walking, drinking, dancing, etc.). Action recognition deals with action categories. The term class rather indicated that more categories that can be very different from each other can be further grouped according to a criterion, which in this paper is violent/non-violent. This can be done by capturing their shared characteristics like, e.g., a generally high gesture speed joined to a closer distance among subsets of subjects. Violence detection in videos is especially useful in the context of video surveillance. Precisely, video surveillance typically involves the act of observing a scene and looking for improper behaviors or events. These may include violence and robbery among other crimes. Traditional methods for surveillance-based crime detection still involve the human intervention. This is not effective for two reasons: the often not negligible security staff costs and the risk of failure by human error due to distraction or fatigue. Lately, artificial intelligence is increasingly being integrated with video surveillance systems to overcome these issues. Of course a human-in-the-loop approach, i.e., the intervention of a human operator, is still needed to confirm alarms. The advantage is that these can be automatically raised by an automatic system therefore relieving the operator from the burden of a continuous attention. This is especially useful with multiple surveillance cameras, e.g., to decide which surveillance video stream to display on the main monitor for anomaly confirmation [1,2].

✉ Paola Barra
barra@di.uniroma1.it

[1] SIANI, Universidad de Las Palmas de Gran Canaria, Las Palmas, Spain

[2] Sapienza University of Rome, Rome, Italy

In this regard, many approaches in the literature consider only specific regions of interest (ROIs) (those actually containing human subjects) and they can be consequently considered as "context-free". To this aim, some works apply some pre-processing step to extract specific ROIs, therefore losing all the context along the process [5,6]. Other works rather consider the overall scene context, i.e., they compute the absolute image difference between consecutive frames [7,8] before the features extraction step. All these works achieve remarkable results by using regular machine learning classifiers [6–8] or deep neural networks [5]. While in some cases the context can be of help, in other cases it could negatively influence the final performance. In any case, it could represent a bias when running a method on a dataset different from the one used to develop and train the classifier.

Paper contributions. This work takes a step towards the context-dependence analysis in violent scenes by four main contributions.

- We introduce a violence classifier built on top of a pre-trained deep neural network that reports highly competitive results in action recognition. The classifier provides a binary violence/non-violence response on input video clips. The results achieved by the proposed classifier on the original videos improve or equal the state-of-the-art baselines not only on the previously commented datasets, but also on other datasets collected in crowded scenarios.
- We devise an analysis protocol to investigate whether or not the context affects the performance of the proposed classifier. To this aim, the classifier pipeline includes a preliminary parameterized context removal operation based on people detection and tracking techniques (see Fig. 1) that evaluates and exploits the amount of overlap between pairs of bounding boxes (BBs) enclosing single detected subjects. The parameter somehow reflects the adopted notion of "context". When a 0% overlap is allowed, it is the case of complete background removal despite the amount of overlap of the BBs detected in the scene. When the parameter value increases, the procedure discards not only the background but also the "isolated" subject BBs or BBs that do not present a sufficient amount of overlap (for the aim of violence detection). Therefore, the proposal assumes that BBs must *touch* each other in violence episodes.
- The obtained results are used to analyze the context influence based on the overlap threshold. We anticipate that context removal causes lower performance, but such performance stabilizes notwithstanding the value of the overlap parameter chosen and can be useful to decrease the computational burden, so that a deeper analysis of frames can be triggered only when a "suspect" of violence is detected.

Fig. 1 Different levels of context removal. We analyze the violence detection performance with a model that efficiently classifies the video into violent/non-violent in a single forward pass and includes a parameter driving an automatic context-removal process. The behavior of the bounding boxes (BBs) enclosing the single subject images plays a key role in the context removal. The two columns show frames of video taken from Hockey Fight Dataset [3] (left) and AVD Dataset [4] (right). The first row shows original frames, while the second and third rows show frames where an increasing overlap between pairs of BBs is imposed

- Experiment further analyze the effect of cross-dataset classification (different training and testing datasets).

Research questions. As anticipated regarding paper contributions, after comparing the performance of the proposed system with the state of the art, further experiments estimate the context influence on two public datasets for violence detection, the widely used Hockey Fight Dataset (1000 clips) [3] and the novel AVD (Automatic Violence Detection) Dataset (350 clips) [4]. The final experiment uses more datasets to evaluate possible performance degradation in cross-dataset classification. The aim is to answer four questions:

1. How important is the context in order to detect violent actions?
2. Is the context equally important for different datasets?
3. At what extent can be the context simplified? Does this simplification come with a cost?
4. Is it possible to collect a training dataset able to support a generalized classification accuracy even when classifying data from different sources?

The reported results assess the negative influence of context removal on the classification accuracy, although this does not significantly depend on the removal extent. The cross-dataset experiment provides interesting insights on cross-dataset violence classification.

The paper is organized as follows. The next section discusses some related work in the state of the art. Section 3 describes the proposed context-removal pipeline. Section 4 reports the experimental setup, the experimental results and the cross-dataset experiment. Finally, Sect. 5 draws conclusions.

2 Related work

The most relevant state-of-the-art methods can be divided into those that use or do not use deep learning.

2.1 Classical approaches

To solve the problem of detecting violent actions within videos, the pixel-by-pixel differences of consecutive frames in a sequence are often used as descriptors to detect movements. The work proposed in [7] introduces the motion blobs of the scene, which are computed by this difference. They are represented by the non-0 pixels after binarizing and clustering them. The following steps only use the K largest ones and their centroids. The analysis of the size of the blobs allows estimating their speed between consecutive frames. Features extracted from motion blobs allow discriminating fight and non-fight sequences. The classification is not linked to the number of people in the video but to the movements detected, so that it also allows detecting acts of vandalism in which the author can even be a single person.

A Gaussian Model of Optical Flow (GMOF) is proposed in [9] to extract the candidate regions in which violent acts occur. Violent acts are recognized as a deviation from the normal behavior of the crowd in the scene. A Support Vector Machine (SVM) using data from a Histogram of Optical Flow (OHOF) descriptor is used to classify violent frames and non-violent ones.

The work in [10] presents an interesting use of optical flow to derive a descriptor called Violence Flows (ViF) that estimates the optical flow between consecutive pairs of frames in a sequence. This descriptor is used to collect the significant information in the video to classify it as violent or non-violent by a SVM.

An original method is proposed in [11]. The authors manually tag videos from the MediaEval dataset [12] into subclasses that are visually related to violence. This information together with audio, motion and image features contributes to the data to train a SVM classifier. The method is not strictly linked to the training dataset so it can also be used on other unlabeled videos. Furthermore, the method is not related to the movement within the video, but rather to the content.

2.2 Deep learning approaches

In the method presented in [5], an entire video sequence is summarized in a single grayscale image describing its movement content. Then, a 2D convolutional neural network is used to classify the obtained image.

The methods proposed in [16] present two video detection schemes based on 3D ConvNet [17], which can learn the spatiotemporal characteristics of the video without using any prior knowledge. The 3D ConvNet consists of a 2D convolutional neural network that takes as input frames in gray scale in which the third dimension is the temporal information.

Several methods mix different solutions to solve the problem. In [18] the authors exploit two ConvNet streams: a temporal stream to describe the violent movements with the features related to the trajectories of the movements in the frames and a spatial stream to analyze the scene through deep learning features. Also the authors in [8] analyze both temporal and spatial changes and introduce an architecture that they call convLSTM: they combine a convolutional neural network with an LSTM (Long short-term memory). The convLSTM architecture takes in input a sequence of video frames that will be classified as violent or not. The latest state-of-the-art method that uses deep learning for video-based violence detection is presented in [19]. The authors use various CNN architectures for feature extraction, such as VGG16 [20] and Xception [21]; then a Fight-CNN is trained for fight detection, with frame labeled fight and non-fight. For the classification a Bi-LSTM is used, to learn the dependency between past and future information. Then, an added attention layer determines the significant input regions.

3 Violence classification pipeline

This work proposes and evaluates a sequential pipeline divided into two main modules, namely the Tracking Loop and the Classification Block, where the former feeds the latter as shown in Fig. 2.

The first module implements the subjects detection and tracking. It provides the necessary information to locate and label all the subjects in the scene. The output of the loop is represented by the BBs enclosing these subjects. In more detail, each subject is located and labeled (same label indicates the same subject across frames) by the Deep SORT algorithm [13], while the SiamRPN+ network [14] supervises this process as will be explained in Section 3.1. The BB context information can be optionally removed according to an overlap parameter σ that will be better described in the following. In this module, the input footage is processed frame by frame in order to generate a temporal sequence of frames as input for the next module.

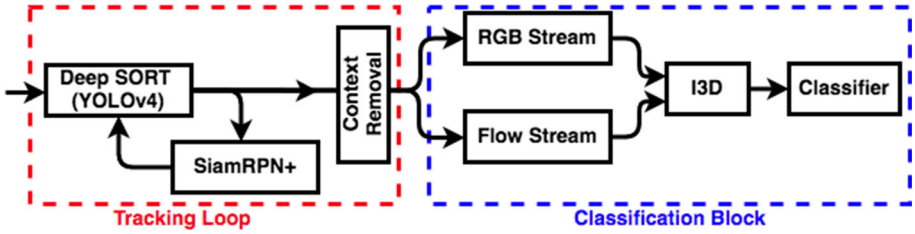


Fig. 2 The proposed pipeline for the violence/non violence context-driven problem. The devised process comprises two main parts: the Tracking Loop and the Classification Block. The Tracking Loop aims to detect and label the subjects through Deep SORT [13] and the visual tracking Siamese Network (SiamRPN+) [14]. The Classification Block implies the generation of two streams of data (RGB and Flow) to feed the Inflated 3D ConvNet [15] and the classification process using the extracted embeddings

More precisely, two different streams are generated from the context-free data and feed the Inflated 3D ConvNet [22] in the second block, namely an RGB Stream and a Flow Stream that will be described in the following. The neural network is used to extract the embeddings considered in the last step. Finally, a classifier decides whether or not the provided content is violent. The classification is not carried out frame by frame, but on a per video basis according to the streams received as input. The following subsections will describe each step in more detail.

3.1 People tracking

Object tracking has played a relevant role in Computer Vision in the past three decades. Several applications has benefited from it such as, e.g., video surveillance [23], human-computer interaction [24], or unmanned vehicle driving [25]. Before deep learning, traditional algorithms such as Kalman filtering [26], the multiple hypothesis tracking [27] and the joint probabilistic data association filter [28] were considered as standards. They mostly use image edge features and probability density to make the object search direction agree with the direction of the rising probability gradient.

As for other fields, the evolution of the recent deep learning techniques represents a real breakthrough also for the visual object tracking. Lately, tracking-by-detection has become prevalent [29]. In this regard, the Simple Online and Realtime Tracking (SORT) [30] has shown a remarkable performance in comparison with other tracking algorithms such as TDAM [31] and MDP [32]. An extension of that algorithm, SORT with deep association metric (Deep SORT) [13] has been proposed for pedestrian detection. Recently, Deep SORT has reported the most stable tracking results in a qualitative evaluation of these algorithms in the sports domain [33].

In the proposed approach (see Fig. 2), the Deep SORT algorithm [13] is exploited as the first tracking step. The goal is not only to track people, but also to correctly label the subjects in the scene. The core idea of this algorithm is to combine the Kalman filtering and Hungarian algorithm for tracking purposes. Wojke et al. assume that the Mahalanobis distance is a suitable association metric when motion uncertainty is low. However, unaccounted camera motion can introduce rapid displacements in the image plane, making the Mahalanobis distance a rather uninformed metric for tracking across occlusions. Therefore, the algorithm integrates a second metric into the assignment problem that underlies tracking by computing an appearance descriptor of each BB and then measuring the smallest cosine distance between the i-th track and j-th detection in the appearance space [13]. The cost function can be expressed as shown in Eq. (1):

$$c_{i,j} = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j) \tag{1}$$

where $d^{(1)}$ denotes the Mahalanobis distance of the detected BB from the position predicted according to the previously known position of the corresponding object, while the visual distance $d^{(2)}$ considers the appearance of the presently detected object compared with the history of appearance of the tracked object to which it is expected to correspond. The present proposal exploits a recent version of the algorithm.[1]

Even though Deep SORT achieves overall good performance in terms of tracking precision and accuracy, the kind of situations considered in violence detection can raise peculiar problems. As a matter of fact, fight scenes present aggressive human pose changes and occlusions that lead to a relatively high number of identity switches (see the left column of Fig. 3). For this reason, the Tracking Loop includes a second tracker. If the Deep SORT fails to properly identify a subject, then the SiamRPN+ network [14] feeds the Deep SORT in order to adjust the tracking process. This neural network has been introduced as an evolution of SiamRPN and two key-features characterize the new version. First, a residual unit with cropped operation is added to address the limitation of the bottleneck convolution, allowing to neatly remove padding-affected features in the residual unit.

[1] https://github.com/theAIGuysCode/yolov4-deepsort.

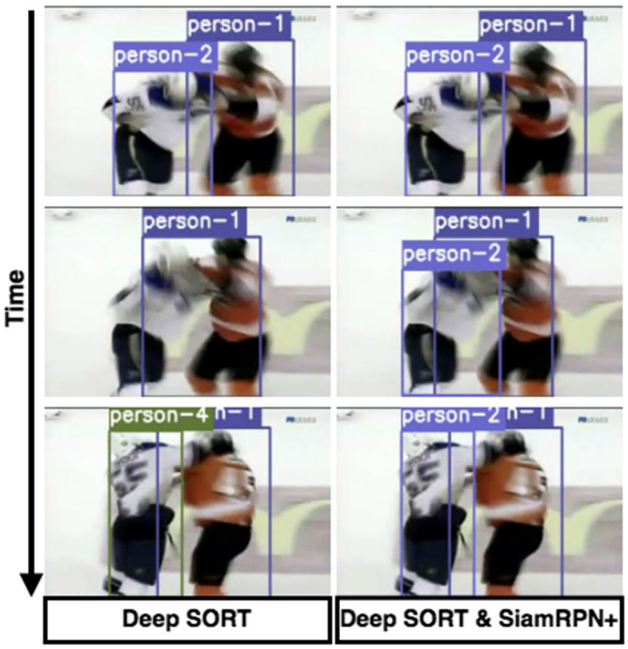


Fig. 3 Example of the tracking process. Deep SORT detected subjects (their BBs) are shown in the left column, while the right column shows the results when the SiamRPN+ is used as backup for the Deep SORT detection

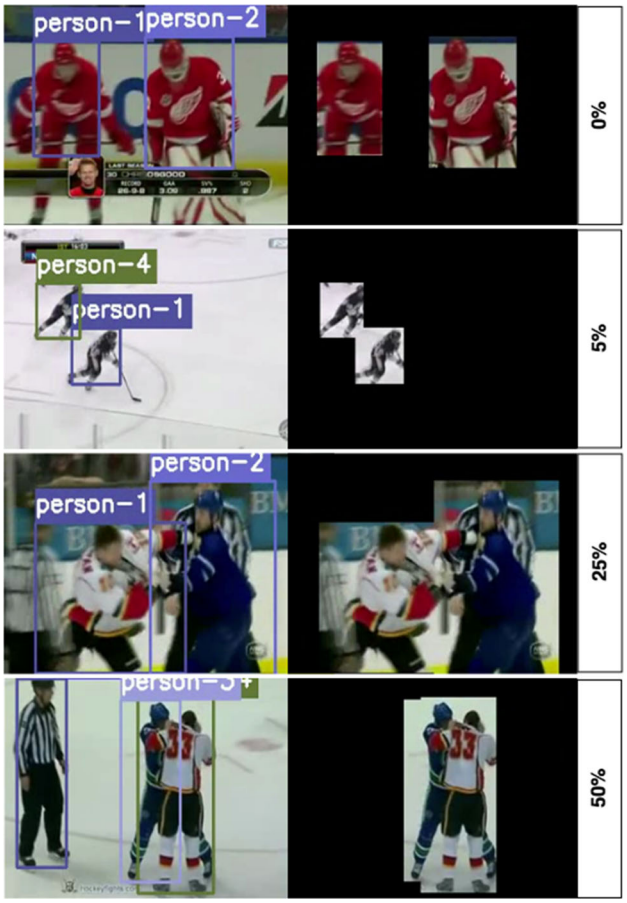


Fig. 4 Samples with BBs with different amount of overlapping. The effect of different overlapping thresholds that, when applied to the original frames (on the left), determine their associated context-free images (on the right). The overlapping threshold parameter indicates that only the detected bounding boxes (BBs) that overlap by at least σ% of their area are shown in the resulting image. Hence, $\sigma = 0\%$ means that all BBs are further processed, while $\sigma = 50\%$ causes that only overlapped BBs by at least 50% of their area enter the next step

Second, SiamRPN+ benefits from a deeper backbone like ResNet, leading to a remarkable performance and robustness [34].

The way Deep SORT and SiamRPN+ interact in the Tracking Loop can be observed in detail in Fig. 2. The latter acts as a backup for the former. It continuously updates the state and only comes into play if the former loses the track. Deep SORT provides the people labeling (e.g., person-1, person-2, etc.). SiamRPN+ only follows tracks with no prior labeling process. When BBs overlap, then a reference can be lost and the Deep SORT possibly assigns a new label to an already labeled BBs. The adopted solution is this backup labeling Siamese network that feeds Deep SORT when the reference is lost. Thus, the detection of the i-th track in the current frame ($d_t(i)$) can be formulated as follows:

$$d_t(i) = \rho \times \Psi_{DS}(\tau_{t-1}(i)) + (1-\rho) \times \Psi_{SRPN}(\tau_{t-1}(i)) \quad (2)$$

where ρ is a binary value that denotes the positive detection of the i-th track in the current frame, $\tau_{t-1}(i)$ is the i-th track in the previous frame and Ψ represents both tracking approaches, Deep SORT (Ψ_{DS}) and SiamRPN+ (Ψ_{SRPN}), respectively. As can be seen in Fig. 3, the integrated system exhibits a higher labeling consistency and consequent robustness.

The final optional step in the Tracking Loop, namely Context Removal in Fig. 2, removes the context depending on the behavior of each positive detection (a BB has been successfully detected in the current frame during tracking). The parameter σ mentioned above determines the "useful neighborhood" of a BB, i.e., its relevant surrounding region, that causes it to be considered in the following steps or to be treated as a part of the "background" (intended as information not taken into account for violence detection). In detail, the percentage of overlap between pairs of BBs is compared with this threshold σ, which is one of the parameters that have been taken into account in the presented experiments. Those BBs with an overlap below the selected threshold are not considered and they are consequently masked from the frame like the rest of the background (see Fig. 4). More formally, given two positive BBs m and n, they are both included in the processed frame (ctxf_t) if the intersection between their areas is bigger or equal to the chosen threshold; this is determined for each pair $< m, n >$ by the following Boolean function:

$$\begin{aligned} \mathrm{ctx}_t(m,n) &= (\mathrm{Ad}_t(m) \cap \mathrm{Ad}_t(n)) \\ &\geq \sigma\, \mathrm{Ad}_t(m) \forall m, n \in \Omega \end{aligned} \tag{3}$$

where Ω represents the space of all possible BBs, while $\mathrm{Ad}_t(m)$ and $\mathrm{Ad}_t(n)$ denote the areas detected in the current frame for the m-th and n-th tracks, respectively.

3.2 Two-stream inflated 3D ConvNets for action recognition

Computer vision algorithms for human action recognition have achieved remarkable progress in the last years. In particular, action recognition accuracy has been significantly improved. The collection of large-scale video datasets and the developments of methodologies and architectures based on convolutional neural networks (CNNs) mainly contribute to this progress [35,36]. As an interesting example, the work by Simonyan and Zisserman [37] proposes a two-stream 2D CNNs that uses both RGB and optical flow frames to process both appearance and motion information, respectively. The experimental results show that the combination of the two streams can significantly improve the action recognition accuracy.

A few years later, Carreira and Zisserman proposed the Inflated 3D Convnet (I3D) also based on a two-stream network [22]. Unlike its predecessors, the I3D applies the two-stream structure for RGB and optical flow to the Inception-v1 [38] along with 3D CNNs. It uses these 3D CNNs to learn spatiotemporal information directly from videos. To do so, it converts 2D classification models into 3D ones by training with multiple frames at once instead of one by one. From the implementation perspective, it starts with a 2D network using asymmetrical filters for max-pooling, maintaining time while pooling over the spatial dimension. Then, it inflates all the filters and pooling kernels so that they become cubes instead of squares. Hence, it can learn from multiple frames at once. In terms of performance, accuracy on representative action recognition collections such as UCF-101 [39] and HMDB-51 [40] improves from 88.0 and 59.4% [37] to 97.9% and 80.2%, respectively [22].

At present, I3D is one of the most common feature extraction methods for video processing. The approach presented in this work exploits the pre-trained model on the Kinetics dataset as a backbone model [22]. The Kinetics dataset [41] is a large action recognition dataset that includes a large number of action categories. In the present proposal, the backbone model has been trained with the Kinetics version of 400 action categories, each action category being represented by approximately 400 video clips. Consequently, the I3D (see Fig. 2) acts as a feature extractor to encode the network input into a 400 vector feature representation that feeds the classifiers described and evaluated in the next section.

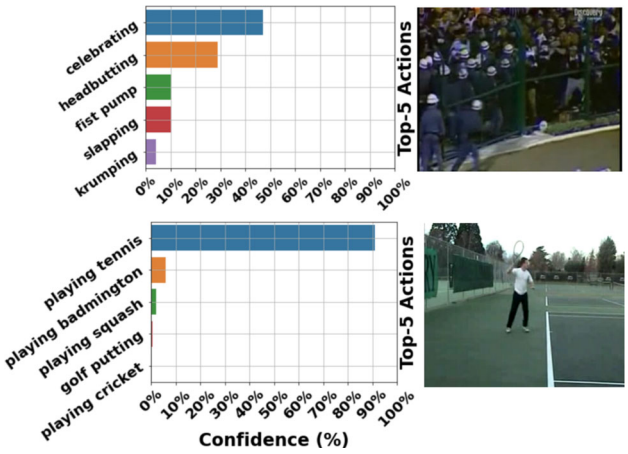


Fig. 5 Examples of I3D action recognition predictions. I3D top-5 predictions for two video clips. The top sample belongs to a violence video from the Crowd Violence Dataset [10], while the bottom sample belongs to a non-violence video of the Kaggle Movies Dataset. The shown predictions are computed on the entire video clips

Each element in a vector (prediction) represents the probability returned by I3D that an entire video clip represents the corresponding action (see Fig. 5).

3.3 Classification approaches

In the last part of the proposed pipeline (see Fig. 2), the feature vectors extracted as described in Sect. 3.2 feed the selected classifier in order to provide a two-class prediction (violent/non-violent). The goal of the experiments has been to evaluate some well-known state-of-the-art supervised classifiers, also considering the influence of the BBs' overlap parameter (σ). This section lists and briefly explains the used classifiers:

- *Decision Tree (DT)*. It is a widely used non-linear machine learning technique [42]. Given a n-dimensional space, the decision tree tries to partition this space into regions while trying to approximate the solution. It is a popular estimation method that exploits a tree-like structure and can complete a separate classification task for each branch. Therefore, in this model the data are divided into smaller groups and a decision tree is created.
- *Random Forest (RF)*. Unlike the decision tree, this technique does not rely on a single decision tree but on many of them [43]. In fact, the Random Forest algorithm builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- *XGBoost*. The acronym stands for eXtreme Gradient Boosting. It is an ensemble machine learning algorithm that builds a strong model based on many weaker ones applied sequentially. To do so, it uses gradient descent with decision trees [44].

– *Linear SVM (LSVM)*. It is a linear classifier that attempts to find a hyperplane with the largest margin that splits the input space into two regions [45].
– *Logistic Regression (LR)*. This algorithm examines the relationship between dependent and independent variables [46]. It has a low variance due to its simple operation structure and is less prone to overfitting.

4 Experimental evaluation

4.1 Experimental setup

This work not only aims to present the performance under different context conditions but also to establish a valid baseline that shows the robustness of the proposed classifier pipeline (see Fig. 2). For this reason, the experiments have been carried out on several state-of-the-art datasets for violence detection. All the considered datasets were explicitly designed for evaluating violence detection performance, and they are all freely available for scientific purposes (see Sect. 4.2).

The results presented in the following refer to the average performance computed over 100 iterations. For each iteration, train and test data are chosen randomly and the results are averaged after considering a stratified fourfold cross validation.

4.2 Datasets

The first two datasets were presented in the same work [3]. The first one, the Hockey Fight Dataset, consists of 1000 clips extracted from hockey games of the National Hockey League. They are divided into two groups, 500 labeled as "fight" and 500 labeled as "non-fight". The second one is the Movies Dataset and it consists of 200 video clips (100 samples per class) in which fights were extracted from action movies. The third dataset is the Crowd Violence Dataset [10] that consists of 426 clips (123 samples per class) in which events occur in crowded environments. The last dataset is a novel proposal for Automatic Violence Detection (AVD Dataset) in videos [4]. It is composed of 350 clips, labeled as "non-violent" (120 clips) and "violent" (230 clips) depending on the represented behavior.

Out of these datasets, only the Hockey Fight Dataset and the AVD Dataset were considered for the context-removal experiment due to the feasibility of humans detection. It is worth underlining the different kind of context/background of videos in the two collections. In the first one, the background is more noisy and represents a stadium scenario. The videos in the second one have been recorded in an empty room. The other two datasets, namely the Movies Dataset and the Crowd Violence Dataset, have a wider variety of scenes that were captured at different resolutions and they are just used to establish a valid baseline between our classifier and the different state-of-the-art proposals.

4.3 Experimental results

The first set of experiments aimed to establish a valid baseline of the described proposal. As stated before, all the previously described datasets were considered. Moreover, these experiments used the original videos of each dataset in the Classification Block (see Fig. 2), i.e., no background information was discarded (σ =None), as explained in Sect. 3.

Table 1 The best results achieved by each considered classifier

Dataset	Overlap th. (σ)	#Samples (v/nv)	DT	RF	XGBoost	LSVM	LR
Crowd violence	None	123/123	84.81%	94.13%	98.08%	99.14%	**99.45**%
Movies	None	100/100	92.27%	98.03%	99.57%	**100.00**%	**100.00**%
Hockey fight	None	500/500	90.62%	97.29%	98.59%	99.42%	**99.43**%
	0%	488/495	85.28%	94.26%	97.48%	97.65%	97.44%
	5%	407/375	87.71%	95.31%	97.67%	97.73%	97.31%
	25%	386/350	85.77%	94.71%	97.49%	97.65%	97.30%
	50%	332/309	85.68%	94.37%	97.28%	97.40%	97.22%
AVD	None	230/120	74.03%	78.27%	94.85%	97.15%	**97.54**%
	0%	230/120	67.85%	74.55%	90.35%	92.38%	92.95%
	5%	226/120	67.04%	73.17%	89.11%	91.61%	92.23%
	25%	223/118	67.99%	73.93%	90.08%	92.12%	92.25%
	50%	208/109	66.49%	73.71%	88.62%	91.91%	91.97%

The table is organized in terms of datasets and overlap threshold (σ), when the latter is investigated. Moreover it reports the number of violence/non-violence samples that survive after context deletion, depending on the overlap threshold. When overlap is None, it means that the input used is the original video

Table 2 Comparison of different approaches on the datasets used in the present work

Approach	Crowd violence	Movies	Hockey fight	AVD
Blob features + RF [7]	–	97.8 ± 0.4%	82.4 ± 0.6%	–
Extended IFV [47]	**96.4**%	99.0%	93.4%	–
Dense HOG + OR-VLAD [48]	93.1 ± 1.4%	**100.0**%	98.2 ± 0.76%	–
Dense HOG + VLAD [48]	91.1 ± 2.77%	**100.0**%	97.6 ± 0.08%	–
3D CNNs [49]	94.3%	99.97%	**99.62**%	–
CNNs + LSTM [8]	94.6 ± 2.34%	**100.0**%	97.1 ± 0.55%	–

Since AVD is a relatively new dataset, there are no works yet reporting classification accuracy rates (−). See Sect. 4.1 for a detailed description of these datasets. The rows corresponding to similar conditions in Table 1 are those labeled as "None" in the column reporting the overlap threshold

Table 1 summarizes the results in the rows corresponding to $\sigma = None$. To better compare the present work with state-of-the-art proposals, the results in these rows can be compared with those in Table 2, which summarizes the performance reported in recent literature on the mentioned datasets. The two tables show in bold the methods in which the accuracy is highest for a given dataset. From an overall perspective, our classifier outperforms or equals other considered prior works on those datasets as well as reports remarkable accuracy rates on the novel AVD Dataset.

Table 1 also shows the model performance on different datasets under the overlapping threshold variations, i.e., how the context reduction affects the model performance. Clearly, when considering the whole image (σ =None), the model performs best in any considered case. Three related issues are worth highlighting.

First, reducing the context may come with a computational advantage, i.e., the system may be faster if it only needs to process a small fraction of the scenes [50]. As can be seen in the third column, the number of samples to classify is reduced along with the increasing value of σ. The reason for this reduction is because there may be some video frames that do not fit the overlapping conditions, i.e., a single subject in the scene or cases when Eq. 3 is not satisfied by the overlapped detections (the equation returns many False results). These situations lead to empty video frames that are automatically discarded. It can be also appreciated that this reduction in the number of samples is not the same on both context-analyzed datasets. In this sense, the Hockey Fight Dataset has a reduction of a 36% of the original number of frames, while the AVD Dataset has a reduction of just the 10%. This can be explained in terms of context diversity that has already been sketched above. Whereas the Hockey Fight Dataset collects clips taken by moving cameras in a wider sportive scenario, the AVD Dataset just contains clips recorded by static cameras in an empty room.

The second issue is related to the first and is represented by the fact that the reduction of the computational demand comes with a cost. Table 1 shows a significant loss of performance under context removal. It can be appreciated that the performance loss is higher on the AVD Dataset than the Hockey Fight Dataset. However, this loss is stable for any given σ (see Fig. 6). For instance, the loss is roughly a 2% in the case of LR on the Hockey Fight Dataset and a 5% in the case of the same classifier on the AVD Dataset.

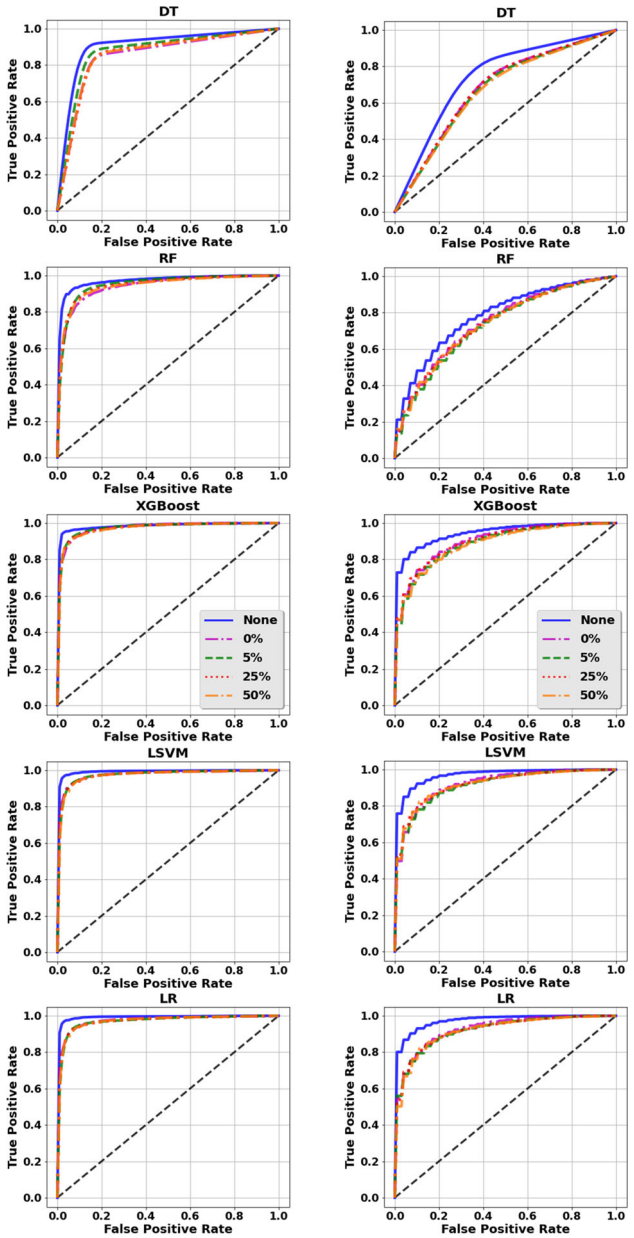
Finally, regarding the classifiers, it is possible to observe that LSVM and LR outperform any tree-based approach, and rates are quite similar for both of them. Among the tree-based approaches, XGBoost reports the best rates, specially on the AVD Dataset. This makes sense: RF builds trees in parallel, while in boosting trees are built sequentially, i.e., each tree is grown and boost using information from previously grown trees.

A final remark regards processing times. The extensive experiments with four GTX 1080Ti GPUs show that a model trained on the proposed set of images from different sources can achieve accurate results. The average processing time distribution is as follows: given a 100% of prediction time, the Flow-Stream computation requires a 72%, the RGB-Stream computation just a 1% and the I3D prediction a 27%, respectively (see Fig. 2). The Flow-Stream computation is therefore the bottleneck for our proposal (0.2 s per frame).

In summary, in this section we have shown how the I3D has exhibited a remarkable performance on some action recognition datasets. We have provided an extensive study on how context constraints affect this deep neural network. Moreover, training deep learning models on a single dataset leads to good performance on the corresponding test split of the same dataset (same camera configuration and same environment) [51]. This may have some limitations in terms of generalization capabilities to unseen data with different characteristics. In the following subsection, we propose a cross-dataset experiment to evaluate whether our approach exhibits a promising generalization capability by testing on diverse datasets not seen during training.

4.4 Cross-dataset experiment

Recently, Ullah et al. proposed an interesting work using spatiotemporal features with 3D CNNs [52]. The work considers



(a) Hockey Fight **(b)** AVD

Fig. 6 ROC curves computed from the results of the different approaches on both datasets. The left column shows the ROC curves for the Hockey Fight Dataset, while the right column shows the ROC curves for the AVD Dataset

three datasets also included in our proposal: the Hockey Fight Dataset, the Crowd Violence Dataset and the Movies Dataset. The discussion section of the cited paper presents the results of a cross-dataset experiment: the training set includes one of the dataset while the test set includes the remaining collections. The reported performance notably drops in comparison to the reported rates when the model is trained and tested on the same dataset. This suggests to get further insight into this issue.

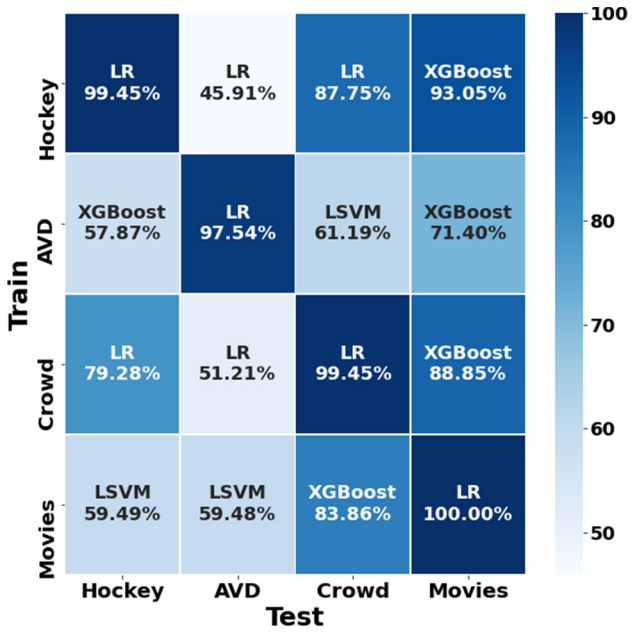


Fig. 7 Cross-dataset results using the entire videos. The labels on the *y*-axis indicate the training dataset, while those on the *x*-axis indicate the test dataset. Each cell reports the result of the best classifier for each training–test pair. The main diagonal corresponds to the results reported in Table 1

The results reported in Table 1 are remarkable. However, a question arises when the model is trained and tested on the same dataset: is the used dataset biased? Can a specific kind of context make the trained classification model little or not generalizable to other datasets? This subsection discusses a final extensive cross-dataset experiment to address this issue. It followed the same procedure described in Sect. 4.1. Figure 7 shows the best rate reported for each experiment configuration considering the original video clips of each collection. This matrix shows a blue heatmap where a darker color represents better rates than a lighter color. As can be seen and expected, the darkest cells are located in the main diagonal. Precisely, this diagonal shows the best results of Table 1, when a stratified fourfold cross validation was considered to split each dataset into training and test subsets. The remaining cells show the best results when the models are trained and tested on different datasets.

The matrix provides interesting insights regarding the considered datasets. The reported rates suggest that the AVD Dataset achieves a low performance in both cross-dataset cases, when it is used for training and when it is used for test. These rates seem to suggest that classes are not well-separable considering the extracted features from that collection. Therefore, the AVD Dataset is not suitable for cross-dataset violence detection generalization, notwithstanding the definitely neutral scenario. Another interesting aspect can be appreciated by observing the rates provided by the Hockey Fight Dataset. This dataset is challenging for test-

ing purposes, but it works really good as a training collection (except when AVD is the test collection, as for other training collections). In this regard, the Crowd Violence Dataset also provides a very interesting framework. This collection is not only suitable to be used for training but also to be used as a test dataset. Finally, the Movies Dataset exhibits a good performance when it is used for test, but this collection is not worthy for training due to the small number of samples that it contains (200 samples, 100 per class).

Probably, the Crowd Violence Dataset and the Hockey Fight Dataset are the most generalizable collections according to the reported rates when they are used for training. Unlike the former, the latter seems to be a more challenging dataset for test. However, there is an imbalance between the number of samples of each collection (see Table 1). The Hockey Fight Dataset has roughly 5× more samples than the Crowd Violence Dataset, and that issue must be taken into consideration.

4.5 Responses to research questions

According to the presented results, we can shortly answer the four research questions in the Introduction:

1. **RQ**: How important is the context in order to detect violent actions? **A**: The context is important even when limited.
2. **RQ**: Is the context equally important for different datasets? **A**: The context relevance depends on its relationship with the video actions, i.e., a neutral context like an empty room has a lower effect on the classification.
3. **RQ**: At what extent can be the context simplified? Does this simplification come with a cost? **A**: The simplification allows a significant processing speed up, but negatively and significantly affects the classification accuracy, although the negative effect is not proportional to the amount of simplification.
4. **RQ**: Is it possible to collect a training dataset able to support a generalized classification accuracy even when classifying data from different sources? **A**: This is an open problem. Cross-dataset classification achieves definitely lower performance. Torralba and Efros [53] conclude their analysis on possible dataset bias supposing that collections that are gathered automatically from the Internet are far better than the ones collected manually, which, quoting the authors, can "*become closed worlds unto themselves*". In this regard, even though the cited paper refers to object recognition datasets, our work supports this statement in a more general way.

5 Conclusions

This paper presented a novel approach to determine whether a video clip contains violence content or not. The proposed classification pipeline generally outperforms state-of-the-art techniques on publicly available datasets. To achieve tracking of the relevant subjects in the scene, the presented pipeline exploits two relevant tracking techniques such as Deep SORT and SiamRPN+, respectively. These allow to determine the possible overlap of subjects' BBs that drives the context removal process. Then I3D is used as feature extractor to feed several tested classifiers. In addition, the reported experiments have demonstrated that context plays a key role during the classification process. The results show that accuracy drops on a regular basis when context-free or context-reduced video clips are considered as input to the classifiers. However, accuracy stabilizes no matter the level of context removal, and this is counterbalanced by the gain represented by a reduced computational effort. A final study investigates which datasets are generalizable and suitable to train classifiers for violence detection, meaning that they can be used for training whatever is the data used for testing or in real operation. The results of the conducted cross-dataset experiment show, as expected, that the classification performance on each collection decreases when another dataset is used during the training step. In addition, they further reveal that such performance decrease is not constant but depends on the specific training collection. Of course, among the most relevant uses it is obvious mentioning CCTV video surveillance. It can benefit from our proposal and, in general, from any further achievements in the field by relieving the operators from the need of a tiring continuous attention. But there are other possible uses that are becoming more and more desirable. For instance, TV parental control allows excluding violent or inappropriate content in advance, but the occurrence of these situations must be foreseen. On the contrary, a real-time detection would be even more beneficial even in sudden appearance of violent scenes. In summary, the achieved results of the presented extensive study show that, though the research is advancing, several open problems still call for further investigation about this challenging topic. This is especially engaging due to the many practical applications.

Acknowledgements This work is partially funded by the ULPGC under Project ULPGC2018-08, the Spanish Ministry of Economy and Competitiveness (MINECO) under Project RTI2018-093337-B-I00 and the Gobierno de Canarias and FEDER funds under Project ProID2020010024.

References

1. Pawlowski, P., Dabrowski, A., Balcerek, J., Konieczka, A., Piniarski, K.: Visualization techniques to support cctv operators of smart city services. Multimed. Tools Appl. **79**(29), 21095–21127 (2020)
2. Wang, J., Kankanhalli, M.S., Yan, W., Jain, R.: Experiential sampling for video surveillance. In: First ACM SIGMM International Workshop on Video Surveillance, pp. 77–86 (2003)
3. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) Computer Analysis of Images and Patterns, pp. 332–339 (2011)
4. Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S., Contardo, P., Lombardi, M., Dragoni, A.F.: A dataset for automatic violence detection in videos. Data Brief **33**, 106587 (2020)
5. Serrano, I., Deniz, O., Espinosa-Aranda, J.L., Bueno, G.: Fight recognition in video using hough forests and 2d convolutional neural network. IEEE Trans. Image Process. **27**(10), 4787–4797 (2018)
6. Zhang, T., Jia, W., He, X., Yang, J.: Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans. Circuits Syst. Video Technol. **27**(3), 696–709 (2017)
7. Serrano Gracia, I., Deniz Suarez, O., Bueno Garcia, G., Kim, T.-K.: Fast fight detection. PLOS ONE **10**, 1–19 (2015)
8. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2017)
9. Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., He, X.: A new method for violence detection in surveillance scenes. Multimed. Tools Appl. **75**(12), 7327–7349 (2016)
10. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: 3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
11. Li, X., Huo, Y., Xu, J., Jin, Q.: Detecting violence in video using subclasses. CoRR (2016). abs/1604.08088
12. Sjöberg, M., Baveye, Y., Wang, H., Quang, V., Ionescu, B., Dellandréa, E., Schedl, M., Demarty, C., Chen, L.: The mediaeval 2015 affective impact of movies task. In: Multimedia Benchmark Workshop, CEUR Workshop Proceedings, CEUR (2015)
13. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017)
14. Zhang, Z., Peng, H.: Deeper and wider SIAMese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4591–4600 (2019)
15. Huang, Y., Guo, Y., Gao, C.: Efficient parallel inflated 3d convolution architecture for action recognition. IEEE Access **8**, 45753–45765 (2020)
16. Ding, C., Fan, S., Zhu, M., Feng, W., Jia, B.: Violence detection in video by using 3d convolutional neural networks. In: Advances in Visual Computing, pp. 551–558. Springer (2014)
17. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497 (2015)
18. Meng, Z., Yuan, J., Li, Z.: Trajectory-pooled deep convolutional networks for violence detection in videos. In: Liu, M., Chen, H., Vincze, M. (eds.) Computer Vision Systems, pp. 437–447 (2017)
19. Aktı, Ş., Tataroğlu, G. A., Ekenel, H. K.: Vision-based fight detection from surveillance cameras. In: 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6 (2019)
20. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730–734 (2015)
21. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807 (2017)
22. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733 (2017)
23. Freire-Obregón, D., Castrillón-Santana, M., Barra, P., Bisogni, C., Nappi, M.: An attention recurrent model for human cooperation detection. Comput. Vis. Image Underst. **197–198**, 102991 (2020)
24. Sridhar, S., Mueller, F., Oulasvirta, A., Theobalt, C.: Fast and robust hand tracking using detection-guided optimization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3221 (2015)
25. Tian, Y., Pei, K., Jana, S., Ray, B.: Deeptest: automated testing of deep-neural-network-driven autonomous cars. In: 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), pp. 303–314 (2018)
26. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**(1), 35 (1960)
27. Reid, D.: An algorithm for tracking multiple targets. IEEE Trans. Autom. Control **24**(6), 843–854 (1979)
28. Fortmann, T., Bar-Shalom, Y., Scheffe, M.: Sonar tracking of multiple targets using joint probabilistic data association. IEEE J. Ocean. Eng. **8**(3), 173–184 (1983)
29. Ciaparrone, G., Luque Sánchez, F., Tabik, S., Troiano, L., Tagliaferri, R., Herrera, F.: Deep learning in video multi-object tracking: a survey. Neurocomputing **381**, 61–88 (2020)
30. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP) (2016)
31. Yang, M., Jia, Y.: Temporal dynamic appearance modeling for online multi-person tracking. Comput. Vis. Image Underst. **153**, 16–28 (2016)
32. Xiang, Y., Alahi, A., Savarese, S.: Learning to track: online multi-object tracking by decision making. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4705–4713 (2015)
33. Host, K., Ivašić-Kos, M., Pobar, M.: Tracking handball players with the DeepSORT algorithm. In: 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 1, 593–599 (2020)
34. Huang, B., Xu, T., Jiang, S., Chen, Y., Bai, Y.: Robust visual tracking via constrained multi-kernel correlation filters. IEEE Trans. Multimed. **22**(11), 2820–2832 (2020)
35. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4305–4314 (2015)

36. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L. V.: Temporal segment networks: towards good practices for deep action recognition. CoRR, abs/1608.00859 (2016)
37. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. CoRR, abs/1406.2199 (2014)
38. Szegedy, C., Liu, Wei., Jia, Yangqing, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015)
39. Soomro, K., Zamir, A. R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402 (2012)
40. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision, pp. 2556–2563 (2011)
41. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. CoRR, abs/1705.06950 (2017)
42. Kaminski, B., Jakubczyk, M., Szufel, P.: A framework for sensitivity analysis of decision trees. Cent. Eur. J. Oper. Res. **26**, 03 (2018)
43. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
44. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. CoRR, abs/1603.02754 (2016)
45. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
46. Tolles, J., Meurer, W.J.: Logistic regression: relating patient characteristics to outcomes. JAMA **316**(5), 533–534 (2016)
47. Bilinski, P., Bremond, F.: Human violence recognition and detection in surveillance video. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 30–36 (2016)
48. Deb, T., Arman, A., Firoze, A.: Machine cognition of violence in videos using novel outlier-resistant VLAD. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 989–994 (2018)
49. Song, W., Zhang, D., Zhao, X., Yu, J., Zheng, R., Wang, A.: A novel violent video detection scheme based on modified 3d convolutional neural networks. IEEE Access **7**, 39172–39179 (2019)
50. Adhikari, B., Peltomaki, J., Puura, J., Huttunen, H.: Faster bounding box annotation for object detection in indoor scenes. In: 2018 7th European Workshop on Visual Information Processing (EUVIP), pp. 1–6 (2018)
51. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) (2020)
52. Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W.: Violence detection using spatiotemporal features with 3d convolutional neural network. Sensors **19**(11),(2019)
53. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. CVPR **2011**, 1521–1528 (2011)

David Freire-Obregón received the M.Sc. and Ph.D. degrees in Computer Science from Las Palmas de Gran Canaria University (ULPGC) in 2010 and 2014, respectively. His research activities focus mainly on biometrics and digital forensics problems and cover different topics related to machine learning, generative models, image processing, and computer graphics. Currently, he is an associate professor at ULPGC.

Paola Barra received the B.S. degree in computer science from University of Salerno and the M.S. degree in Business Informatics from University of Pisa. She then obtained her PhD from the Universisty of Salerno. Her research interests include machine learning techniques in facial and gait recognition, image processing and video games development. She is member of IEEE and GIRPR/IAPR.

Modesto Castrillón-Santana received the M.Sc. and Ph.D. degrees in computer science from the Las Palmas de Gran Canaria University (ULPGC), in 1992 and 2003, respectively. He is currently a Full Professor with the Department of Computer Science and Systems, ULPGC. His main research activities focus particularly on the automatic facial analysis, covering also different topics related to image processing, perceptual interaction, human-machine interaction, biometrics, and computer graphics. He is a member of the AEPIA and AERFAIIAPR, having co-authored more than 150 articles, including peer-reviewed international journals, book chapters, and conference proceedings. He has acted as an external expert for the Chilean, Italian, and Qatar Research Agencies and the Spanish Accreditation Agency. He serves to the community in different conference programmes and technical committees. He is currently an Associate Editor of Pattern Recognition Letters and the IEEE Biometrics Newsletter.

Maria De Marsico is an Associate Professor (with qualification as Full Professor) at Computer Science Department of Sapienza University of Rome. Her research interest focuses on image and signal processing, especially biometric techniques, and on human-computer interaction, in particular, multimodal interaction. She has published about 190 papers on international conferences and high-ranked journals. She has been Associate Editor for Pattern Recognition Letters since 2015, has been Area Editor for the same journal from January 2018, and is the Editor in Chief for Special issues since 2019. She is Associate Editor for Pattern Recognition. She has been the Area Editor for IEEE Biometrics Compendium from 2012 up to its termination in 2019, member of the Editorial Board for IEEE Biometrics Newsletter since 2015 and Editor in Chief since 2018. She is Program co-chair for the International Conference on Pattern Recognition Applications and Methods since 2013. She has been guest co-editor of several special issues for high-ranked journals. She is Senior member of IEEE, ACM, CVPL (formerly GIRPR - Gruppo Italiano dei Ricercatori in Pattern Recognition-national chapter of IAPR), EAB (European Association for Biometrics) and INSTICC (Institute for Systems and Technologies of Information, Control and Communication).