Soft Computing (2020) 24:11007–11017
https://doi.org/10.1007/s00500-019-04310-x

FOCUS

A soft computing approach to violence detection in social media for smart cities

Francisco A. Pujol[1] · Higinio Mora[1] · Maria Luisa Pertegal[2]

Published online: 11 September 2019

Abstract
In recent years, social media has become an everyday tool for the distribution of videos in which signs of violence appear in different ways. Citizens of smart cities are demanding increasing efforts to authorities in order to maintain public safety, as well as to be efficient in an emergency response. The complexity of monitoring automatically the enormous amount of information generated through social networks results in the need for the development of systems that allow for the automatic detection of violent content in videos. This fact is becoming increasingly important in order to guarantee security for the citizens in any smart city. As a result, this work proposes the development of a system for detecting violence in videos by combining different descriptors that calculate the acceleration produced between two frames of a video. To do this, different techniques, such as the Radon transform or optical flow, are used. The trained system then performs the classification using support vector Machines. The results are promising, with accuracy rates between 85 and 97%, depending on the complexity of the databases used, which demonstrates the validity of our proposal.

Keywords Violence detection · Smart cities · Social media · Radon transform · Optical acceleration

1 Introduction

With the growth and increasing popularity of social media, the presence of users on the Internet has become almost permanent. Through social media, users express feelings, opinions or emotions that are commented or discussed in public or private conversations. In recent years, there is an emergence of new services that allow people to stream live videos of any kind of event. This trend has seen the rise of the publication of a great amount of enjoyable videos, but videos of brutal aggressions to disabled people or terrorist attacks, among others, have been also posted and even broadcast live (Studer 2017).

Facebook, Twitter, YouTube and other social networks try to block or remove publications that encourage violence. They use automated programs to detect publications that violate conditions of use, such as those that show or promote violence. However, it is a challenging task, even harder when live streaming plays a key role to expand violent activities.

An automated violence detection system has immediate applicability in smart cities to monitor citizens welfare (Visvizi et al. 2018; Visvizi and Lytras 2018). Cameras installed in train, metro or bus stations, sports stadiums, airports, prisons, schools or on the streets help alert authorities of potentially dangerous situations for the integrity of the citizens. In recent years, acts of terrorism in big cities or acts of vandalism and fights have generated a need for the use of cameras to monitor people, causing, as a consequence, an exponential increase in security measures in many cities around the world. For this reason, the detection of violence requires specific systems that act in the fastest and most efficient way.

At the same time, there is a growing demand for automatic classification systems for videos uploaded to the Internet.

Communicated by Miltiadis D. Lytras.

✉ Francisco A. Pujol
fpujol@ua.es

Higinio Mora
hmora@ua.es

Maria Luisa Pertegal
ml.pertegal@ua.es

[1] Department of Computer Technology, University of Alicante, Alicante, Spain

[2] Department of Developmental and Educational Psychology, University of Alicante, Alicante, Spain

As mentioned before, these systems will help to combat the harmful effects of videos showing violent contents, such as recordings of violent scenes on football matches or riots in the street. The implementation of real-time detection systems in case of possible violent actions is a very useful method to control public security.

Therefore, it would be convenient for a smart city governance to implement some kind of measures to monitor and potentially detect violence on videos uploaded or streamed in social networks (Lytras and Visvizi 2018). However, using a human control of all the information on a social network is unfeasible due to the huge amount of information that can be produced at any given time. Therefore, the main objective of this work is to propose and develop an automated violence detection system on social media by means of the analysis of suspicious images and videos. To do this, as violent movements can be typically produced by an increasing acceleration of the attackants' bodies, our proposal is based on the estimation of the acceleration between frames in a video. So the research question for this paper is: Is the acceleration between frames a reliable cue to address the detection of violence in videos? To answer this question, the main contributions of this paper are:

- A new method to calculate the acceleration between frames is proposed. Since the Fourier transform of frames with accelerated movements produces an ellipse whose direction is orthogonal to the direction of the motion, this acceleration is estimated by dividing each frame into blocks and calculating the local eccentricity of the ellipses that appear in each block. Radon transform is used to calculate the eccentricity.
- Two recent descriptors, the histogram of optical acceleration (HOA) and the histogram of spatial gradient of acceleration (HSGA) (Edison and Jiji 2017), are adapted and utilized for the first time to detect violence in videos. Both descriptors make use of optical flow frames.
- The combination of the three methods, local eccentricity, HOA and HSGA, with the popular soft computing method of support vector machines to classify input videos, achieves very good detection rates and proves to be a suitable technique for detecting violence in videos from social media.

This paper is organized as follows: Sect. 2 summarizes some related works; Sect. 3 explains our proposal to detect violence in videos, introducing both the calculation of the local eccentricity of the frames in a video and the two descriptors, HOA and HSGA; Sect. 4 describes the experimental setup and the set of experiments completed with different databases; and finally, conclusions and some future works are discussed in Sect. 5.

2 Related works

Automatic surveillance systems, based on audio and video recordings, are systems that, in addition to being capable of detecting situations of risk, prevent possible situations of violence thanks to their persuasive effect. Thus, some studies show that the presence of Closed Circuits of TV (CCTV) reduced security interventions by 3% while security controls increased by 11% (Sivarajasingam et al. 2003).

With the growth of surveillance systems and the development of image analysis methods, investigations are arising aimed at automatically detecting possible risk situations. At first, heuristic methods were used based on the position and acceleration of the movement or acceleration measure vector (AMV) (Datta et al. 2002) that allowed to detect anomalous movements. However, situations of crime or violence can have different forms of behavior, so it is necessary to start from a broad database that can help in the recognition of this type of actions. The space–time interest points (STIP) and motion scale-invariant feature transform (MoSIFT) systems (Bermejo Nievas et al. 2011) precisely allowed and characterized these scenes with 90% success, despite being in restricted contexts. Violence is frequently linked to rapid movements and accelerations that demand a high computational effort from an automatic system. Deniz et al. (2014) propose an efficient method using Radon transform that allows characterizing and detecting extreme acceleration patterns, achieving speedup rates of up to 15×.

The detection of violence or risky situations in crowds follows completely different patterns from those that may arise in small groups. In Manfredi et al. (2014), this type of groupings is characterized where flows of people are identified. These people are monitored by means of a one-class support vector machine. Computational intelligence has also been used for smart environments, as in Fahmi et al. (2017), Amin et al. (2018) and Amin et al. (2019).

Maturity in recognition and vision techniques enabled the creation of broad-spectrum security systems such as the rotation-invariant feature modeling motion coherence (RIMOC) (Ribeiro et al. 2016). This system is based on inference methods that are applied at a multi-scale level and makes it possible to recognize the erratic movements that aggressive movements usually involve.

Artificial intelligence has been present in automatic violence detection systems. The first approaches were based on defining a vector of spatiotemporal features that would enable us to discern violent actions from those that were not (De Souza et al. 2010). The new researches incorporate more features such as audio in order to be able to discern among other situations that do not have such a clear violent pattern; such is the case of situations of harassment in CCTV (Gao et al. 2016; Sidhu and Sharad 2016).

The monitoring of various camera systems with the aim of detecting situations of violence in large areas has been a challenge that has been addressed from the techniques of Big Data and Data Mining. Thus, Shao et al. (2018) propose a system that is capable of detecting this type of situation from the monitoring of detection events and alarm messages from intelligent cameras. In the study by Dorogyy et al. (2018), the detection goes further: Verbal and nonverbal behaviors are extrapolated so that it is possible to detect the violent situation some time before it occurs.

Violence is also present in cyberspace, and social networks provide a context that can foster such situations because of the anonymity they provide. In fact, several studies have been carried out to analyze YouTube videos or Twitter messages that seek to limit this type of violence (Giannakopoulos et al. 2010; Osborne et al. 2014).

Violence in cities is one of the main problems facing our society. In the modern environment of intelligent cities, there is a need to develop a system capable of detecting whether a violent situation occurs or not. The large amount of data available in these cities from a large number of information sources, together with the variety of situations to be controlled, require an agile cyber-infrastructure, oriented to Big Data and Deep Learning that allow an efficient and effective use for all areas of smart cities, including security (Shams et al. 2018). In Bautista-Duran et al. (2017) and García-Gómez et al. (2016), systems are presented that allow acoustic violence detection for intelligent cities, integrating both signal processing and pattern recognition techniques. The new techniques, in addition to audio, incorporate video, and a centralized analysis in the Cloud, an interesting compilation of the main researches for the detection of violence in the context of smart cities, is presented in Srivastava et al. (2017).

3 Methodology

In this section, the method followed to determine whether there is a violent behavior in a video is described. As mentioned before, different studies about human behavior and how humans perceive actions performed by others have shown that kinematic patterns of movement are a suitable indicator for the perception of actions.

Therefore, aspects such as acceleration, suddenness or excitement are often associated with states of high activation (happiness, anger), whereas soft or slow movements are considered as states of low activation (depression, sadness). This way, it can be assumed that violence in videos can be detected thanks to these kinematic features that represent accelerated movements, such as blows, punches or kicks. Knowing that high acceleration plays a key role for violence detection, our method will calculate acceleration in different ways to verify violent actions. Figure 1 shows how our system works.

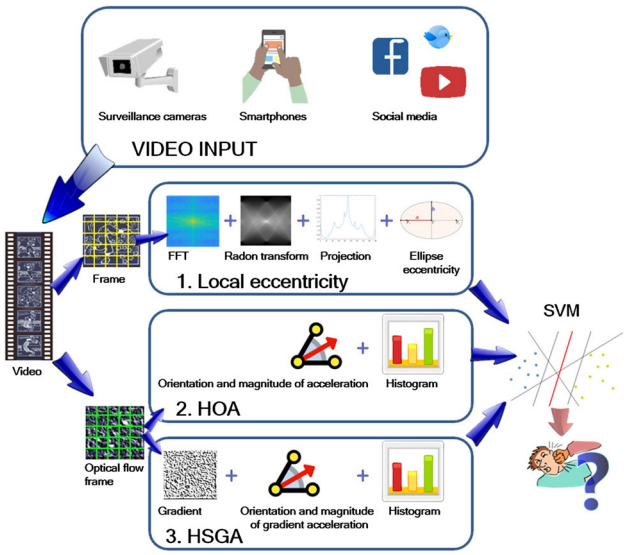

Fig. 1 Proposed violence detection process

Two different approaches are used in order to estimate acceleration. Firstly, the system will compare any two consecutive frames and will calculate the eccentricity of the ellipse that appears in the fast Fourier transform when some blur from motions is found between frames. This eccentricity is related to the acceleration.

Secondly, optical flows are extremely popular as input for human actions recognition. As a consequence, optical flow frames are utilized to compute the so-called optical acceleration by using two recent descriptors, the histogram of optical acceleration (HOA) and the histogram of spatial gradient of acceleration (HSGA).

Finally, the system will classify the video as violent or nonviolent by using a trained support vector machine (SVM).

Let us point out that in any case it is assumed that any motion blur in a video will come exclusively from the sequence of actions that happen inside the video and not from any blur introduced by the camera from where the video was recorded.

3.1 Acceleration and eccentricity of an ellipse

Let us consider a video sequence captured from any of the sources shown in Fig. 1. Let us assume in all cases that any blur in the video comes from the scene being filmed only; that is, the source camera does not introduce any blur. According to Deniz et al. (2014), for each pair of consecutive frames in the video, the power spectrum of both frames will be obtained using the fast Fourier transform (FFT). When there is a sudden accelerated movement between the two frames, the second frame will show a blur in the area of the image with movement (see Fig. 2). This blur will be more or less noticeable depending on how abrupt the movement has been.

Fig. 2 Two consecutive frames of a video from an action movie, where there appears a blur on the arms and leg of the attacking person

Consequently, it can be detected more easily when it involves aggressive movements.

When this happens, the spectrum of the second frame will show an ellipse, whose orientation will be perpendicular to the direction of motion, as shown in Fig. 3. In this figure, there is a vertical movement with different accelerations. The resulting ellipse is depicted in each case.

Let $I_{t-1}(x, y)$ and $I_t(x, y)$ be two consecutive frames of a video, the blur caused by movement is equivalent to applying a convolution with a low-pass filter $h(x, y)$:

$$I_t(x, y) = I_{t-1}(x, y) * h(x, y) \tag{1}$$

Let $F(\cdot)$ indicate the Fourier transform, then the equation can be written as:

$$F(I_t) = F(I_{t-1}) \cdot F(h) \tag{2}$$

Here, $F(h)$ refers to the ellipse that appears in Fig. 3d, f. So, if two frames have slow movements between them or are static, their Fourier transform will be almost the same and, consequently, no ellipse will appear. The more abrupt a movement is, the closer to 1 the eccentricity of the ellipse will be, as shown in Fig. 3d, f. Therefore, the calculation of this eccentricity e will give an approximation of the acceleration produced between two consecutive frames and it is an estimator of the presence of a potential aggressive behavior. A definition of eccentricity e is (Bennett et al. 1999):

$$e = \sqrt{\frac{a^2 - b^2}{a^2}} \tag{3}$$

where a and b are the major and minor axis of the ellipse, respectively. As known, eccentricity has values between 0 and 1 for ellipses.

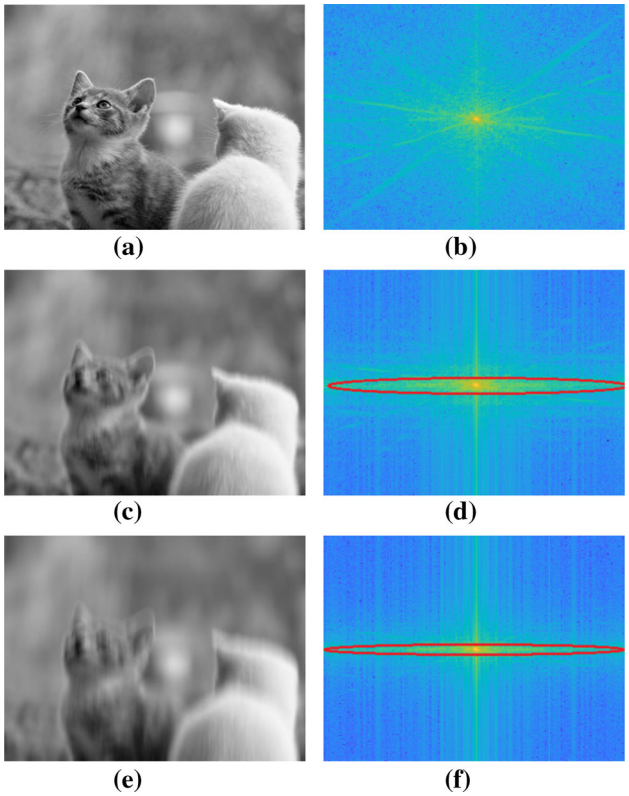

Fig. 3 Power spectrum density (PSD) using the FFT for different versions of an image. **a** Original image; **b** PSD of the original image; **c** image with a vertical motion of 20 pixels; **d** corresponding PSD for image **c**; **e** image with a vertical motion of 50 pixels; **f** corresponding PSD for image **e**

In order to calculate eccentricity, Radon transform will be used. The mathematician Joseph Radon developed in 1917 an analytical method, based on Radon's inverse transformation, of reconstructing the image of an object by its projections. It is an integral transformation consisting of the integral of a function on a set of straight lines. This transform indicates that the image of an object is precisely and unequivocally determined by the infinite set of all its projections. Radon transform is having an increasing interest in image processing, since it is extremely robust to noise and presents scale and rotation invariance (Hoang and Tabbone 2012; Tizhoosh 2015; Tizhoosh and Rahnamayan 2016). This way, it has been used as a powerful descriptor for image retrieval, especially in medical image applications. Essentially, Radon transform consists of an integral transform which projects all pixels from different orientations to a single vector.

The Radon transform R of an image I_t at pixel (x, y) is defined as the line integral along an inclined line with angle θ from the x-axis and at a distance ρ from the origin (Hoang and Tabbone 2012):

$$R(\rho,\theta)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_t(x,y)\delta(x\cos\theta+y\sin\theta-\rho)\mathrm{d}x\mathrm{d}y, \tag{4}$$

where δ is the Dirac delta function and ρ, θ are the position and the orientation of the Radon projection vector, respectively.

Therefore, once the Radon transform has been calculated, a matrix whose number of columns equals the number of angles θ_i, for $i = 0, 1, 2, \ldots, 179$, will be generated. After this, a row vector with the maximum value of each column will be created and, then, the maximum value of this vector will be calculated. This vector represents the maximum vertical projection vector v_p, which is normalized to 1. When an ellipse appears, this vector will show a peak, representing the major axis of the ellipse (see Fig. 4).

Knowing the position $p_{\max}$ of the maximum value in the projection vector v_p, which is related to the major axis of the ellipse, the minor axis of the ellipse will be located at position $p_{\min}$:

$$p_{\min}=\begin{cases} p_{\max}+90, & \text{if } (p_{\max}+90)\leq 180 \\ p_{\max}-90, & \text{otherwise} \end{cases} \tag{5}$$

The method described in Deniz et al. (2014) follows the steps described above. That is, eccentricity is calculated globally, using the complete frame. However, if the blur that may describe an acceleration related to violence is located in a small area of the frame, this will not be taken into account in the global calculation, resulting in a false negative for the violence detector. Our method tries to overcome this problem by dividing each frame into blocks of $n \times n$ pixels and performing all the steps mentioned before in each block, i.e., locally. Our algorithm can be described like this:

1. Divide frame I_t into blocks of size $n \times n$ pixels.
2. For each block j:
   (a) Calculate the power spectrum density (PSD) using the FFT.
   (b) Compute the Radon transform $R_j(\rho, \theta)$.
   (c) Find the positions $(p_{\max_j}, p_{\min_j})$ of the major and minor axis of the corresponding ellipse.
   (d) Calculate the eccentricity e_{b_j} for block j.
3. The eccentricity of frame I_t will be $e_t = \max_j(e_{b_j})$.

The calculation of the local eccentricity of each block will help to accurately identify violence, since local accelerations may be dismissed if eccentricity is computed globally over the whole image. The greater the eccentricity between frames, the greater the blur between them and, in other terms, the greater the acceleration produced. The key point from this calculation is to find a threshold for the eccentricity to determine whether there has been any violence in the video. This will be discussed in Sect. 4.

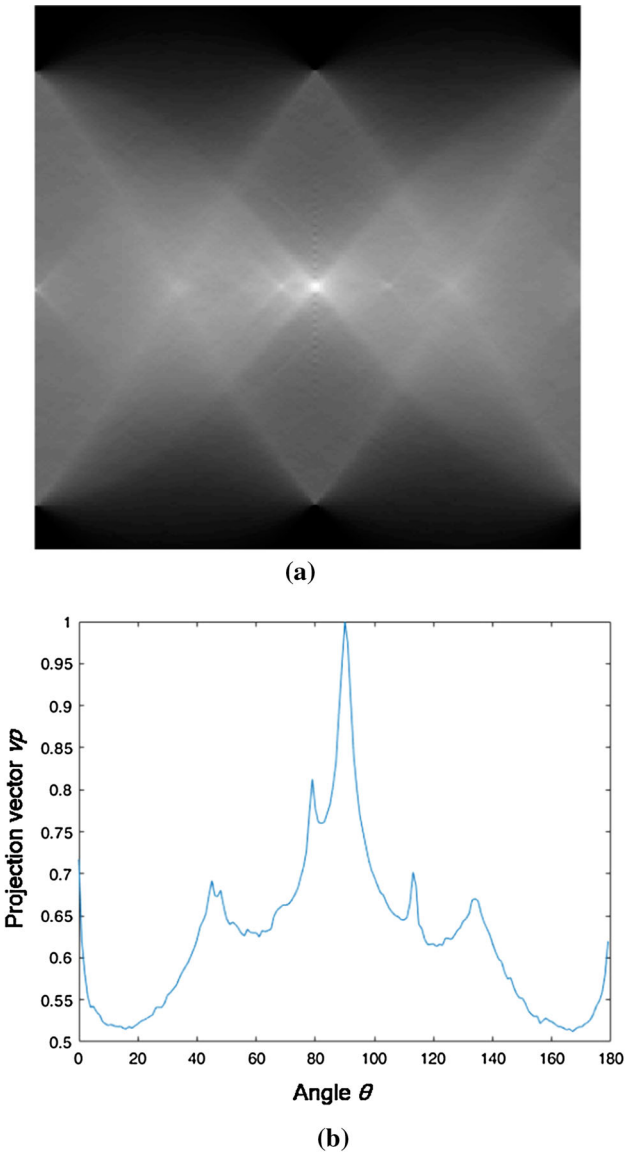

Fig. 4 **a** Radon transform of Fig. 3c; **b** maximum vertical projection normalized. The peak represents the major axis of the ellipse

3.2 Computing optical acceleration

Recently, Edison and Jiji (2017) have developed a new set of descriptors to calculate optical acceleration, the so-called histogram of optical acceleration (HOA) and histogram of spatial gradient of acceleration (HSGA). The authors define optical acceleration as the rate of change of optical flow that shows a perceptible acceleration of each pixel in a frame. Both descriptors make use of optical flow to compute the histograms.

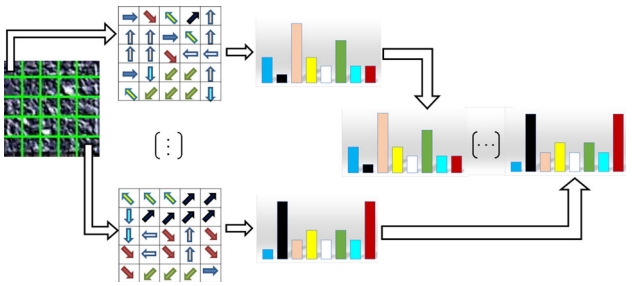

Fig. 5 Graphical representation of the process of extraction of features for HOA. Image is divided into blocks. After computing the orientation of all the pixels in the block, the histogram of the orientations of the optical acceleration for each block is obtained. Finally, all the orientation histograms are concatenated to obtain the final HOA descriptor

Let F_t and F_{t-1} be two consecutive optical flow frames. The difference d_t between them can be estimated as (Zhou et al. 2017):

$$d_t = F_t - F_{t-1} \tag{6}$$

Therefore, the instant acceleration a_t will be:

$$a_t = \begin{cases} 0, & \text{if } (d_t + 128) < 0 \\ d_t + 128, & \text{if } 0 \leq d_t + 128 \leq 255 \\ 255, & \text{if } d_t + 128 > 255 \end{cases} \tag{7}$$

Let us consider that $\{a_{x_t}(x, y), a_{y_t}(x, y)\}$ are, respectively, the horizontal and vertical optical acceleration components for pixel (x, y) in an optical flow frame at time t in the video that have been calculated from Eq. (7). The magnitude O_t and orientation θ_t for the acceleration in that pixel can be calculated as:

$$O_t = \sqrt{|a_{x_t}(x, y)|^2 + |a_{y_t}(x, y)|^2} \tag{8}$$

$$\theta_t = \arctan\left(\frac{a_{y_t}(x, y)}{a_{x_t}(x, y)}\right) \tag{9}$$

In order to calculate the histogram of optical acceleration (HOA), optical flow frames will be divided into blocks of $m \times m$ pixels. The orientation of optical acceleration θ_t will be calculated for each block, and a histogram of its magnitude O_t will be computed, considering eight bins. The histograms are then concatenated to obtain the final HOA descriptor, as shown graphically in Fig. 5.

In recent years, several works have appeared considering that the estimation of motion on object boundaries is an important feature for optical flow applications in action recognition (Weinzaepfel et al. 2015; Fu et al. 2016; Rezazadegan et al. 2017). As a result, computing acceleration on object boundaries can be a reliable feature to find violent behaviors in videos. To do this, the histogram of spatial gradient of acceleration (HSGA) is defined in Eqs. (10) and (11), where A_{x_t} and A_{y_t} are the horizontal and vertical components of the spatial gradient of acceleration, and a Sobel operator is applied for the computation of the gradient.

$$A_{x_t} = \begin{bmatrix} a_{x_t}(x-1, y-1) & a_{x_t}(x, y-1) & a_{x_t}(x+1, y-1) \\ a_{x_t}(x-1, y) & a_{x_t}(x, y) & a_{x_t}(x+1, y) \\ a_{x_t}(x-1, y+1) & a_{x_t}(x, y+1) & a_{x_t}(x+1, y+1) \end{bmatrix} \times \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{10}$$

$$A_{y_t} = \begin{bmatrix} a_{y_t}(x-1, y-1) & a_{y_t}(x, y-1) & a_{y_t}(x+1, y-1) \\ a_{y_t}(x-1, y) & a_{y_t}(x, y) & a_{y_t}(x+1, y) \\ a_{y_t}(x-1, y+1) & a_{y_t}(x, y+1) & a_{y_t}(x+1, y+1) \end{bmatrix} \times \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{11}$$

As for the case of the HOA, optical flow frames will be divided in blocks of $m \times m$ pixels. The orientation of both horizontal and vertical gradients will be calculated for each block, and a histogram of its magnitude will be computed, considering eight bins. The histograms are then concatenated to obtain the final HSGA descriptor in a similar way as shown in Fig. 5.

4 Experiments

4.1 Experimental setup

In this section, some experiments to validate our proposal will be carried out. Three databases have been used to perform the tests.

- The SBU Kinect Interaction dataset (Yun et al. 2012), which consists of 54 video clips, 16 of which have violent content and the remaining 38 do not have high acceleration movements and could therefore be considered as nonviolent. This database contains both violent (punching, kicking, pushing) and nonviolent (walking, hugging, shaking hands) videos of two people in one room, with the camera recording frontally. Some examples are shown in Fig. 6.
- The UT-Interaction dataset (Ryoo and Aggarwal 2010). This database includes videos taken with a surveillance camera located at a certain distance. These videos contain different human actions in two realistic environments. The first one is a parking lot with mostly static background. The second one is a garden with a slight movement in the background due to the wind. In both cases, one can find both violent and nonviolent actions, with six interactions: hugging, handshaking, pointing,

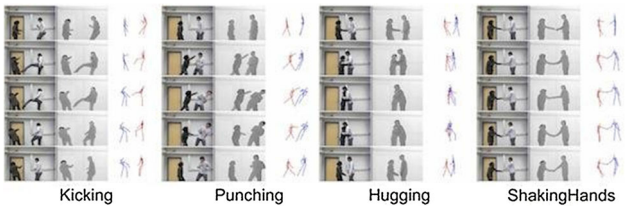

Fig. 6 Examples of some movements of videos from the SBU Kinect Interaction dataset

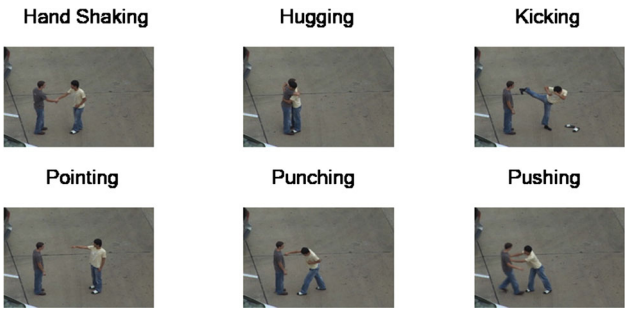

Fig. 7 Different actions from videos located in the parking lot from the UT-Interaction dataset

Fig. 8 Violence (first row)/nonviolence (second row) examples from the Violent-Flows database

pushing, punching and kicking. There are 60 videos in total, so that each interaction is represented using ten videos. See Fig. 7 for some examples of interactions of this dataset.

– The Violent-Flows—Crowd Violence/Nonviolence database (Hassner et al. 2012). This is a real-world, challenging crowd violence dataset collected from YouTube. There are 246 videos in the database, with 123 clips of crowd violence events and 123 clips of crowd nonviolent actions; see Fig. 8 for some sample frames from this database.

All the videos have been resized to 160 × 120 pixels. For the training process, a total of 40 videos have been chosen, where 20 of them contain violent actions and the remaining 20 include nonviolent actions. In particular, ten videos from both the SBU Kinect Interaction and UT-Interaction datasets and 20 videos from the Violent-Flows database have been selected. In order to reduce the computation time of the whole process, five frames per second are randomly chosen for each video.

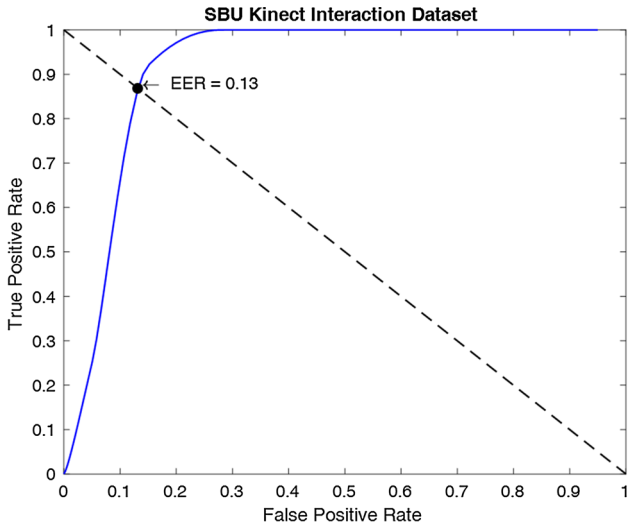

Fig. 9 ROC curve for the local eccentricity method with the SBU Kinect Interaction dataset

The process to test our method will be the following. First, for each dataset, the value of the eccentricity to determine when the action involved in the video is violent or nonviolent will be computed. To do this, the local eccentricity of each pair of consecutive frames as described in Sect. 3.1 is calculated and, then, the maximum eccentricity value is computed for the whole video. After computing the false positive rate (FPR) and false negative rate (FNR) for each training set obtained with this maximum, a threshold for the eccentricity value will be selected. This threshold will be then applied to the test set for each database.

In addition, both HOA and HSGA will be computed for each video in the training set and concatenated to obtain the feature vector. Once obtained, a support vector machine (SVM) with a linear kernel has been chosen to classify the video as violent/nonviolent (Chang and Lin 2011).

4.2 Results for the SBU Kinect Interaction Dataset

The ROC curve to determine the threshold for the eccentricity is shown in Fig. 9. The equal error rate (EER) value from the ROC curves has been chosen to calculate the threshold, $\mathrm{Th_e}$. That is, the eccentricity value that obtains the EER value is considered as the threshold for the rest of experiments. The black dashed diagonal line in the ROC curve represents the EER line. For this database, $\mathrm{Th_e} = 0.7734$, which gives an accuracy rate of 87.02%.

Table 1 summarizes the results on the recognition accuracy for this dataset with the three proposed methods, as well as combinations of them. We also include the global eccentricity results.

Table 1 Results for the SBU Kinect Interaction dataset

Method	Accuracy (%)
Global eccentricity	81.35
Local eccentricity (LE)	87.02
HOA	90.52
HSGA	92.47
LE + HOA	92.35
LE + HSGA	94.68
HOA + HSGA	95.56
LE + HOA + HSGA	97.85

Table 2 Results for the UT-Interaction dataset

Method	Accuracy (%)
Global eccentricity	74.21
Local eccentricity (LE)	76.35
HOA	80.82
HSGA	82.02
LE + HOA	81.14
LE + HSGA	84.03
HOA + HSGA	84.78
LE + HOA + HSGA	89.68

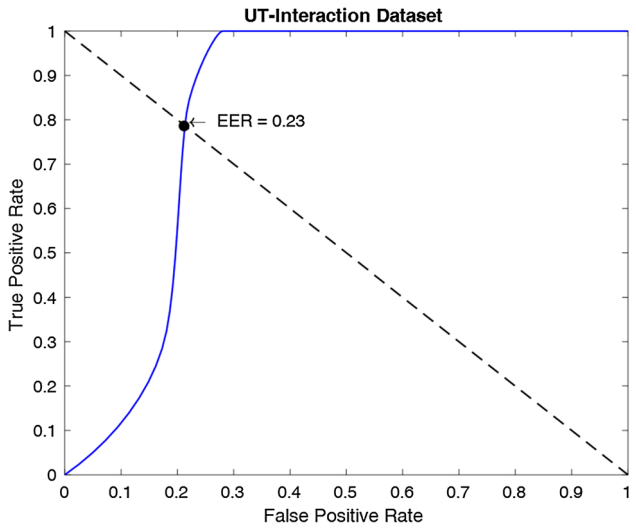

Fig. 10 ROC curve for the local eccentricity method with the UT-Interaction dataset

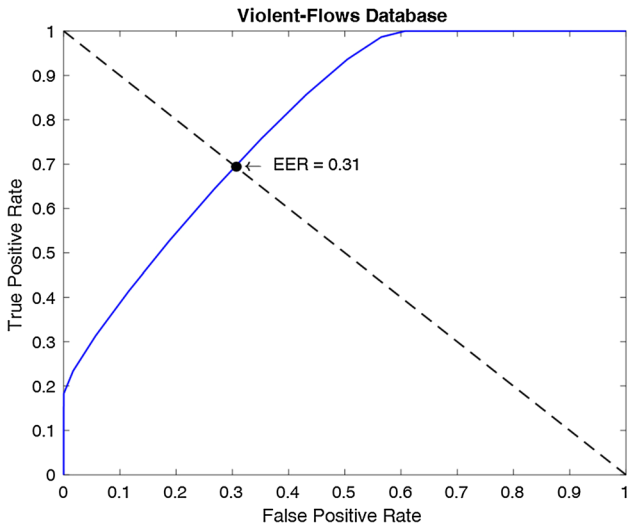

Fig. 11 ROC curve for the local eccentricity method with the Violent-Flows database

From these results, it can be seen that all the methods proposed have high accuracy rates for this database. The local eccentricity calculation gives better accuracy than the original global eccentricity, as expected. On the other hand, LE by itself does not achieve a good recognition rate compared to HOA and HSGA, but its combination with the other two methods improves their rates, and the combination of all the three algorithms increases the accuracy of using only one method by 5–10%.

4.3 Results for the UT-Interaction Dataset

The same experimental procedure has been followed for the UT-Interaction dataset. The ROC curve to determine the threshold for the eccentricity can be seen in Fig. 10. The equal error rate (EER) value has been selected again in order to estimate the threshold, $\mathrm{Th_e}$. In this case, $\mathrm{Th_e} = 0.8117$, which gives an accuracy rate of 76.35%.

The results obtained for this dataset are shown in Table 2. Again, the combination of the three methods gives the best results.

Table 3 Results for the UT-Interaction dataset

Method	Accuracy (%)
Global eccentricity	64.71
Local eccentricity (LE)	69.48
HOA	70.66
HSGA	75.39
LE + HOA	73.43
LE + HSGA	77.12
HOA + HSGA	77.50
LE + HOA + HSGA	84.35

4.4 Results for the Violent-Flows Database

Figure 11 shows the ROC curve for the local eccentricity method. From the equal error rate (EER) value, the threshold $\mathrm{Th_e} = 0.8368$, which gives an accuracy rate of 69.48%.

The results of the experiments completed in this database are shown in Table 3.

From these results, it becomes clear that the combination of the three methods gives the better results on the recognition of violence in all cases, whereas taking into account only the eccentricity or the optical acceleration does not result in a reliable accuracy rate.

4.5 Comparison with other methods

After having found the optimal combination of our three algorithms to achieve a reliable accuracy, a set of experiments to compare our method with other state-of-the-art algorithms for violence recognition have been completed. Most of the selected algorithms make use of similar descriptors as the ones presented in this work.

In particular, many of the works that make use of the SBU Kinect dataset are based on the positions of human skeletons, such as category-blind human action recognition method (CHARM) (Li et al. 2015) or global context-aware attention LSTM networks (GCA-LSTM) (Liu et al. 2018). We also include the Relation History Image (RHI) descriptor, which simplifies the classification step by extracting multiple activities (Gori et al. 2016).

For the UT-Interaction dataset, the combination of histogram of gradients (HOG), histogram of optical flow (HOF) and motion boundary histograms (MBH) is used in Rota et al. (2015), whereas Fisher vectors (FV) are the base for the method described in Kantorov and Laptev (2014). Finally, in Yuan et al. (2012) the spatiotemporal context for activity recognition is defined.

Regarding the Violent-Flows database, optical flow and, in particular, a novel descriptor, the orientation histogram of optical flow (OHOF) is used in Zhang et al. (2016). Similarly, Gao et al. (2016) introduced the orientation information of optical flow which they called Oriented VIolent Flows (OVIF). And optical flow is taken into account again in the simplified histogram of oriented tracklets (sHOT) algorithm (Rabiee et al. 2018).

The results for all the algorithms involved are shown in Table 4. In all cases, the mean accuracy is presented.

After having implemented the experiments, let us comment some remarks. First, results show that the best performance belongs to our approach, outperforming all the other methods in the selected databases. Thus, the best results for our system come from the SBU Kinect dataset, where videos are recorded in highly constrained situations. On the contrary, the challenging Violent-Flows database, where there are real-life videos with crowds of people, achieves the worst results, although the accuracy obtained is still above 84%. The related works used for the comparison, which in most cases make use of descriptors based on optical flow, obtain comparable results, but in all cases they are 2–3% below the accuracy of our approach. As a consequence, we consider that the developed violence detector is a reliable detector and it constitutes a valid alternative to detect violence in videos uploaded to social media. Thus, it can make a successful contribution to enhance security in a near future in Smart cities.

Table 4 Comparison of the accuracy (%) of our proposal (LE + HOA + HSGA) with other state-of-the-art methods

SBU dataset	
CHARM (Li et al. 2015)	83.9
RHI (Gori et al. 2016)	93.08
GCA-LSTM (Liu et al. 2018)	94.9
Our proposal	97.85
UT dataset	
STCK (Yuan et al. 2012)	82.5
FV (Kantorov and Laptev 2014)	87.6
HOG + HOF + MBH (Rota et al. 2015)	83.54
Our proposal	89.68
Violent-Flows database	
OViF + SVM (Gao et al. 2016)	76.80
OHOF (Zhang et al. 2016)	82.79
sHOT (Rabiee et al. 2018)	82.2
Our proposal	84.35

5 Conclusions

Technology is naturally integrated into the management of smart cities, becoming the most effective means of addressing their problems and needs. Technology is beneficial for many applications, such as urban mobility, water or waste management, among others. However, one of the highest priorities for any smart city is citizen safety. In this line, security systems are a basic element to be taken into account for any smart city. Current technologies allow citizens to give alarm signals in a much more exhaustive way. Sensors of all kinds and intelligent surveillance cameras can monitor people at potential risk at all times, reviewing trouble spots. In this line, our approach introduces a new way to monitor videos in social media, potentially helping to increase civil security in smart cities.

In this work, a new method to detect violence in videos using soft computing methods has been proposed. The proposed method provides promising results for video surveillance contexts, as well as for rather more complicated scenarios such as those with crowds of people. The results from the experiments give accuracies above 97% for the SBU Kinect database, where only two people appear and videos are recorded close to where actions are actually happening. Even in a much more challenging database, such as the Violent-Flows database, where one can find unconstrained environments and actions performed by crowds of people, the recognition rate is above 84 %. As a result, our proposal

may be applied to monitor automatically violent behaviors in smart cities using both video surveillance cameras and videos uploaded to social media.

Future works aim at applying the proposed algorithm to real-time videos, such as the ones streamed live in social networks. We are currently adapting the method to work with GPUs and parallelizing the most time-consuming steps in the algorithm, such as computing the Radon transform.

Another improvement under development is trying to minimize the blur caused by global motion, such as camera motion, which can also cause blurring in the image and cause the local blurring produced by the movement of a person to be clearly distinguished.

To sum up, a smart city must provide its citizens with a protective comfort with the help of access control systems and civil security, until ensuring transparent navigation on social and collaborative platforms based on the Internet. As a conclusion, we think that in a near future our contribution may help to build new solutions to improve security in smart cities.

Acknowledgements This work has been partially supported by the Spanish Research Agency (AEI) and the European Regional Development Fund (FEDER) under project CloudDriver4Industry TIN2017-89266-R.

Compliance with ethical standards

Conflict of interest All authors state that there is no conflict of interest.

References

Amin F, Fahmi A, Abdullah S (2019) Dealer using a new trapezoidal cubic hesitant fuzzy TOPSIS method and application to group decision-making program. Soft Comput 23(14):5353–5366

Amin F, Fahmi A, Abdullah S, Ali A, Ahmad R, Ghani F (2018) Triangular cubic linguistic hesitant fuzzy aggregation operators and their application in group decision making. J Intell Fuzzy Syst 34(4):2401–2416

Bautista-Duran M, Garcia-Gomez J, Gil-Pita R, Mohino-Herranz I, Rosa-Zurera M (2017) Energy-efficient acoustic violence detector for smart cities. Int J Comput Intell Syst 10(1):1298–1305

Bennett N, Burridge R, Saito N (1999) A method to detect and characterize ellipses using the Hough transform. IEEE Trans Pattern Anal Mach Intell 21(7):652–657. https://doi.org/10.1109/34.777377

Bermejo Nievas E, Deniz Suarez O, Bueno García G, Sukthankar R (2011) 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, vol 6855. Springer, Berlin, pp 332–339

Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

Datta A, Shah M, Lobo NDV (2002) Person-on-person violence detection in video data. In: 16th international conference on pattern recognition, 2002. Proceedings, vol 1. IEEE, pp 433–438

De Souza FD, Chavez GC, do Valle EA Jr, Araújo AA (2010) Violence detection in video using spatio-temporal features. In: 2010 23rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 224–230

Deniz O, Serrano I, Bueno G, Kim T (2014) Fast violence detection in video. In: 2014 international conference on computer vision theory and applications (VISAPP), vol 2, pp 478–485

Dorogyy Y, Kolisnichenko V, Levchenko K (2018) Violent crime detection system. In: 2018 IEEE 13th international scientific and technical conference on computer sciences and information technologies (CSIT), vol 1. IEEE, pp 352–355

Edison A, Jiji CV (2017) Optical acceleration for motion description in videos. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 1642–1650

Fahmi A, Abdullah S, Amin F, Siddiqui N, Ali A (2017) Aggregation operators on triangular cubic fuzzy numbers and its application to multi-criteria decision making problems. J Intell Fuzzy Syst 33(6):3323–3337

Fu H, Wang C, Tao D, Black MJ (2016) Occlusion boundary detection via deep exploration of context. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 241–250

Gao Y, Liu H, Sun X, Wang C, Liu Y (2016) Violence detection using oriented violent flows. Image Vis Comput 48–49:37–41. https://doi.org/10.1016/j.imavis.2016.01.006

García-Gómez J, Bautista-Durán M, Gil-Pita R, Mohino-Herranz I, Rosa-Zurera M (2016) Violence detection in real environments for smart cities. In: Ubiquitous computing and ambient intelligence. Springer, Cham, pp 482–494

Giannakopoulos T, Pikrakis A, Theodoridis S (2010) A multimodal approach to violence detection in video sharing sites. In: 2010 20th international conference on pattern recognition (ICPR). IEEE, pp 3244–3247

Gori I, Aggarwal JK, Matthies L, Ryoo MS (2016) Multitype activity recognition in robot-centric scenarios. IEEE Robot Autom Lett 1(1):593–600. https://doi.org/10.1109/LRA.2016.2525002

Hassner T, Itcher Y, Kliper-Gross O (2012) 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), pp 1–6. www.openu.ac.il/home/hassner/data/violentflows/. Accessed 21 June 2018

Hoang TV, Tabbone S (2012) Invariant pattern recognition using the RFM descriptor. Pattern Recognit 45(1):271–284

Kantorov V, Laptev I (2014) Efficient feature extraction, encoding and classification for action recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Li W, Wen L, Choo Chuah M, Lyu S (2015) Category-blind human action recognition: a practical recognition system. In: The IEEE international conference on computer vision (ICCV)

Liu J, Wang G, Duan L, Abdiyeva K, Kot AC (2018) Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Trans Image Process 27(4):1586–1599. https://doi.org/10.1109/TIP.2017.2785279

Lytras MD, Visvizi A (2018) Who uses smart city services and what to make of it: toward interdisciplinary smart cities research. Sustainability 10(6):1998. https://doi.org/10.3390/su10061998

Manfredi M, Vezzani R, Calderara S, Cucchiara R (2014) Detection of static groups and crowds gathered in open spaces by texture classification. Pattern Recognit Lett 44:39–48

Osborne M, Moran S, McCreadie R, Von Lunen A, Sykora M, Cano E, Ireson N, Macdonald C, Ounis I, He Y, Jackson T, Ciravegna F, O'Brien A (2014) Real-time detection, tracking, and monitoring of automatically discovered events in social media. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, association for computational linguistics, Baltimore, Maryland, pp 37–42

Rabiee H, Mousavi H, Nabi M, Ravanbakhsh M (2018) Detection and localization of crowd behavior using a novel tracklet-based model.

Int J Mach Learn Cybern 9(12):1999–2010. https://doi.org/10.1007/s13042-017-0682-8

Rezazadegan F, Shirazi S, Upcroft B, Milford M (2017) Action recognition: from static datasets to moving robots. arXiv preprint arXiv:1701.04925

Ribeiro PC, Audigier R, Pham QC (2016) Rimoc, a feature to discriminate unstructured motions: application to violence detection for video-surveillance. Comput Vis Image Underst 144:121–143

Rota P, Conci N, Sebe N, Rehg JM (2015) Real-life violent social interaction detection. In: 2015 IEEE international conference on image processing (ICIP). IEEE, pp 3456–3460

Ryoo MS, Aggarwal JK (2010) UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html. Accessed 25 June 2018

Shams S, Goswami S, Lee K, Yang S, Park SJ (2018) Towards distributed cyberinfrastructure for smart cities using big data and deep learning technologies. In: 2018 IEEE 38th international conference on distributed computing systems (ICDCS). IEEE, pp 1276–1283

Shao Z, Cai J, Wang Z (2018) Smart monitoring cameras driven intelligent processing to big surveillance video data. IEEE Trans Big Data 4(1):105–116

Sidhu RS, Sharad M (2016) Smart surveillance system for detecting interpersonal crime. In: 2016 international conference on communication and signal processing (ICCSP). IEEE, pp 2003–2007

Sivarajasingam V, Shepherd JP, Matthews K (2003) Effect of urban closed circuit television on assault injury and violence detection. Inj Prev 9(4):312–316

Srivastava S, Bisht A, Narayan N (2017) Safety and security in smart cities using artificial intelligence—a review. In: 2017 7th international conference on cloud computing, data science engineering—confluence, pp 130–133

Studer G (2017) Live streaming violence over social media: an ethical dilemma. Charlest Law Rev 11:621

Tizhoosh HR (2015) Barcode annotations for medical image retrieval: a preliminary investigation. In: 2015 IEEE international conference on image processing (ICIP). IEEE, pp 818–822

Tizhoosh HR, Rahnamayan S (2016) Evolutionary projection selection for Radon barcodes. In: 2016 IEEE congress on evolutionary computation (CEC). IEEE, pp 1–8

Visvizi A, Lytras MD (2018) Rescaling and refocusing smart cities research: from mega cities to smart villages. J Sci Technol Policy Manag 9(2):134–145. https://doi.org/10.1108/JSTPM-02-2018-0020

Visvizi A, Lytras MD, Damiani E, Mathkour H (2018) Policy making for smart cities: innovation and social inclusive economic growth for sustainability. J Sci Technol Policy Manag 9(2):126–133. https://doi.org/10.1108/JSTPM-07-2018-079

Weinzaepfel P, Revaud J, Harchaoui Z, Schmid C (2015) Learning to detect motion boundaries. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2578–2586

Yuan F, Xia GS, Sahbi H, Prinet V (2012) Mid-level features and spatio-temporal context for activity recognition. Pattern Recognit 45(12):4182–4191

Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 28–35

Zhang T, Yang Z, Jia W, Yang B, Yang J, He X (2016) A new method for violence detection in surveillance scenes. Multimed Tools Appl 75(12):7327–7349

Zhou P, Ding Q, Luo H, Hou X (2017) Violent interaction detection in video based on deep learning. In: Journal of physics: conference series, vol 844. IOP Publishing, p 012044