IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 37, No. 5, October, 2020, pp. 687-701
Journal homepage: http://iieta.org/journals/ts

A Novel Spatio-Temporal Violence Classification Framework Based on Material Derivative and LSTM Neural Network

Wafa Lejmi[1], Anouar Ben Khalifa[2*], Mohamed Ali Mahjoub[2]

[1] Université de Sousse, ISITCom, LATIS - Laboratory of Advanced Technology and Intelligent Systems, Sousse 4011, Tunisia
[2] Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS - Laboratory of Advanced Technology and Intelligent Systems, Sousse 4023, Tunisia

Corresponding Author Email: anouar.benkhalifa@eniso.rnu.tn

https://doi.org/10.18280/ts.370501

Received: 18 July 2020
Accepted: 20 September 2020

Keywords:
challenges, classification, derivative, LSTM, motion, recognition, material, violence

ABSTRACT

In the current era, the implementation of automated security video surveillance systems is particularly needy in terms of human violence recognition. Nevertheless, the latter encounters various interlinked difficulties which require efficient solutions as well as feasible methods that provide a relevant distinction between normal human actions and abnormal ones. In this paper, we present an overview of these issues and a literature review of the related works and current research on-going efforts on this field and suggests a novel prediction model for violence recognition, based on a preliminary spatio-temporal features extraction using the material derivative which describes the rate of change of a particle while in motion with respect to time. The classification algorithm is then carried out using a deep learning LSTM technique to classify generated features into eight specified violent and non-violent categories and a prediction value for each class of action is calculated. The whole model is trained on a public dataset and its classification capacity is evaluated on a confusion matrix which assembles all the predictions made by the system with their actual labels. The obtained results are promising and show that the proposed model can be potentially useful for detecting human violence.

1. INTRODUCTION

Violence in its many different forms is pervasive in the lives of many people around the world. No nation is immune from its reaches and we are all affected to varying degrees. Many protect themselves by locking their homes and avoiding unsafe places. For others, there is no escape and the danger of violence exists even behind these closed doors far away from eyes.

According to the World Health Organization (WHO: World report on violence and health), violence is defined as follows:

"The threat or intentional use of physical force or power against oneself, against another or against a group or community that causes or is at high risk of causing trauma, death, psychological damage, maldevelopment or deprivation."

The WHO report on violence and health uses a typology which divides violence into three broad categories, according to who commits the violent act: Self-directed violence, interpersonal violence and collective violence.

Moreover, the Web is currently the information channel most frequently used by the young people wherever they are, and the mobility handicaps the vigilance of parents regarding websites content which is plagued with violence and harming children. Violence has widespread impacts which are not limited to physical injuries but also extend to psychological effects.

The global human toll of violence is nearly two million lives lost each year, and so many others, innumerable, devastated in ways that are not always apparent. Violence is among the main death reasons worldwide for people aged 15 to 44. Furthermore, a large UN study says that, worldwide, criminal activities claim more victims than terrorism and armed conflict combined. In particular, homicide (in any form whatsoever: war, murder ...) is the most extreme form of violence, and also the most measurable.

The 2019 UNODC Homicide Report (https://www.unodc.org/documents/data-and-analysis/gsh/Booklet2.pdf) states that intentional homicide generated the death of 464,000 persons around the world in 2017. The Americas registered the highest proportion (37%), and then comes Africa which shares over a third (35%) of the total. Although Asia has a large population, it registered less than a quarter (23%) of the total. Europe registered lower rates (4.7%) and Oceania accounted for the lowest share (0.2%). Such rates are alarming and simply hard to ignore. Thereby, video surveillance systems became a daily priority in public spaces around the world inasmuch as they help enhance citizens' security and, hence, reduce risks of becoming a victim of crime. Most of the researches that have been conducted during the last years were devoted to focusing on automatic Human Action Recognition (HAR) [1-4]. However, the automatic characterization of aggressive activities has been comparatively less studied [5, 6]. Furthermore, combined with the ever-increasing amount of captured video content and the growing need to appropriately describe such sensitive information, recognizing suspicious behaviour is more and more challenging and depends on different factors. Indeed, several constraints [6, 7] impede automatic detection of violence due to its complexity and its ambiguous aspect which

inflict certain obstacles in defining what should be specified as aggressive actions [8].

In this paper, we present a problem that has a serious impact on human security. We mainly focus on the suspicious behaviour such as assaults between individuals. Therefore, we suggest a novel framework based on material derivative and LSTM neural network to deal with violence recognition in surveillance videos. The proposed framework starts by taking a set of successive images as input. The first step is a preliminary spatio-temporal features extraction using the material derivative which describes the rate of change of a particle while in motion over time. Further, a classification step is carried out using a deep learning neural network LSTM technique to classify generated features into specified violent and non-violent categories. Our work contributed to the following three aspects:

- We suggest a novel framework to recognize violent actions based on spatio-temporal features extraction using the material derivative concept and a classification technique based on the LSTM neural networks.
- We detail the main challenges likely to be faced by violence detection in video sequences as well as the techniques suggested to overcome such issues.
- We conduct an extensive characterization of the related works in the evolving field of human violence detection.

2. ISSUES OF VIOLENCE DETECTION IN VIDEO

Capturing moving individuals in a video presents various challenges depending on the environment where the video is recorded and the camera used. If the environment is not controllable in terms of lighting and composition, issues are numerous such as complex backgrounds [9], sudden movements [10], occlusions [11] and moving shadows. If a video is filmed with a moving camera [10], objects may appear blurry or the lens may be partially distorted. In the following, we examine in detail some frequent challenges.

2.1 Lighting variations

Lighting conditions of a scene may vary. This can be gradual such as in an outdoor scene, when a cloud passes over the sun or changes from bright sunlight to cloudy or rainy weather. It can be sudden when turning on or off lights inside or a possible motion of the light source. Therefore, it is painful to guarantee a good video quality under varying lighting conditions as shown in Figure 1. To enhance night video, Soumya et al. [13] implemented a daytime coloring transfer method. They used moving pixel-based background estimation approach during night time. Then, a tone mapping is performed to prepare the night video pixels enhancement. This allowed dissociating the illumination and reflectance map from images with day and night background. Finally, the preprocessed night video is fused with background illuminations and a statistical color transfer is applied.

Zhou et al. [14] tackled illumination variation using a local histogram of oriented gradient descriptor. To deal with lighting changes, they adopted two techniques: Going halfway through the block, and then normalizing each LHOG. Recently, Kim et al. [15] suggested a new tip to enhance low-light image by considering as an illumination component the maximal value resulting from the diffusion process. Such value represents the bright pixel which adapts well to the illumination property in the dark. Thus, in accordance with the Retinex theory, the selection of the highest value at each pixel position of diffusion spaces allows separating the estimated illumination component from the scene reflectance.

Figure 1. Examples of violent scenes from RLVS dataset [12] showcasing a great diversity of lighting conditions

2.2 Moving background

Among the causes giving rise to a moving background, may be mentioned camera motion or change of the ambient light, which requires removing the noise. Dynamic background might be defined as a set of image sequences of complex scenes which includes moving elements such as patterned grounds, or sea waves and moving trees [9]. For example, as shown in Figure 2, in hockey games and movies, non-stationary backgrounds are due to filming tips and tricks like zoom in/out [16]. In an attempt to resolve this issue, the optical flow approach is used by Fu et al. [17] to analyze the movement, as its vector's magnitude provides a very useful indication to measure the motion. Also, the direction of the flow gives further information about the motion. A background algorithm which is motion-resilient was deployed for an efficient determination of optical flow between each couple of successive images. The movement of the camera produces a relative background motion at a uniform speed and actions seem to change position less uniformly. This made it possible to filter background movement noise.

Various techniques were described such as using a 3x3 Gaussian kernel in order to minimize noise effect, or a histogram equalization which is an enhancement of the image contrast through a better use of the range of possible pixel values, or a background subtraction to separate objects that are not related to the scene [19].

Figure 2. Sample fight-scene from Hockey dataset [18] with background motion

2.3 Camera motion

A camera may be handled manually and shaken by the person holding the device or mounted on something that moves. Its displacement makes that everything appears moving (egomotion) as shown in Figure 3. In such situation, it is less easy to separate moving objects and static ones. The background subtraction methods previously presented and usually conceived for static cameras do not apply directly to

the moving ones. Consequently, in case of moving camera, most of the commonly known methods are adaptations of background subtraction concept [10]. To deal with this challenge, certain works [20-24] opted for selecting points from a grid then using an optical flow or several tracking approaches such as Kanade-Lucas-Tomasi (KLT) feature tracker. Zhao et al. [23] suggested a new framework named IFB (Integration of Foreground and Background cues). They used the GMM (Gaussian Mixture Model) and then image alignment to extract foreground cues. They got the background cues from spatio-temporal characteristics which have been filtered by the homography transformation. To check videos, Febin et al. [25] filtered the movement by means of temporal derivative and did not extract features from most of the non-violent activities.

Figure 3. Example of moving camera effect seen on a sample fight from Peliculas dataset [18]

2.4 Occlusion

Occlusion is an issue that cannot be avoided in any tracking system. When a person moves, it may become hidden behind trees or other obstacles. These individuals in motion may be fully or partially occluded in a video stream and their actions may not be fully visible. Figure 4 shows some samples of partially occluded persons fighting each other.

Figure 4. Sample fight-scenes of persons occluded behind trees or vehicles from RLVS dataset [12]

Since the scene geometry has a fundamental impact on handling occlusion [26], Geiger et al. [27] applied the technique of epipolar lines and disparity map as well as ordering constraint [28]. To deal with occlusion, several techniques have been proposed such as those reported in the review presented by Chandel et al. [11] for object detection under occlusions. They used a couple of occlusion maps to describe occluded regions, and presented the field of movement between frames as a main element. Besides, cross checking and extrapolation are among the simplest techniques that were adopted for this issue. Fehrman et al. [29] completely removed the occluding objects to observe their background. They used a canny edge detector, and then generated the disparity map. Occlusion is properly addressed by using the classical Kalman filtering method. Niknejad et al. [30] used 2-layers classifiers based on CRF (conditional random field) and DPM (deformable parts model). Zhang et al. [31] adopted an approach using KLT (Kanade-Lucas-Tomasi) technique. Later, MoWLD (Motion Weber local descriptor) was suggested [32]. In fact, it offers better tolerance to partial occlusion by capturing local aspect using a histogram group of gradients from neighbour regions. Li et al. [33] tackled detecting human actions through walls and occlusions using Wi-Fi signals together with deep learning techniques. Relying on radio signals, they could precisely distinguish actions and interactions despite limited lighting conditions using solely radio frequency (RF) signals as input.

2.5 Motion blur

Motion blur is a well-known phenomenon that results from shooting moving objects or individuals with long shutter speeds. It arises if the exposure time is great relatively to the movement velocity. Indeed, motion blur occurs if singular images are displayed with persistence of important parts of the image duration. Figure 5 shows an example of motion blur. To resolve this issue, Marziliano et al. [34] presented an approach to estimate perceptual blur. They measured the average width of the frame's vertical edges. Such blur measurement is described in the spatial domain. Sometimes, it is obvious along edges and prominent in textured areas. First, a vertical edge detector is applied (such as vertical Sobel filter) to extract vertical edges in the frame. Next, each line of the edge image is scanned to get the edge position. At last, the total measure of the blur for the whole image results from the average of local values over all edge positions. Later, Kadim et al. [35] suggested another approach to estimate the blurriness of the image. They merged the Wrońskian's change detection technique [36] and the concept of neighbouring pixels to attenuate the noise generated by the imperfect arrangement of consecutive images. Deniz et al. [37] presented another approach without resorting to either tracking process or optical flow. They suggested a technique based on extreme acceleration patterns and then performed the Radon transform to the power spectrum of successive images. They noticed that great acceleration causes image blur and therefore tracking becomes less relevant or not feasible. To remove the blur, they carried out a deconvolution pre-processing step after performing an initial correlation to extract motion between each two successive images. More recently, Pujol et al. [38] computed acceleration between images assuming that their movement generates blur. They applied Radon transform to determine eccentricity related to acceleration that arises in fast Fourier transform if blur occurs.

Figure 5. Motion blur in three successive images in a battle clip from HMDB51 dataset [39]

2.6 Body shapes

In reality, all objects are three-dimensional, and may change their appearance when they move. For example, the front view differs from the side view as shown in Figure 6. Besides, there can be non-rigid objects like human hands whose shape changes over time. Many methods used two-dimensional body parts positions in a monocular image in order to estimate three-

dimensional human pose [40, 41]. Several studies have been carried out by considering manually labelled two-dimensional body parts positions. Later, numerous two-dimensional CNN-based pose estimation approaches have been suggested [42-45] and could be used to estimate three-dimensional human pose. The two-dimensional poses of multiple persons were successfully detected thanks to a non-parametric description which helped to learn body parts combinations. Elhayek et al. [40] automatically estimated the people number in the scene, and for each image, every three-dimensional skeleton was fitted to equivalent two-dimensional body parts positions calculated thanks to a well-known CNN based two-dimensional pose estimation approach. Their method is used to track many individuals in outdoor scenes and in case of low-quality scenes filmed with mobile-phone cameras.

Figure 6. Example of different body shapes from real time fight detection dataset [46]

2.7 Varying scales and multi-views

Individuals may appear at different scales in different videos yet perform the same action. Figure 7 highlights an example of varying scale for a sample of fight scene. Scale variance issues are mainly related to the distance that separates a subject from the camera [1]. Depending on such distance, many scales and descriptions of the same subject could exist. Aiming to tackle these challenges, many techniques have been proposed. Zhu et al. [47] presented an extended Random transform for invariant representation to geometrical transformations such as rotation, scaling and translation. They used binary silhouette information to recognize human action. Their approach can be used when the camera is unstable.

Goudelis et al. [48] used various Trace transform functionals to compute robust features for human action recognition that are efficient and invariant to scaling. Later, a recent method to detect real-time multi-scale action was presented by Sharaf et al. [49] using a descriptor relying on angular velocities of the three-dimensional joint data taken from depth sensors. To handle the intra-class actions variations, like temporal scale variations, the descriptor is calculated using various window scales per action. To recognize actions, Chen et al. [50] proposed a temporal scale-invariant deep learning model. Assuming that an action is made up of a number of ordered sub-actions, they found that sampling keyframes from every sub-action sequence is temporal scale-invariant to action quickness and helps to better recognize actions than the traditional serial keyframe sampling strategy.

Figure 7. Example of a varying scale for a sample of fight scene extracted from BEHAVE dataset [51]

More recently, Singh et al. [52] presented another model to recognize view-invariant human activities. During extraction step, they combined some tricks with the uniform LBP (local binary patterns) which is invariant to rotation and thus offers a view-invariant recognition of multi-views activity (Figure 8). The scale invariance is obtained, while calculating distance signal feature. The last module includes the use of HMMs (Hidden Markov models) which helped to provide view-invariant action recognition as well as time-scale invariability.

Figure 8. Multi-views of a kicking action from WVU dataset [53]

2.8 Changes in execution rate of activity

Each person conducts an action at their own pace. Besides, nothing guarantees that an individual will redo the action at the same speed each time. This change in the execution rate of an action is described in Figure 9 and should be considered in a violence action recognition system. According to Veeraraghavan et al. [54], few has been achieved to correct the effect of changes in the execution rate of an activity. So, they provided a systematic model-based method to learn such variations. They designed a Bayesian algorithm that considers the execution rate function as a variable of nuisance and integrates it out through a Monte Carlo sampling, in order to produce estimates of posterior classes. To deal with the variation in temporal execution rate, Abdelkader et al. [55] used an advanced DTW (dynamic time warping) algorithm for learning warping functions between various occurrences of each action based on geodesic distances on the shape space when calculating the temporal warping functions. Moreover, Amor et al. [56, 57] presented a comprehensive framework for analysing human actions using shapes of skeletons, based on relevant geometric tools which help maintain desired invariances based on elastic distance. The latter is invariant to random execution rates of activities. Ghorbel et al. [58] suggested a human action descriptor which relies on interpolating the joints kinematics such as position, velocity and acceleration. To deal with execution rate variations, they used skeleton normalization as well as temporal normalization.

A recap of the challenges previously explained is provided in Table 1, as well as the works that tackled the issues that have been addressed.

Figure 9. Variation in the execution rate of pushing activity from SBU Kinect interaction dataset [59]

Table 1. Main violence detection issues and related works

Challenge	Related work	Year
1. Lighting Variations	- Soumya et al. [13]	-2010
	- Zhou et al. [14]	-2018
	- Kim et al. [15]	-2019
2. Moving Background	- Fu et al. [17]	-2015
	- De Souza et al. [19]	-2017
	- Akti et al. [16]	-2019
3. Camera Motion	- Minematsu et al. [20]	-2015
	- Kurnianggoro et al. [21]	-2016
	- Minematsu et al. [22]	-2017
	- Zhao et al. [23]	-2018
	- Yu et al. [24]	-2019
	- Febin et al. [25]	-2020
4. Occlusion	- Cho et al. [28]	-2012
	- Zhang et al. [31]	-2012
	- Niknejad et al. [30]	-2013
	- Fehrman et al. [29]	-2014
	- Chandel et al. [11]	-2015
	- Zhang et al. [32]	-2015
	- Li et al. [33]	-2019
5. Motion Blur	- Marziliano et al. [34]	-2002
	- Kadim et al. [36]	-2013
	- Deniz et al. [37]	-2014
	- Pujol et al. [38]	-2019
6. Body Shapes	- Toshev et al. [45]	-2014
	- Insafutdinov et al. [43]	-2016
	- Bulat et al. [44]	-2016
	- Cao et al. [42]	-2017
	- Elhayek et al. [40]	-2018
	- Chen et al. [41]	-2020
7. Varying Scales and Multi-Views	- Zhu et al. [47]	-2009
	- Goudelis et al. [48]	-2013
	- Sharaf et al. [49]	-2015
	- Chen et al. [50]	-2017
	- Singh et al. [52]	-2019
8. Changes in execution rate of activity	-Veeraraghavan et al. [54]	-2009
	- Abdelkader et al. [55]	-2011
	- Amor et al. [56]	-2016
	- Ghorbel et al. [58]	-2016
	- Amor et al. [57]	-2019

3. RELATED WORKS

Over the past several years, various violence recognition and detection approaches have been proposed. We can basically gather them into five categories detailed as follows: approaches relying on local descriptors, approaches using optical flow descriptors, approaches using acceleration descriptors, approaches using dynamic textures and approaches using deep learning models.

3.1 Approaches relying on local descriptors

Such approaches commonly use the standard Scale Invariant Feature Transform algorithm [60] which detects and describes images local features and extracts remarkable interest points in the spatial domain. Afterwards, there is a rejection of candidate points which have limited optical flow around the neighbourhood. Nievas et al. [18] assessed the performance of the approaches of recognizing violence in videos using two current approaches: STIP (Space-Time Interest Points) [61] as an improvement of Harris corner detector, and MoSIFT (Motion Scale Invariant Feature Transform) as an extension of SIFT with an aggregated Histogram of Optical Flows which defines local motion. A versatile fight detector is built by local descriptors method which efficiently detects violence, even in case of a moving camera. It demonstrated encouraging results retaining 90% accuracy levels using MoSIFT features. Xu et al. [62] presented a reliable method of violent video detection that uses MoSIFT algorithm as well as sparse coding. They employed many techniques to generate an extremely distinctive video representation based on local features: MoSIFT detects the local shape and motion patterns of an action. The most indicative features of this descriptor are selected using a KDE (Kernel Density Estimation) based feature selection technique. Interesting results are obtained using two datasets: The violence detection performance of this approach reaches 94.3% on the first violence dataset and 89.05% on the second one. Senst et al. [63] proposed a local descriptor that uses SIFT algorithm which incorporates motion models based on appearance and Lagrangian. They evaluated the LaSIFT algorithm using a Bag-of-Words technique and great improvements were achieved in comparison with SIFT and MoSIFT approaches. The obtained accuracies are very good on a pair of datasets reserved for violent video detection: 93.32% on the first violence dataset and 92.42% on the second one. Later, Senst et al. [64] proposed a particular Lagrangian approach to automatically detect video footage violent scenes. They used Lagrangian direction fields relying on spatio-temporal representation and then applied an extended bag-of-words technique in a late-fusion way for classification. They demonstrated that capturing the temporal scale via the Lagrangian integration time parameter is a main key to detect violence. They tested LaSIFT algorithm with Bag-of-Words on four various violence datasets. The obtained results are very encouraging: 94.42% on the first dataset, 94.95% on the second one, 93.12% on the third one and 84.00% on the last one. Recently, Febin et al. [25] proposed a cascaded approach to detect violence thanks to a MoBSIFT algorithm, i.e. a mix between motion SIFT and motion boundary histogram. In this approach, this algorithm of movement filtering relying on temporal derivative allows to check the videos. Accuracy obtained using individual features of MoBSIFT with Random Forest classifier reached 98.2% on the first dataset. Using combinations of MoBSIFT and MF with Adaboost as well as Random Forest classifiers, it reached 98.9 % on the second dataset.

In general, MoSIFT features are shown to be mighty and employed for generic action recognition. Nevertheless, extracting such features is a computationally expensive process which takes almost one second per image if running on a high-performance CPU, and thus prevents exploiting them in heuristic contexts, if several camera streams need a real-time processing.

3.2 Approaches using optical flow descriptors

Optical flow methods are widely employed for action recognition, as they are very common for assessing motion detection from a set of images. Several approaches have been developed. To detect violence in crowded scenes, Hassner et al. [65] suggested an approach based on statistics describing flow-vector magnitudes change over time. Such statistics are presented through the Violent Flows (ViF) descriptors produced when calculating the optical flow between pairs of successive images, then a classification as either violent or non-violent is performed thanks to a linear SVM. This approach is computationally effective to detect violence in

crowded contexts and its accuracy rate is impressive (around 82.9%). Huang et al. [66] studied the optical flow and found that its variance change increases in case of a crowd violence. Therefore, they suggested a statistical approach relying on optical flow fields to detect the behaviours of a violent crowd. Their approach exploits the optical flow fields' statistical properties to obtain a SCOF (statistical characteristic of the optical flow) descriptor for images. Then, actions are classified thanks to a linear SVM. This method showed a competitive accuracy of 86.9% on the first dataset and 83.35% on the second one. To localize violence in surveillance scenes, Zhang et al. [67] proposed a robust framework based on GMOF (Gaussian Model of Optical Flow) to withdraw violence areas that were represented as a deflection from an ordinary behaviour of crowd noticed in the video. They densely sampled the candidate violence areas using an OHOF (Orientation Histogram of Optical Flow) descriptor then a linear SVM to classify events. The detection accuracy found on three different datasets (82.79%, 85.29% and 86.75%) demonstrated that this method is very efficient. However, it is discriminatorily inefficient in case of a disordered and dynamic background. Fu et al. [17] presented a method to identify violence using motion analysis tricks, which are reliable in case of low video resolution. They applied an optical flow approach to get motion information, and performed a two-level statistical aggregation to build high-level features from simple motion signals. Machine learning classifiers were employed to detect fights. The accuracies obtained on two public datasets are quite reasonable for the first one (84.5%) and quite encouraging for the second one (98.5%). A performance of 85.4% is obtained for the combined dataset with a drop of 6%, then 89.0% with a degradation of only 2.5%. Later, an OViF (Oriented Violent Flows) feature extraction approach using this same descriptor is proposed by Gao et al. [68] who attempted to make some improvements using combined features and multi-classifier combination strategies (AdaBoost+Linear-SVM) to achieve a better performance: The best rates of violence detection are 87.50% and 88.00% on the first and second datasets. Zhou et al. [14] suggested a novel approach using low-level features. To simplify the features and decrease the noise, they segmented the motion areas according to the optical flow fields' distribution. Then, LHOG descriptor was extracted from RGB images to capture the appearance information, and LHOF descriptor was extracted from optical flow images to obtain dynamic information of the objects. Finally, a late-fusion classification strategy was performed using BoW (bag-of-words) model, and then both vectors were combined and a SVM classifier was utilized. Such approach reached high accuracies on different datasets: 95.1%, 100% and 94.31%. Mahmoodi et al. [69] proposed HOMO (Histogram of Optical flow Magnitude and Orientation) descriptor to identify violent behaviour in both crowded and uncrowded situations, and a SVM classifier was adopted to get the classification. Accuracy rates of this approach are generally satisfying: (89.3% and 76.83% for the first and second dataset). Overall, optical flow methods are the best motion representation for action recognition and represent the best way to consider the temporal motion of the video. Nevertheless, almost of them are computationally demanding, sensitive to brightness change, and would need a specially designed hardware for real-time applications.

3.3 Approaches using acceleration descriptors

The concept of acceleration is closely related to motion, speed, and velocity. More precisely, acceleration is the rate of change of an object velocity. To detect violence, Datta et al. [70] exploited human limbs motion trajectory and orientation information during violence to compute jerk which is an effective mean to identify such behaviour. They presented an Acceleration Measure Vector where a jerk represents its temporal derivative. They merged analysis of two separate methods tested on eight different persons with various physical builds and under different backdrop conditions and which gave reliable results (around 87%) when combined. Deniz et al. [37] proposed an approach that considers intensive acceleration models as an essential distinctive feature of violent behaviour. These patterns were efficiently evaluated using the Radon transform to the power spectrum of two successive images through the 2D Fast Fourier Transform. Indeed, they assumed that when a sudden motion occurs between a pair of images, the power spectrum frame of the second one depicts an ellipse. The proposed approach aspires to detect the sudden presence of this ellipse and to estimate its eccentricity. The latter defines the acceleration magnitude. Impressive accuracy results of the presented method were obtained on first and second datasets using Adaboost classifier (98.9% and 90.1%), and on the third dataset using SVM classifier (93.4%). Mohammadi et al. [71] presented a new video descriptor using substantial derivative which represents a crucial point in fluid mechanics. They exploited the spatio-temporal characteristics of the substantial derivative when calculating convective and local accelerations estimated from the optic flow for each video. Then, each video was described using the bag-of-words, and SVM classifier is used with Histogram intersection kernel to form the final descriptor. The effectiveness of the suggested method was extensively evaluated on five benchmarks, including three standard datasets and a couple of YouTube video-surveillance sequences. The average accuracies obtained on the first and second datasets were very satisfactory (96.89 % and 85.43%) and prove that such descriptor performs well in the densely crowded situations.

Generally, acceleration patterns are highly informative in the task of violence recognition. However, few works studied them experimentally in developing violence detectors.

3.4 Approaches using dynamic textures

Dynamic texture recognition techniques were successfully applied to different scenes. For instance, Kellokumpu et al. [72] developed the Local Binary Patterns (LBP) technique, initially suggested to recognize texture in two-dimensional images and extended for three-dimensional videos, and it has shown its efficiency for dealing with recognizing motion patterns. Indeed, texture is extracted thanks to local comparisons between a pixel and those surrounding it. Such relations are encoded as short binary strings whose frequencies are merged to represent the describe image area. Yeffet and Wolf [73] suggested LTP (local trinary patterns) and encoded local motion information by considering self-similarity in three neighbourhood circles at a specific spatial position. Lloyd et al. [74] noticed that violence in city centre environments mainly happens in crowded places and consequently, human actions are occluded by other crowd individuals. Hence, they proposed the VCT (violent crowd

texture) method by modelling crowd dynamics using GLCM matrix typically used to estimate crowd density and which applies temporal encoding to represent crowd dynamics. A classification using Random Forest was performed to assess the VCT approach's power to distinguish violent and non-violent behaviour. This approach provides regularly high performance across various kinds of data. It attained ROC performance values of 0.98, 0.91, 0.97 and 0.93 for the four datasets respectively. Later, Lloyd et al. [75] improved the previous work by temporal summaries of GLCM matrix features. Indeed, they measured inter-image uniformity and demonstrated that the violent behaviour appearance varies in a less uniform way in comparison with other kinds of crowd behaviour. Evaluating their approach with a privately held dataset and three public datasets, they reported a ROC score of 0.9782, 0.9403, 0.8218 and 0.9956. Recently, one more texture feature descriptor approach is suggested by Lohithashva et al. [76] and based on LBP (Local Binary Pattern) and GLCM (Gray Level Co-occurrence Matrix). Prominent features were used with five different supervised classifiers on two standard benchmark datasets. Their proposed texture features fusion descriptor achieved better results than other existing approaches and showed that SVM was superior the other classifiers, with an accuracy and AUC results of respectively 91.51% and 93.60% on the first dataset, then, 89.06% and 93% on the second dataset. Overall, this kind of approach is computationally cheap and provides real-time description.

3.5 Approaches using deep learning models

The recent development of machine learning methods, especially deep learning, provided new opportunities to boost violence recognition. Xu et al. [77] presented the AMDN (Appearance and Motion DeepNet) framework using deep neural networks for an automatic learning of feature descriptions. They exploited the additional information of appearance and motion patterns through a double fusion framework. Many one-class SVM models were applied to forecast each input's anomaly scores. This method was evaluated on a pair of public datasets and showed promising performance in terms of Equal Error Rate (16% on the first dataset) and AUC (Area under ROC) (92%). Fang et al. [78] used a deep learning approach by respectively extracting saliency information (SI) and optical flow of images as main spatio-temporal features. They used PCANet deep learning network to imitate the human brain in extracting high-level features from SI and MHOF for suspicious event detection. Classification rates exceeded 99% on the used dataset for a PCANet with a filter size 5 × 5. To detect intensive violent actions, Dong et al. [79] presented a descriptor relying on a 3-stream deep neural network framework with LSTM (Long Short-Term Memory) to describe long-term temporal information. The resulting accuracy, applied on a public dataset and based on three streams and LSTM, was impressive (93.9%). Sudhakaran et al. [80] developed an end-to-end trainable deep neural network model. They used a CNN (convolutional neural network) for frame level features extraction. An agregation of frame level features is then carried out in the temporal domain thanks to a convLSTM which adopts convolutional gates. The CNN along with the convLSTM can capture the localized spatio-temporal characteristics and enable analyzing local motion taking place in the video. The suggested method was assessed on three public datasets and showed high performance in terms of accuracy that attained 97.1% on the first dataset, 100% on the second one and 94.57% on the third one. Carneiro et al. [81] suggested a model using a multi-stream classification and high-level features. They adopted a multi-stream learner where the streams are Visual Geometry Group (VGG-16) networks, i.e., each one is a highly optimized neural network trained on the ImageNet set of data. This approach is evaluated on two public datasets. On the first one, the combinations had an above 80% in metrics (reaching 89.10%) in case of an ordered dataset division, which indicates their pertinence considering a classification problem. Accuracies reached 95.76% with random dataset division. On the orderly divided second dataset, ideal rates were achieved (100%) and 99.67% in case of a random dataset division. Overall, the achieved results showed that combining correlated descriptor information with a multi-stream approach improves the classification accuracies of the deep learning method. A summary of the five-direction reference related works previously explained is provided in Table 2.

Table 2. Main groups of violence detection related works

Group of violence detection methods	Related work	Year
1. Local descriptors	- Nievas et al. [18]	-2011
	- Xu et al. [62]	-2014
	- Senst et al. [63]	-2015
	- Zhang et al. [32]	-2015
	- Senst et al. [64]	-2017
	- Febin et al. [25]	-2019
2. Optical flow descriptors	- Hassner et al. [65]	-2012
	- Huang et al. [66]	-2014
	- Zhang et al. [67]	-2015
	- Fu et al. [17]	-2015
	- Gao et al. [68]	-2016
	- De Souza et al. [19]	-2017
	- Zhou et al. [14]	-2018
	- Mahmoodi et al. [69]	-2019
3. Acceleration descriptors	- Datta et al. [70]	-2002
	- Deniz et al. [37]	-2014
	- Mohammadi et al. [71]	-2015
4. Dynamic textures techniques	- Kellokumpu et al. [72]	-2008
	- Yeffet and Wolf [73]	-2009
	- Lloyd et al. [74]	-2016
	- Lloyd et al. [75]	-2017
	- Lohithashva et al. [76]	-2020
5. Deep learning models	- Xu et al. [77]	-2015
	- Fang et al. [78]	-2016
	- Dong et al. [79]	-2016
	- Sudhakaran et al. [80]	-2017
	- Carneiro et al. [81]	-2019

4. PROPOSED METHOD

Due to the little number of descriptors relying on the acceleration concept summarized in the previous Table 2, the idea underlying our approach is inspired from the movement of a particle in fluid mechanics [82] using optical flow and a material derivative in order to calculate local (L) and convective (Cv) accelerations.

Further, we will classify the extracted features using a deep learning RNN (Recurrent Neural Network) which is LSTM (Long-Short Term Memory). The general layout of the suggested approach is shown in Figure 10.

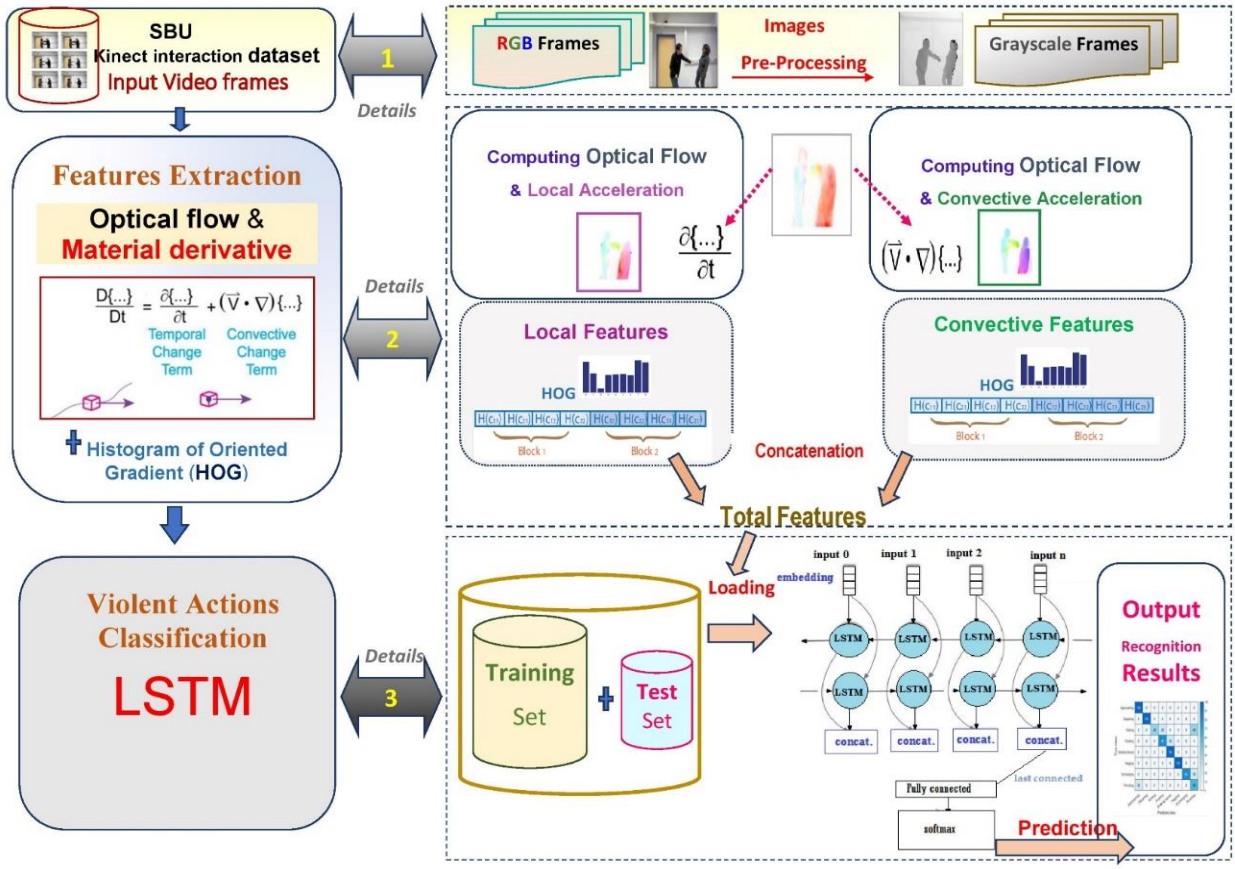

Figure 10. General layout of the suggested approach

4.1 Features extraction

This first step consists in computing the optical flow of each video sequence, then, using the material derivative we calculated the local acceleration (a^{L}) and convective acceleration (a^{Cv}). The extraction of primitives is performed using the Histogram of Oriented Gradient (HOG) for each kind of acceleration. The total acceleration (a^{Tot}) is then obtained by concatenating histograms of both accelerations.

To make reference easy, we listed in Table 3 the main mathematical notations adopted in the next section.

4.1.1 Material derivative as a physical concept of fluid mechanics

Fluid mechanics is a physics branch that often deals with properties which vary in space and change over time. A fluid particle velocity is defined as a position and time function (Figure 11). Thus, we need to consider the differentials of multivariable functions. If we consider the scalar function f(x,y,z,t) as a physical characteristic of the fluid at coordinates (x,y,z) where t is time, it is possible to describe the fluid motion by following a parcel along its trajectory [x(t), y(t), z(t)]. The material derivative is physically defined as the rate of change of a quantity being experienced by an observer who moves along with the flow [82, 84]. Indeed, what is observed is influenced by the stationary time-rate of change of the property ($\frac{\partial f}{\partial t}$), but is also depending on where the observer goes as it floats along with the flow. A material derivative notation is a derivative written with a capital D:

$$\frac{Df}{Dt} = \frac{\partial f}{\partial x} * v_x + \frac{\partial f}{\partial y} * v_y + \frac{\partial f}{\partial t} = \frac{\partial f}{\partial t} + \vec{v}\,\overrightarrow{\text{grad}}\, f \qquad (1)$$

We may apply this derivative to any fluid property, scalar or vector. This is the Gibbs notation of the material derivative:

$$\frac{Df}{Dt} \equiv \frac{\partial f}{\partial t} + \underline{v}.\nabla f \qquad (2)$$

We notice that this full rate of increase $\frac{Df}{Dt}$ for a certain particle defines the sum of two terms: $\frac{\partial f}{\partial t}$ defines the local or temporal acceleration, i.e., the velocity's rate of increase over time at a specific point in the flow. It results when the flow is unsteady. The second term ($\underline{v}.\nabla f$) represents the convective acceleration, i.e., the rate velocity's rate of increase due to the particle's change of position. It results when the flow is non-uniform, i.e., if the velocity changes along a streamline.

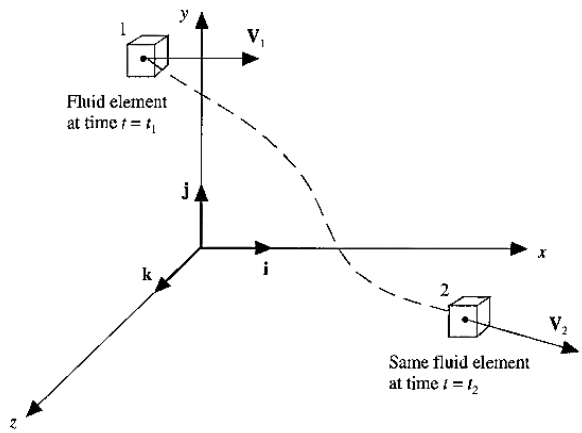

Figure 11. A fluid element moving in a flow field [83]

4.1.2 Local and convective accelerations estimation from videos

By analogy to the material derivative arising from the physics of fluid mechanics explained above, we estimated local and convective accelerations from a video.

Table 3. Mathematical notations adopted in this article

Mathematical notation	Definition
$f(x, y, z, t)$	Physical characteristic of the fluid at coordinates (x, y, z) and time t
$\frac{Df}{Dt} \equiv \frac{\partial f}{\partial t} + \underline{v} \cdot \nabla f$	Material derivative notation (General notation)
$\frac{Df}{Dt}$	Material derivative (Gibbs notation)
$= \frac{\partial f}{\partial x} * v_x + \frac{\partial f}{\partial y} * v_y + \frac{\partial f}{\partial t}$	
$= \frac{\partial f}{\partial t} + \vec{v}\,\overrightarrow{\text{grad}}\, f$	
$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \dots \end{bmatrix}$	Gradient operator of the scalar-valued multivariable f function, called "nabla"
$\frac{\partial f}{\partial t}$	Temporal or local acceleration (Lagrangian)
$(v \cdot \nabla f)$	Convective acceleration (Eulerian)
$I(x,y,t)=I(x+dx, y+dy, t+dt)$	Brightness Conservation Equation
$f_x U + f_y V + f_t = 0$	Optical flow equation
f_x, f_y	Pixel intensity gradients
U , V	Flow vector maps for perceived motion in x and y coordinate plane
(v_x^t, v_y^t)	Velocity v of each pixel in "x" and "y" directions
$a_x^t = v_x^t - v_x^{t-1}$	Local acceleration in an "x" direction
$a_y^t = v_y^t - v_y^{t-1}$	Local acceleration in a "y" direction
$a^L = \sqrt{{a_x^t}^2 + {a_y^t}^2}$	Magnitude of local acceleration of two successive optical flows
$a_x = (\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}) * v_x$	Convective acceleration in an "x" direction
$a_y = (\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}) * v_y$	Convective acceleration in a "y" direction
$a^{Cv} = \sqrt{a_x^2 + a_y^2}$	Magnitude of the convective acceleration

First, we need to calculate the optical flow of the video sequences. An optical flow represents a group of vector fields which relates a frame to an upcoming one. Each vector field describes the obvious movement of each pixel from frame to frame. Assuming that pixels intensity is conserved, we apply the "Brightness Conservation Theorem" which means that "The brightness of an object is constant from one image to another." To seek the displacement vector [dx, dy] at an x position of the image, so that the following one allows getting the same luminance and consequently the same grayscale, this concept can be written as follows:

$$I(x, y, t) = I(x + dx, \quad y + dy, \quad t + dt) \tag{3}$$

where, I is an image sequence, dy and dx are the displacement vectors for the pixel with coordinates $[x, y]$ and t and dt are the frame and temporal displacement of the image sequence. To calculate the optical flow, a standard Horn-Schunck method [85] is used. The optical flow equation is derived from the Eq. (3) as follows:

$$f_x U + f_y V + f_t = 0 \tag{4}$$

where, f_x, f_y are pixel intensity gradients and f_t is the first temporal derivative. Solving Eq. (4), we get two flow vector maps U and V that dictate perceived motion in both the x and y coordinate plane. In general, for each frame $\{I^t\}_{t=1}^{N}$ of the video, the optical flow $\{f^t\}_{t=1}^{N-1}$ represents each pixel's velocity in x and y directions:

$$f^t(x, y) = (v_x^t, v_y^t) \tag{5}$$

By applying (5), the local acceleration gets the value of the rate of velocity change over time at a fixed point in a flow field. We consider a_x^t as the local acceleration in an "x" direction and a_y^t as the local acceleration at "y" direction as detailed below:

$$a_x^t = v_x^t - v_x^{t-1} \tag{6}$$

$$a_y^t = v_y^t - v_y^{t-1} \tag{7}$$

The local acceleration of two successive optical flows is calculated as the following magnitude:

$$a^L = \sqrt{{a_x^t}^2 + {a_y^t}^2} \tag{8}$$

To get the rate of a velocity change with respect to position at a fixed time in a flow field, we calculate the convective acceleration. It is combined with spatial velocity gradients in the flow field. We consider a_x as the convective acceleration in an "x" direction and a_y as the convective acceleration in a "y" direction:

$$a_x = (\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}) * v_x \tag{9}$$

$$a_y = (\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}) * v_y \tag{10}$$

The convective acceleration magnitude is defined as follows:

$$a^{Cv} = \sqrt{a_x^2 + a_y^2} \tag{11}$$

Given the accelerations calculated for each video, we extracted features using the Histogram of Oriented Gradient feature descriptor. The HOG algorithm divides an image into many small connected regions, i.e. cells. For each region, HOG counts occurrences of gradient orientation. To help a descriptor to be invariant to illumination and shadowing, the gradient values are contrast-normalized over larger overlapping spatial blocks. The groups of adjacent cells correspond to spatial regions called blocks. The final HOG descriptor is built by grouping all normalized groups of histograms into a single feature vector. We chose HOG features extraction with a 2×2 block size and 16×16 cell size.

4.2 Classification

To classify our data sequence, we trained a deep neural

network [86]. Thus, we created an LSTM classification network [87]. It is a model that extends the memory of recurrent neural networks (RNN). Typically, recurrent neural networks have 'short-term memory' as they use determined previous information to be employed in the current neural network. Besides the advantages of using neural networks for time series, LSTM offers the capability of learning the items temporal dependencies in a sequence. This is appropriate for a time-series forecasting problem. As shown in Figure 12, an LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. A cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. Thus, computing f_t and i_t at each time step depends on the outputs (h_{t-1}, C_{t-1}) of its previous time step.

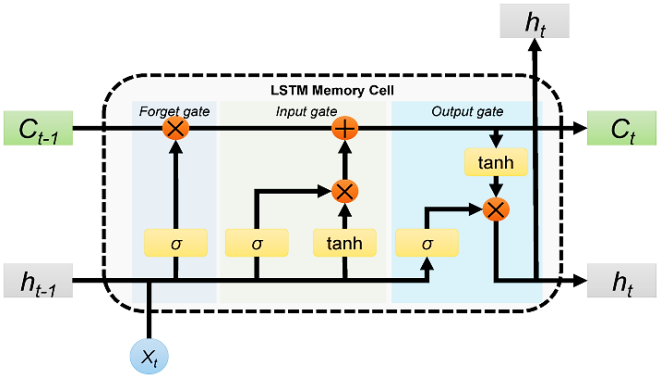

Figure 12. Computational graph of an LSTM block [88]

An LSTM layer can look at the time sequence in the forward direction. However, BiLSTM layer can look at the time sequence in both forward and backward directions. We mainly focus on the bidirectional LSTM layer. LSTM architecture is illustrated in Figure 13. During the classification phase, the extracted features are concatenated and fed to a Softmax classifier via a fully-connected operator.

Figure 13. LSTM network architecture

5. EXPERIMENTS AND RESULTS

We used MATLAB R2019a software to perform experimental work on an Intel (R) Core (TM) i-7 6500U, 2.5GHz and 8GB RAM under Windows 10 operating system (64-bit).

5.1 Dataset

To evaluate the proposed approach, we used a public dataset named SBU (Stony Brook University) Kinect Interaction dataset [59] that represents the collection of interactions of two people using the Microsoft Kinect sensor. Eight kinds of interactions are filmed between two persons indoor and belong to the following classes of activity: Approaching, Departing, Pushing, Kicking, Punching, Object exchanging, Hugging and Shaking hands. The videos have been filmed in the same environment. Seven actors accomplished these actions and this dataset consists of 21 sets where each one includes videos of a different pair of humans performing the eight interactions. In most interactivities, an individual acts and the other reacts. Each set includes one or two sequences per action category. We have a total of about 300 interactions. The depth map is 640×480 pixels. A commonly used split is to assign three-quarters of the data for training and the remaining one-quarter for tests. Thereby, for classification step, as the SBU Kinect Interaction dataset is composed of 21 sets, we used 16 sequences of the dataset for the learning phase and the five remaining sequences for the test. We noticed that the eighth action is missing for the second test sample.

5.2 Parameters and settings

First, we loaded both data sequences used for training and test. Then, we defined the network architecture and the training options: Indeed, we specified a sequence-to-sequence LSTM classification network with 500 hidden units. We set the feature dimension of training data as the input size, and the number of categories in the responses as the output size of the fully connected layer. We specified the solver as 'sgdm' and set the maximum number of training epochs to '100' with an initial default learning rate '0.01'. To prevent the gradients from exploding, we set the gradient threshold to '1'. We avoided shuffling the data every epoch by setting the 'Shuffle' option to 'never'. Given that mini batches are small with short sequences, a typical CPU is better suited for training. We specified a mini-batch size that evenly divides the data to ensure that the function uses all observations for training. Otherwise, the function ignores observations that do not complete a mini batch. We set the mini-batch size to 200. We trained the network, predicted the labels of the data and calculated the classification accuracy.

5.3 Results and discussion

We evaluated our suggested approach against three previous approaches tested on the same dataset. We recapitulate the results in terms of recognition rates in Table 4.

As depicted in Table 4, using LSTM classifier, our descriptor achieved the highest classification accuracy of **84.62%** over eight action classes, which outperforms SURF, STIP and Hierarchical Bidirectional Recurrent Neural Network (HBRNN) descriptors previously evaluated on this same dataset. To examine the recognition rate of each action and thus better understand the performance of the proposed method, we generated three confusion matrices (Figures 14, 15, 16). By analyzing the final confusion matrix (Figure 16) we find that the model can better differentiate between the classes. However, there is some confusion between similar actions. The latter are challenging since they are non-periodic with very resembling body movements. We explain below the origin of the errors presented in this confusion matrix: For instance, the confusion between "Kicking" and "Punching" or "Pushing" is due to the similarity of the bodies' behavior when performing such activities. The confusion between "Pushing" and "Shaking hands" is partially explained by the identical behavior of the hands. Indeed, these actions differ in the speed of execution which varies from person to person. Similarly, "Exchanging" could be confused with "Punching" as these actions contain joint body movements where both people extend and withdraw their arms. Generally, the system managed to completely identify these four actions and to relevantly recognize them: "Approaching", "Departing", "Shaking hands", "Hugging".

Table 4. Classification results using SVM then LSTM

Features descriptor	SURF [8]	STIP [89]	HBRNN [90]	$\mathbf{a^{Cv}}$	$\mathbf{a^{L}}$	$\mathbf{a^{Tot}}$
Classifier	SVM	SVM		LSTM	LSTM	LSTM
Recognition rate (%)	42.42	72	80.4	76.92	64.1	**84.62**

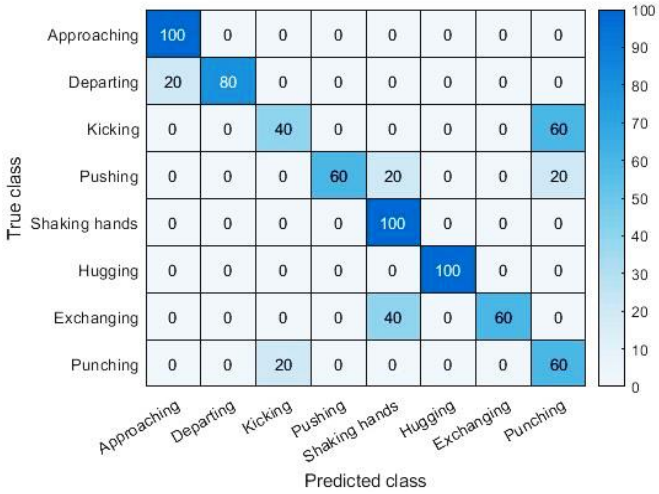

Figure 14. Confusion matrix after extracting a^{Cv} features

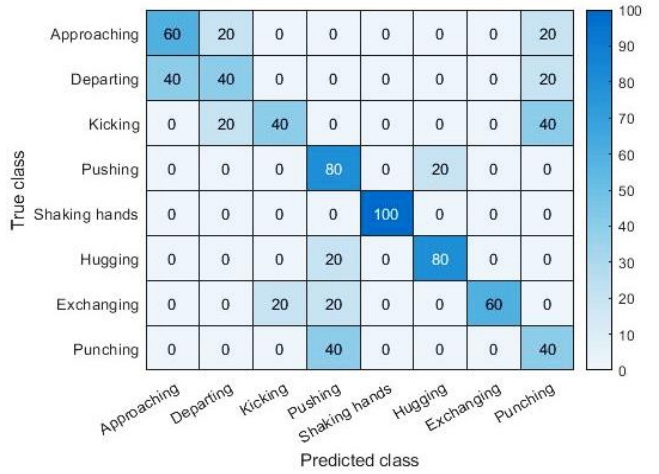

Figure 15. Confusion matrix after extracting a^{L} features

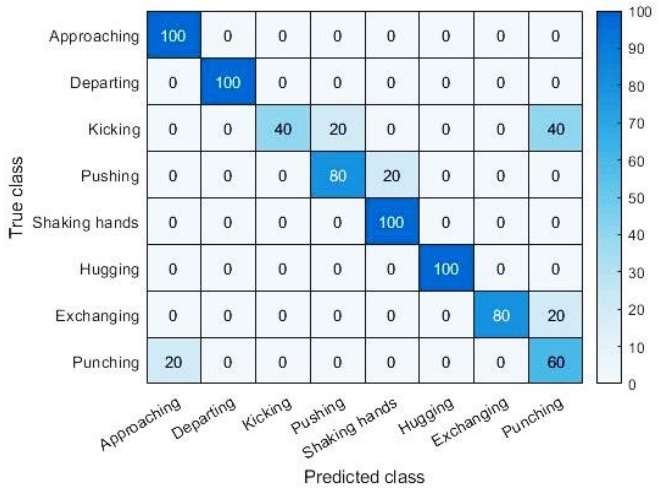

Figure 16. Final confusion matrix for a^{Tot} features

Comparing the results presented in Table 4, we notice that our method outperforms SURF descriptor which suffers from locality and long computation time despite its great number of detected features, and outstrips STIP and HBRNN descriptors which are not visibly as discriminative as the total acceleration descriptor. Moreover, LSTM classifier allows an almost clear discrimination between the actions and offers better performance for this dataset than SVM classifier which only uses local features and does not consider spatio-temporal ones. The specific advantages of our work consist in enhancing the limited number of descriptors based on the acceleration concept by exploiting some particle properties used in fluid mechanics with a deep learning classifier. The disadvantages are a lack of recognizing some actions such as "Kicking" (40%) and a significant computation time needed for LSTM compared to other methods. Possible future works which motivate further investigating violence classification are visual attention models based on deep learning.

6. CONCLUSION

After carrying out an exhaustive presentation of the issues mainly related to the field of violence detection and the solutions to counteract them as well as the broad groups of related works including the existing methods commonly employed in violence detection, we conceptualized a novel prediction framework for violent scenes recognition, based on a preliminary spatio-temporal features extraction using the material derivative which describes the rate of change of a particle while in motion with respect to time. Then, a classification algorithm was conducted using a deep learning neural network LSTM method to classify generated features of input images into specific violent and non-violent classes and a prediction value for each class of action was calculated. We trained the model on a public dataset and to evaluate its classification capacity some confusion matrices were then calculated according to the actual classes and gathering all the predictions made by the system with their actual labels.

REFERENCES

[1] Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A. (2020). Vision-based human action recognition: An overview and real world challenges. Forensic Science International: Digital Investigation, 32: 200901. https://doi.org/10.1016/j.fsidi.2019.200901

[2] Jegham, I., Khalifa, A., Alouani, I., Mahjoub, M.A. (2018). Safe driving: Driver action recognition using SURF keypoints. 2018 30th International Conference on Microelectronics (ICM), Sousse, Tunisia, pp. 60-63. https://doi.org/10.1109/icm.2018.8704009

[3] Mimouna, A., Khalifa, A., Ben Amara, N.E. (2018). Human action recognition using triaxial accelerometer data: Selective approach. 2018 15th International Multi-Conference on Systems, Signals & Devices (SSD), Hammamet, pp. 491-496. https://doi.org/10.1109/ssd.2018.8570429

[4] Jegham, I., Khalifa, A.B. (2017). Pedestrian detection in

poor weather conditions using moving camera. 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, pp. 358-362. https://doi.org/10.1109/aiccsa.2017.35

[5] Mabrouk, A., Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Systems with Applications, 91: 480-491. https://doi.org/10.1016/j.eswa.2017.09.029

[6] Lejmi, W., Khalifa, A.B., Mahjoub, M.A. (2019) Challenges and methods of violence detection in surveillance video: A survey. In: Vento M., Percannella G. (eds) Computer Analysis of Images and Patterns. CAIP 2019. Lecture Notes in Computer Science, 11679: 62-73. https://doi.org/10.1007/978-3-030-29891-3_6

[7] Lejmi, W., Mahjoub, M.A., Ben Khalifa, A. (2017). Event detection in video sequences: Challenges and perspectives. 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, pp. 682-690. https://doi.org/10.1109/fskd.2017.8393354

[8] Lejmi, W., Khalifa, A.B., Mahjoub, M.A. (2017). Fusion strategies for recognition of violence actions. 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, pp. 178-183. https://doi.org/10.1109/aiccsa.2017.193

[9] Ramzan, M., Abid, A., Khan, H.U., Awan, S.M., Ismail, A., Ahmed, M., Mahmood, A. (2019). A review on state-of-the-art violence detection techniques. IEEE Access, 7: 107560-107575. https://doi.org/10.1109/access.2019.2932114

[10] Chapel, M., Bouwmans, T. (2020). Moving objects detection with a moving camera: A comprehensive review. ArXiv, abs/2001.05238.

[11] Chandel, H., Vatta, S. (2015). Occlusion detection and handling: A review. International Journal of Computer Applications, 120(10): 33-38. https://doi.org/10.1.1.695.7078

[12] Soliman, M.M., Kamal, M.H., Nashed, M.A., Mostafa, Y.M., Chawky, B.S., Khattab, D. (2019). Violence recognition from videos using deep learning techniques. 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, pp. 80-85. https://doi.org/10.1109/icicis46948.2019.9014714

[13] Soumya, T., Thampi, S.M. (2015). Day color transfer based night video enhancement for surveillance system. 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Kozhikode, pp. 1-5. https://doi.org/10.1109/spices.2015.7091556

[14] Zhou, P., Ding, Q., Luo, H., Hou, X. (2018). Violence detection in surveillance video using low-level features. PLOS ONE, 13(10): 1-15. https://doi.org/10.1371/journal.pone.0203668

[15] Kim, W., Lee, R., Park, M., Lee, S. (2019). Low-light image enhancement based on maximal diffusion values. IEEE Access, 7: 129150-129163. https://doi.org/10.1109/ACCESS.2019.2940452

[16] Akti, S., Tataroglu, G.A., Ekenel, H.K. (2019). Vision-based fight detection from surveillance cameras. 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, pp. 1-6. https://doi.org/10.1109/ipta.2019.8936070

[17] Fu, E.Y., Leong, H.V., Ngai, G., Chan, S. (2015). Automatic fight detection based on motion analysis. 2015 IEEE International Symposium on Multimedia (ISM), Miami, FL, pp. 57-60. https://doi.org/10.1109/ism.2015.98

[18] Nievas, E.B., Suarez, O.D., García, G.B., 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. CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23678-5_39

[19] De Souza, F., Pedrini, H. (2017). Detection of violent events in video sequences based on census transform histogram. 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Niteroi, pp. 323-329. https://doi.org/10.1109/SIBGRAPI.2017.49

[20] Minematsu, T., Uchiyama, H., Shimada, A., Nagahara, H., Taniguchi, R.I. (2015). Adaptive search of background models for object detection in images taken by moving camera. In: 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, pp. 2626-2630. https://doi.org/10.1109/ICIP.2015.7351278

[21] Kurnianggoro, L., Yu, Y., Hernandez, D., Jo, K. (2016). Online background-subtraction with motion compensation for freely moving camera. In: Huang DS., Jo KH. (eds) Intelligent Computing Theories and Application. ICIC 2016. Lecture Notes in Computer Science, vol 9772. Springer, Cham. https://doi.org/10.1007/978-3-319-42294-7_51

[22] Minematsu, T., Uchiyama, H., Shimada, A., Nagahara, H., Taniguchi, R.I. (2017). Adaptive background model registration for moving cameras. Pattern Recognition Letters, 96: 86-95. https://doi.org/https://doi.org/10.1016/j.patrec.2017.03.010

[23] Zhao, C., Sain, A., Qu, Y., Ge, Y., Hu, H. (2019). Background subtraction based on integration of alternative cues in freely moving camera. IEEE Transactions on Circuits and Systems for Video Technology, 29(7): 1933-1945. https://doi.org/10.1109/TCSVT.2018.2854273

[24] Yu, Y., Kurnianggoro, L., Jo, K. (2019). Moving object detection for a moving camera based on global motion compensation and adaptive background model. International Journal of Control, Automation and Systems, 17(7): 1866-1874. https://doi.org/10.1007/s12555-018-0234-3

[25] Febin, I.P., Jayasree, K., Joy, P.T. (2019). Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Analysis and Applications, 23: 611-623. https://doi.org/10.1007/s10044-019-00821-3

[26] Tang, S., Andriluka, M., Schiele, B. (2014). Detection and tracking of occluded people. International Journal of Computer Vision, 110(1): 58-69. https://doi.org/10.1007/s11263-013-0664-6

[27] Geiger, D., Ladendorf, B., Yuille., A. (1995). Occlusions and binocular stereo. International Journal of Computer Vision (IJCV), 14(3): 211-226 https://doi.org/10.1007/BF01679683

[28] Cho, S.Y., Sun, I., Ha, J., Jeong, H. (2012). Occlusion detection and filling in disparity map for multiple view

synthesis. 8th International Conference on Computing and Networking Technology (INC, ICCIS and ICMIC), Gueongju, pp. 425-432.
[29] Fehrman, B., McGough, J. (2014). Handling occlusion with an inexpensive array of cameras. 2014 Southwest Symposium on Image Analysis and Interpretation, San Diego, CA, pp. 105-108. https://doi.org/10.1109/ssiai.2014.6806040
[30] Niknejad, H.T., Kawano, T., Oishi, Y., Mita, S. (2013). Occlusion handling using discriminative model of trained part templates and conditional random field. 2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast, QLD, pp. 750-755. https://doi.org/10.1109/ivs.2013.6629557
[31] Zhang, C., Xu, J., Beaugendre, A., Goto, S. (2012). A KLT-based approach for occlusion handling in human tracking. 2012 Picture Coding Symposium, Krakow, pp. 337-340. https://doi.org/10.1109/pcs.2012.6213360
[32] Zhang, T., Jia, W., Yang, B., Yang, J., He, X., Zheng, Z. (2015). MoWLD: A robust motion image descriptor for violence detection. Multimedia Tools and Applications, 76(1): 1419-1438. https://doi.org/10.1007/s11042-015-3133-0
[33] Li, T., Fan, L., Zhao, M., Liu, Y., Katabi, D. (2019). Making the Invisible Visible: Action Recognition Through Walls and Occlusions. IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 872-881. https://doi.org/10.1109/ICCV.2019.00096
[34] Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T. (2002). A no-reference perceptual blur metric. International Conference on Image Processing, 3: 57-60.
[35] Kadim, Z., Daud, M.M., Radzi, S.S.M., Samudin, N., Woon, H.H. (2013). Method to detect and track moving object in non-static PTZ camera. Int MultiConf Eng Comput Sci, 1. Lecture Notes in Engineering and Computer Science, 472-477.
[36] Durucan, E., Ebrahimi, T. (2001). Change detection and background extraction by linear algebra. Proceedings of the IEEE, 89(10): 1368-1381. https://doi.org/10.1109/5.959336
[37] 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), 2: 478-485.
[38] Pujol, F.A., Mora, H., Pertegal, M.L. (2019). A soft computing approach to violence detection in social media for smart cities. Soft Computing, 24: 11007-11017. https://doi.org/10.1007/s00500-019-04310-x
[39] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T. (2011). HMDB: A large video database for human motion recognition. 2011 International Conference on Computer Vision, Barcelona, pp. 2556-2563. https://doi.org/10.1109/ICCV.2011.6126543
[40] Elhayek, A., Kovalenko, O., Murthy, P., Malik, J., Stricker, D. (2018). Fully automatic multi-person human motion capture for VR applications. EuroVR.
[41] Chen, Y., Tian, Y., He, M. (2020). Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding, 192: 102897. https://doi.org/10.1016/j.cviu.2019.102897
[42] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 1302-1310. https://doi.org/10.1109/cvpr.2017.143
[43] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B. (2016). DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9910. Springer, Cham. https://doi.org/10.1007/978-3-319-46466-4
[44] Bulat, A., Tzimiropoulos, G. (2016). Human pose estimation via convolutional part heatmap regression. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9911. Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_44
[45] Toshev, A., Szegedy, C. (2014). DeepPose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, pp. 1653-1660. https://doi.org/10.1109/CVPR.2014.214
[46] Fu, E.Y., Leong, H.V., Ngai, G., Chan, S.C.F. (2017). Automatic fight detection in surveillance videos. International Journal of Pervasive Computing and Communications, 13(2): 130-156. https://doi.org/10.1108/ijpcc-02-2017-0018
[47] Zhu, P., Hu, W., Li, L., Wei, Q. (2009). Human activity recognition based on R transform and Fourier Mellin transform. In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2009. Lecture Notes in Computer Science, vol 5876. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10520-3_60
[48] Goudelis, G., Karpouzis, K., Kollias, S. (2013). Exploring trace transform for robust human action recognition. Pattern Recognition, 46(12): 3238-3248. https://doi.org/10.1016/j.patcog.2013.06.006
[49] Sharaf, A., Torki, M., Hussein, M.E., El-Saban, M. (2015). Real-time multi-scale action detection from 3D skeleton data. 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, pp. 998-1005. https://doi.org/10.1109/wacv.2015.138
[50] Chen, H., Chen, J., Hu, R., Chen, C., Wang, Z. (2017). Action recognition with temporal scale-invariant deep learning framework. China Communications, 14(2): 163-172. https://doi.org/10.1109/cc.2017.7868164
[51] Blunsden, S., Fisher, R.B. (2010). The BEHAVE video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA, 4(1-12): 4.
[52] Singh, R., Kushwaha, A.K.S., Srivastava, R. (2019). Multi-view recognition system for human activity based on multiple features for video surveillance system. Multimedia Tools and Applications, 78(12): 17165-17196. https://doi.org/10.1007/s11042-018-7108-9
[53] Kavi, R., Kulathumani, V. (2013). Real-time recognition of action sequences using a distributed video sensor network. Journal of Sensor and Actuator Networks, 2(3): 486-508. doi:10.3390/jsan2030486
[54] Veeraraghavan, A., Srivastava, A., Roy-Chowdhury, A.K., Chellappa, R. (2009). Rate-invariant recognition of humans and their activities. IEEE Transactions on Image Processing, 18(6): 1326-1339. https://doi.org/10.1109/tip.2009.2017143
[55] Abdelkader, M.F., Abd-Almageed, W., Srivastava, A., Chellappa, R. (2011). Silhouette-based gesture and action recognition via modeling trajectories on

Riemannian shape manifolds. Computer Vision and Image Understanding, 115(3): 439-455. https://doi.org/10.1016/j.cviu.2010.10.006

[56] Amor, B., Su, J., Srivastava, A. (2016). Action recognition using rate-invariant analysis of skeletal shape trajectories. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1): 1-13. https://doi.org/10.1109/tpami.2015.2439257

[57] Amor, B., Srivastava, A., Turaga, P., Coleman, G. (2019). A framework for interpretable full-body kinematic description using geometric and functional analysis. IEEE Transactions on Biomedical Engineering, 67(6): 1761-1774. https://doi.org/10.1109/tbme.2019.2946682

[58] Ghorbel, E., Boutteau, R., Bonnaert, J., Savatier, X., Lecoeuche, S. (2016). A fast and accurate motion descriptor for human action recognition applications. 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, pp. 919-924. https://doi.org/10.1109/icpr.2016.7899753

[59] Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D. (2012). Two-person interaction detection using body-pose features and multiple instance learning. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, pp. 28-35. https://doi.org/10.1109/cvprw.2012.6239234

[60] Lowe, D.G. (1999). Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, pp. 1150-1157. https://doi.org/10.1109/iccv.1999.790410

[61] Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2-3): 107-123. https://doi.org/10.1007/s11263-005-1838-7

[62] Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L. (2014). Violent video detection based on MoSIFT feature and sparse coding. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, pp. 3538-3542. https://doi.org/10.1109/icassp.2014.6854259

[63] Senst, T., Eiselein, V., Sikora, T. (2015). A local feature based on Lagrangian measures for violent video classification. 6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15). https://doi.org/10.1049/ic.2015.0104

[64] Senst, T., Eiselein, V., Kuhn, A., Sikora, T. (2017). Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. IEEE Transactions on Information Forensics and Security, 12(12): 2945-2956. https://doi.org/10.1109/tifs.2017.2725820

[65] Hassner, T., Itcher, Y., Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, pp. 1-6. https://doi.org/10.1109/cvprw.2012.6239348

[66] Huang, J.F., Chen, S.L. (2014). Detection of violent crowd behavior based on statistical characteristics of the optical flow. 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, pp. 565-569. https://doi.org/10.1109/fskd.2014.6980896

[67] Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., He, X. (2015). A new method for violence detection in surveillance scenes. Multimedia Tools and Applications, 75(12): 7327-7349. https://doi.org/10.1007/s11042-015-2648-8

[68] Gao, Y., Liu, H., Sun, X., Wang, C., Liu, Y. (2016). Violence detection using oriented violent flows. Image and Vision Computing, 48: 37-41. https://doi.org/10.1016/j.imavis.2016.01.006

[69] Mahmoodi, J., Salajeghe, A. (2019). A classification method based on optical flow for violence detection. Expert Systems with Applications, 127: 121-127. https://doi.org/10.1016/j.eswa.2019.02.032

[70] Datta, A., Shah, M., Lobo, N. (2002). Person-on-person violence detection in video data. Object Recognition Supported by User Interaction for Service Robots, 1: 433-438. https://doi.org/10.1109/icpr.2002.1044748

[71] Mohammadi, S., Kiani, H., Perina, A., Murino, V. (2015). Violence detection in crowded scenes using substantial derivative. 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Karlsruhe, pp. 1-6. https://doi.org/10.1109/avss.2015.7301787

[72] Kellokumpu, V., Zhao, G., Pietikainen, M. (2008). Human activity recognition using a dynamic texture based method. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 1-10.

[73] Yeffet, L., Wolf, L. (2009). Local trinary patterns for human action recognition. 2009 IEEE 12th International Conference on Computer Vision, Kyoto, pp. 492-497. https://doi.org/10.1109/ICCV.2009.5459201

[74] Lloyd, K., Marshall, A.D., Moore, S.C., Rosin, P.L. (2016). Detecting violent crowds using temporal analysis of glcm texture. CoRR abs/1605.05106

[75] Lloyd, K., Rosin, P.L., Marshall, D., Moore, S.C. (2017). Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. Machine Vision and Applications, 28(3-4): 361-371. https://doi.org/10.1007/s00138-017-0830-x

[76] Lohithashva, B.H., Aradhya, V.M., Guru, D.S. (2020). Violent video event detection based on integrated LBP and GLCM texture features. Revue d'Intelligence Artificielle, 34(2): 179-187. https://doi.org/10.18280/ria.340208

[77] Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. Proceedings of the British machine vision conference (BMVC '15), Swansea, UK.

[78] Fang, Z., Fei, F., Fang, Y., Lee, C., Xiong, N., Shu, L., Chen, S. (2016). Abnormal event detection in crowded scenes based on deep learning. Multimedia Tools and Applications, 75(22): 14617-14639. https://doi.org/10.1007/s11042-016-3316-3

[79] Dong, Z., Qin, J., Wang, Y. (2016). Multi-stream deep networks for person to person violence detection in videos. In: Tan T., Li X., Chen X., Zhou J., Yang J., Cheng H. (eds) Pattern Recognition. CCPR 2016. Communications in Computer and Information Science, vol 662. Springer, Singapore. https://doi.org/10.1007/978-981-10-3002-4_43

[80] Sudhakaran, S., Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, pp. 1-6.

https://doi.org/10.1109/avss.2017.8078468
[81] Carneiro, S.A., da Silva, G.P., Guimaraes, S.J.F., Pedrini, H. (2019). Fight detection in video sequences based on multi-stream convolutional neural networks. 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Rio de Janeiro, Brazil, pp. 8-15. https://doi.org/10.1109/SIBGRAPI.2019.00010
[82] Morrison, F.A. (2013). An Introduction to Fluid Mechanics. Cambridge University Press, UK. https://doi.org/10.1017/CBO9781139047463
[83] Rizvi, Z.A. (2017). A study to understand differential equations applied to aerodynamics using CFD technique. International Journal of Scientific & Engineering Research, 8(2): 16-19.
[84] Cushman-Roisin, B. (2013). Environmental Fluid Mechanics. United States of America: John Wiley & Sons, Inc.
[85] Horn, B.K.P., Schunck, B.G. (1981). Determining optical flow. Artificial Intelligence, 17(1-3): 185-203. https://doi.org/10.1016/0004-3702(81)90024-2
[86] Neelapu, R., Devi, G.L., Rao, K.S. (2018). Deep learning based conventional neural network architecture for medical image classification. Traitement du Signal, 35(2): 169-182. https://doi.org/10.3166/TS.35.169-182
[87] Olah, C. (2015). Understanding LSTM Network. Colah's Blog. Github, 27 Aug. 2015. Web. 04 May 2016.
[88] Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J., Jiang, J. (2020). Comparison of long short-term memory networks and the hydrological model in runoff simulation. Water, 12(1): 175. https://doi.org/10.3390/w12010175
[89] Mhiri, I., Khalifa, A. (2018). Violence classification algorithm based on substantial derivative. International Conference on Sensors, Systems, Signals and advanced Technologies (SSS'18), pp. 1-6.
[90] Du, Y., Wang, W., Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 1110-1118. https://doi.org/10.1109/CVPR.2015.7298714

NOMENCLATURE

∂ The partial derivative "del", the rate of change of a multi-variable function when we allow only one of the variables to change.

Subscripts

L Local
Cv Convective
Tot Total