Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines

Simone Accattoli, Paolo Sernani, Nicola Falcionelli, Dagmawi Neway Mekuria & Aldo Franco Dragoni

To cite this article: Simone Accattoli, Paolo Sernani, Nicola Falcionelli, Dagmawi Neway Mekuria & Aldo Franco Dragoni (2020): Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines, Applied Artificial Intelligence

To link to this article: https://doi.org/10.1080/08839514.2020.1723876

Published online: 06 Feb 2020.

APPLIED ARTIFICIAL INTELLIGENCE
https://doi.org/10.1080/08839514.2020.1723876

Taylor & Francis
Taylor & Francis Group

Violence Detection in Videos by Combining 3D Convolutional Neural Networks and Support Vector Machines

Simone Accattoli, Paolo Sernani, Nicola Falcionelli, Dagmawi Neway Mekuria, and Aldo Franco Dragoni

Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

ABSTRACT

Video-surveillance has always been a vital tool to enforce safety in both public and private environments. Even though (smart) cameras are nowadays relatively widespread and cheap, such monitoring systems lack effectiveness in most scenarios. In addition, there is no guarantee about a human operator who monitors rare events in live video footages, forcing the use of such systems after unwanted events already took their undisturbed course, as a mere tool for investigations. Having an intelligent software to perform the task would allow to unlock the full potential of video-surveillance systems. To this end, in this paper we propose a solution based on a 3D Convolutional Neural Network that can effectively detect fights, aggressive motions and violence scenes in live video streams. Compared to state-of-the-art techniques, our method showed very promising performance on three challenging benchmark datasets: Hockey Fight, Crowd Violence and Movie Violence.

Introduction

Despite the available technology improved the potential of surveillance systems, there is a general increase in public violence-related issues. For example, according to United Nations Office on Drugs and Crime (2019b), it was estimated that in the American continent the rate of homicide is 16.3 (with a rate of 29.53 for Brazil) per 100,000 inhabitants against 6.0 in entire world (United Nations Office on Drugs and Crime 2019a). This example is just one critical scenario among the many in which surveillance systems to automatically detect violence scenes might help. In fact, violence detection is a specific task of action recognition, and it is a binary problem which consists of recognizing the presence or the absence of violence. In particular, we consider as "violent" a voluntary action, exercised by one subject (or more) over another (or more), to act against the victims' will.

CONTACT Paolo Sernani p.sernani@univpm.it Università Politecnica delle Marche, Via Brecce Bianche 12, Ancona 60131, Italy

Following the trend of data mining techniques (Nissan 2012) at the service of law enforcement, automatic violence detection has the potential of providing immediate response in case of violence, preventing delays in calling for help when this might be a matter of life and death. Moreover, an automatic system significantly reduces the burden of a person who is supposed to monitor hours of videos.

In this work, we propose the use of an existing pre-trained 3D Convolutional Neural Network (CNN), named C3D (Tran et al. 2015), together with a Support Vector Machines (SVM) classifier, to implement a system for automatic violence detection in videos. The rationale for combining CNN and SVM for violence detection is based on the good accuracy achieved in detection and classification tasks on images, obtained in different domains (Niu and Suen 2012; Tao et al. 2016; Xue et al. 2016). Specifically, this paper adds the following contributions to the state of the art in violence detection:

- It demonstrates how a deep neural network, pre-trained on datasets not intended for violence detection, can be used to compute feature descriptors of videos, which can then feed a classifier to discriminate between violent and nonviolent videos.
- It proposes a system to detect both person-to-person and crowd violence with high accuracy.
- It describes the improvement in accuracy with respect to existing algorithms, by means of comparison tests on three benchmark datasets.
- It analyzes how the proposed system might be used for real-time violence detection.

Despite most of the existing techniques rely on hand-crafted features for violence detection (Ben Mabrouk and Zagrouba 2017; Gao et al. 2016; Hassner, Itcher, and Kliper-Gross 2012), this work is based on a deep learning technique. The main advantages of deep learning techniques compared to hand-crafted ones are to learn high level features and achieve a high degree of generalization (Bengio 2009), without prior information about the data, being capable to detect multiple-layer features (Guo, Wu, and Xu 2017). In fact, several studies (Ji et al. 2013; Taylor et al. 2010; Yu et al. 2017) proved that deep neural network-based approaches achieved good accuracy in action recognition tasks. Therefore, in order to directly benefit from these advantages, we exploited deep learning techniques and utilized them to perform violent video recognition.

The rest of paper is organized as follows. In Section 2, we discuss the main violence detection techniques available in state of the art. In Section 3, we provide a detailed explanation of the proposed system architecture. In Section 4 we compare our experimental results with the state of the art. Section 5 concludes the paper and highlights future works.

Related Work

Most of the approaches to automatic violence detection in videos derive from the general domain of action recognition. According to Xu et al. (2014), violence detection techniques can be categorized into two specific classes based on how features are extracted:

(1) Local features: the representation of an action is computed by using Points of Interest (POIs) across the frames of a video.
(2) Global features: the representation of an action is computed by evaluating characteristics from multiple frames as a whole.

Both classes use three kinds of information to detect violence in videos: spatial, motion (such as acceleration), and temporal. We refer to such information as "spatio-temporal" .

In addition to techniques based on hand-crafted features, deep learning-based approaches to violence detection are emerging in recent years. Hence, the following subsections divide the related works into local feature-based, global feature-based and deep learning-based, highlighting the differences with the approach proposed in this paper.

Violence Detection Based on Local Features

POIs are generally called Space-Time Interest Point (STIP) (De Souza et al. 2010) by considering the time domain for searching interest points on videos. The approaches based on local features extract local spatio-temporal descriptors from STIP found by different techniques such as 3D-Harris corner detector (Chen et al. 2008) or Differences-of-Gaussians (DoG) (Lowe 2004). The first approaches using such POIs were based on traditional descriptors for action recognition such as Histograms of Oriented spatial Gradient (HOG) and Histograms of Optical Flow (HOF), in the neighborhood of salient points. However, these were not descriptive enough to represent both spatial and temporal information of violence. Chen and Hauptmann (2009) proposed to use Motion SIFT (MoSIFT) for violence detection. MoSIFT applied standard SIFT algorithm combined with an analogous HOF, being more significant than traditional descriptor. Xu et al. (2014) proposed MoSIFT algorithm to extract the low-level descriptions for videos in addition to a non-parametric Kernel Density Estimation (KDE), in order to delete some irrelevant and redundant features. Moreover, they proposed to use sparse coding instead of Bag of Words (BoW) model to provide a more accurate and discriminative intermediate representation. They obtained good accuracy in detecting person-to-person fights (with standard datasets such as the "Hockey Fight") but worse results in detecting

crowd fights. Deniz et al. (2014) proposed to use acceleration to identify a large variation in speed, which could be a potential aggression. The acceleration is computed with the power spectrum of adjacent frames. This method is very fast, but the obtained accuracy is not good enough compared to other techniques: the main reason is that acceleration data gives only partial information to discriminate violent actions by those nonviolent but cannot represent all the spatio-temporal information in a video. Zhou et al. (2018) proposed to extract two variations of HOG and HOF called LHOG and LHOF, where L stand for 'Local'. Instead of computing HOG and HOF on POIs, the authors suggest to extract the motion regions and compute both descriptors on such regions. The experiments show that this method reaches high performance on some benchmark datasets. Approaches based on local spatio-temporal descriptors usually need to encode the extracted features (such as BoW or Sparse coding information) to get a more significant representation and a fixed dimension descriptor that can be used as input in a classifier, such as SVM. On the contrary, the deep learning-based approach we propose in this paper has a fixed output dimension given by the dimension of the output layer. Furthermore, if there are many moving subjects in the visual scene, they could produce many useless POIs. For these reasons, traditional local feature-based methods do not have good accuracy in crowded situations, while, as shown by the results presented in this paper, our approach is able to extract global information by showing a better generalization capacity compared to the state of the art in violence detection.

Violence Detection Based on Global Features

Most state-of-the-art approaches using global features are based on acoustic or visual features (Chen et al. 2011; Giannakopoulos et al. 2006; Lin and Wang 2009; Nam, Alghoniemy, and Tewfik 1998; Zajdel et al. 2007). Audio-based methods define "violence" as an event that includes shots, explosions, fights and screams, looking for such audio contents in the videos. These methods proof that the audio might be very important to classify a violence scene. Unfortunately, in lots of surveillance system, audio is usually unavailable. Like audio-based methods, traditional visual approaches define as violent a scene containing a particular visual content, such as blood, flame, explosions, and weapons. These methods could produce too many false-positive cases and do not generalize enough the violence in the videos. Hassner, Itcher, and Kliper-Gross (2012) proposed the Violent Flow descriptor (VIF) that is another global feature extractor to identify violence in crowded scenes. They extract features from a sequence of frames and the violence is identified by optical flow magnitude change. The advantage of VIF is the response time, and it can be used for a real-time detection. In addition, VIF is specifically designed for crowds. Gao et al. (2016) proposed

an extension of VIF called Oriented VIF (OVIF). They show that VIF may loss some important information as it considers only the magnitude change of the flow vector; instead, also the orientation adds discriminant information. Although OVIF performs better than VIF on person-to-person violence detection, it has lower results on the crowded scenes. Ben Mabrouk and Zagrouba (2017) proposed a method called DIMOLIF. It detects violence in both crowded and uncrowded scenes. The features are extracted by computing the bivariate distribution of the orientation and magnitude of optical flow vector calculated around STIP points. This method outperforms VIF, OVIF and their combination. However, the approaches based on optical flow have some limitations, such as the aperture problems, discontinuities or motion camera. Contrariwise, techniques based on deep learning do not suffer from these problems. In addition some researches such as Neelakantan et al. (2015) show how adding noise might increase accuracy and limit the overfitting of deep learning models.

Violence Detection Based on Deep Learning Techniques

In the last few years, with the great success of deep neural networks in action recognition, some studies introduced the use of neural networks in violence detection (Ding et al. 2014; Dong, Qin, and Wang 2016; Meng, Yuan, and Li 2017; Sudhakaran and Lanz 2017; Xu et al. 2015; Zhou et al. 2017). Obviously, a simple convolutional neural network can learn only the spatial information, since they were not designed to deal with time. To tackle this problem, the techniques that use neural networks introduced additional components that extract also the temporal information. In Sudhakaran and Lanz (2017), the authors proposed to use a stream of CNNs: each stream takes the difference between two consecutive frames as input, to force the neural network to learn motion features. However, with such an architecture, the network cannot learn long temporal information: to overcome such issues, the authors used a convolutional Long Short-Term Memory (ConvLSTM) unit at the end of each CNN. Dong, Qin, and Wang (2016) proposed to use three stream deep neural networks, namely spatial, temporal and acceleration streams, which extract different types of violence information from raw videos. The spatial stream is used to capture spatial correlations between violent actions and scenes. The temporal stream extracts short term action information by getting in input the optical flow image, and the acceleration stream takes the acceleration flow images as input. Each stream works individually and uses LSTM units to get temporal information returning as output a degree of confidence in range $[0, 1]$ with a SoftMax layer. The output of each stream is combined to get the final score. The main problem of this work is that the system is designed for person-to-person violence, so it has lower performance in crowded scenes. Meng, Yuan, and Li (2017)

proposed a novel method by integrating trajectory and deep convolutional neural networks: in this way they took both the potentiality of hand-crafted features and deep learned features. This method reached the best results of accuracy on person-to-person fight but gets lower results on crowd violence detection than those obtained by our work. Ding et al. (2014) proposed to use a 3D CNN for violence detection. Their network consists of nine layers: an input layer that takes a video of 60x90x40 size, three 3D convolutional layer alternated with two pooling, two fully connected and a softmax layer for classification. The neural network is trained directly on three benchmark datasets for violence detection. Contrariwise, the 3D CNN used in this work is pre-trained on a dataset not related to violence detection and it's used to generate feature descriptors. We compute a binary classification in violent or nonviolent videos by means of a Support Vector Machines (SVM) classifier. The main advantage in using a 3D CNN is that, in addition to the spatial information, it can extract also motion information from the raw input video without using any prior information. Hence, considering the above reasons, in this paper we propose the use of an existing pre-trained 3D CNN architecture called C3D (Tran et al. 2015) to detect violence in videos.

Proposed Method

One of the goals of the proposed study is to demonstrate the potential of C3D, a 3D CNN pre-trained with a large dataset of sport activities (indeed a very different use-case from violence detection), as a feature extractor for violent scenes classification. Without using any prior information, we get better generalization capabilities than state-of-the-art approaches, given that the proposed technique obtains a high accuracy both in person-to-person fights and crowded scenes. Thus, in this section, we explain our usage of the C3D model.

3D Convolution and 3D Pooling

Unlike a 2D CNN, a 3D CNN can model also the temporal information available in sample data, by using 3D convolution and 3D pooling. The 3D convolution is obtained using a 3D kernel on the cube formed by stacking adjacent frames together. The resulting feature map, being connected to several contiguous frames, acquires the information related to the movement of subjects in the video. Formally, as shown in Ji et al. (2013), the value in the position (x, y, z) on the j-th feature map in the i-th layer is given by:

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \quad (1)$$

where b_{ij} is the bias of the feature map, P_i and Q_i are the length and height of the 3D kernel, while R_i is the temporal dimension of the 3D kernel and w_{ijm}^{pqr} is the (p, q, r)-th weight connected to the m-th feature map of the previous layer. The 3D pooling, as the convolution operation, is based on its 2D counterpart, but adding the time dimension in its calculations. A visual example of the 3D convolution operation is provided in Figure 1.

System Architecture

Figure 2 illustrates the architecture of the proposed violence detection system, composed of a customization of the known C3D model and a linear SVM classifier. The expected input consists of 16 frames. The typical frame rate obtained from a video capture is about 30 fps, so we work under the assumption that 16 consecutive frames are enough to represent a violent action. The neural network extracts the features of those 16 frames; therefore, the input dimensions are 3x16xHxW, where H and W represent the height and width of the frames and the first element represents the number of channels. The network is used as a feature extractor, in order to compute a video representation to be given as input to a classifier. Then, in this experiment, we used a linear SVM to perform a binary classification (violent vs nonviolent videos).

The used C3D model is pre-trained on Sport-1M dataset (Karpathy et al. 2014), that is one of the largest benchmark datasets for video classification. Sport-1M includes over one million of sport videos: this is a key element that led us to use this model of neural network. In fact, apart from avoiding the burden of training a neural network from scratch, using a model pre-trained

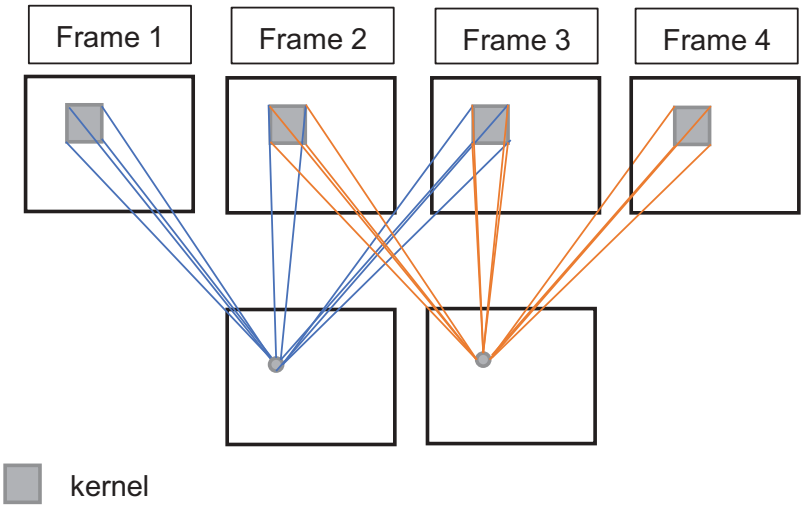

Figure 1. Example of the 3D convolution with 4 frames in input.

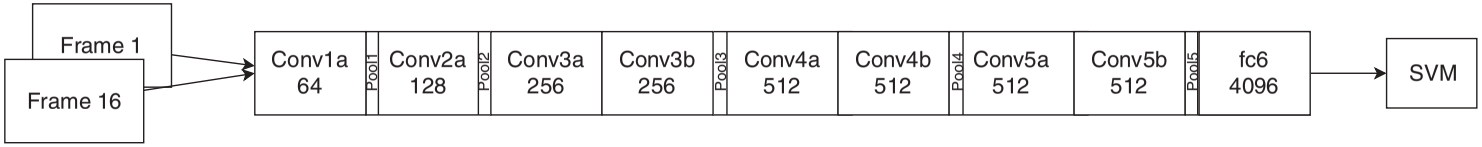

Figure 2. The proposed approach to violence detection: the videos are fed into the C3D (the layers are the same used in Tran et al. (2015), until the first fully connected layer). The resulting feature vector is used for classification with a SVM.

on large datasets (even though different from real usage scenarios) has been proven useful to achieve better generalization and prevent overfitting (Misra, Zitnick, and Hebert 2016; Simonyan and Zisserman 2014).

C3D was designed for action recognition rather than violence detection. As described in Tran et al. (2015), C3D uses a 3x3x3 kernels in a total of eight convolution layers alternated with five pooling layers, followed by two fully connected and a softmax output layer. The two final fully connected layers and the softmax are used for performing classification in action recognition. However, unlike the original architecture, our model took the output of the first fully connected layer to compute feature descriptors. In the architecture proposed in this paper, such descriptors are fed into a linear SVM classifier. The rationale for replacing the second fully connected layer and the softmax layer with a linear SVM classifier derives from the already proven increase of the accuracy in classification tasks (Tang 2013). Moreover, without such change, the subsequent processing would be to assimilate the input to one of the classes defined in the neural network training set, i.e. to a particular sport action. This would have decreased the performance of violence detection, because a violent action usually involves a specific behavior such as kicks and punches, which can make violence different from a sport activity. Figure 2 depicts the architecture of the recognition systems we propose: as a feature descriptor of 16 consecutive frames in a video, we took the output given by the first fully connected layer in the C3D model. Such descriptor, which includes 4096 values, is fed into a linear SVM in order to classify the frame sequence as "violent" or not. Then, a video is considered as including violence if it contains at least a violent sequence of frames.

Experiments and Results

To evaluate the performance of the proposed approach we used data from three benchmark datasets. We report the classification accuracy and the Area Under the Curve of the Receiver Operating Characteristic curve (AUC ROC), showing a comparison with the results obtained by the other approaches available in the state of the art. The experiments were performed with an NVIDIA GTX 1060 GPU, an i7 6700HQ and 16 GB of RAM. We computed the video descriptors with C3D by using CUDA core (version 9.1) and cuDNN acceleration (version 7.1).

Datasets

The three benchmark datasets we used are: Hockey Fight dataset (Bermejo Nievas et al., 2011), Crowd Violence dataset (Hassner, Itcher, and Kliper-Gross 2012) and Movie Violence dataset (Bermejo Nievas et al., 2011). We divided each available video into segments of 16 frames and we tagged each

one with the corresponding label, violent or nonviolent. Then, we randomly splitted the available segments in each dataset into training and testing samples, using a 5-fold cross validation scheme.

Hockey Fight Dataset

This dataset contains 1000 videos of violent and nonviolent (500 fighting and 500 non-fighting) behaviors of ice hockey matches. This is one of the most used datasets for violence detection, as it is applied as a benchmark by all other approaches we compare with. All the videos in the dataset have similar background and subjects.

Crowd Violence Dataset

This dataset contains 246 videos of violence scenes in crowds (123 violent and 123 nonviolent videos). Most of the samples represents football fans during matches. In general, in state-of-the-art algorithms the approaches that achieve high accuracy in Hockey Fight do not have the same accuracy in the Crowd Violence dataset, and viceversa.

Movie Violence Dataset

The Movie Violence dataset contains 100 fight and 100 non-fight videos extracted from several films.

Experimental Settings

The C3D network is implemented using the Caffe library (Jia et al. 2014). Initially, we processed each dataset to be able to work with the C3D model, in order to use all the available videos with their full duration (instead of only 16 frames for each video) and avoid to discard information. Hence, for each dataset and video, we extracted a collection of bags of 16 frames, tagging each bag as violent or nonviolent. We used the results as samples for training and testing purposes.

To compute the accuracy and compare our work with the existing approaches, we used the 5-folds cross validation scheme, which is the protocol usually applied in the state of the art. Thus, each dataset was divided into five different splits: four were used for training and the remaining one was used for testing. For each split we computed the accuracy and the final result was the average of the accuracy obtained at each iteration. In addition to the accuracy, we adopted as an evaluation measure the AUC ROC. As described in the previous section of the paper, we used the SVM classifier with a linear kernel to classify the videos as violent or nonviolent. We computed all the evaluation measures separately on each dataset.

Results and Discussion

Tables 1 and 2 show the comparison between our approach and the other studies in the relevant literature about violence detection. The comparison on the Movie Violence dataset was not reported, due to both a very limited number of samples and the easiness in discriminating videos in violent or not: we achieved 100% accuracy as the LSTM-based method described in Sudhakaran and Lanz (2017). In fact, most of the state-of-the-art approaches to violence detection do not use such dataset for benchmarking.

The results show that our approach reaches high accuracy on both the Hockey Fight and the Crowd Violence datasets, scoring better than the state of the art on person-to-person fights, and in line with the best approach (which has an accuracy equal to 98.6% against our 98.51%) on crowd fights . The advantage of our approach is that we reach high accuracy values in both datasets, differently from other approaches in the literature which achieved lower accuracy or are specific to only one use case between person-to-person and crowd fights. This means that the proposed method has a good generalization capability, being versatile and usable in different cases. In fact, Figures 3 and 4 show very similar ROC (and therefore AUC) for each of the folds in which the datasets were randomly splitted.

Table 1. Comparison of classification results on the Hockey Fight dataset. The AUC ROC of the proposed approach is the average of the AUC ROC of each fold of the experiments, as shown in Figure 3.

	Hockey Fight	
Algorithm	Acc ± SD	AUC
LHOG + LHOF + BoW (Zhou et al. 2018)	95.1 ± 1.15%	0.9798
Three streams + LSTM (Dong, Qin, and Wang 2016)	93.9%	–
CNN + LSTM (Sudhakaran and Lanz 2017)	97.1 ± 0.55%	–
Two-stream + IDT (Meng, Yuan, and Li 2017)	98.6 %	–
MoSIFT + KDE + SC (Xu et al. 2014)	94.3 ± 1.68%	0.9708
DIMOLIF (Ben Mabrouk and Zagrouba 2017)	88.6 ± 1.2%	0.9323
3D Conv Net (Ding et al. 2014)	91%	–
C3D + SVM (Proposed)	98.51 ± 1.05%	0.9832

Table 2. Comparison of classification results on the Crowd Violence dataset. The AUC ROC of the proposed approach is the average of the AUC ROC of each fold of the experiments, as shown in Figure 4.

	Crowd Violence	
Algorithm	Acc ± SD	AUC
LHOG + LHOF + BoW (Zhou et al. 2018)	94.31 ± 1.65%	0.9703
Three streams + LSTM (Dong, Qin, and Wang 2016)	–	–
CNN + LSTM (Sudhakaran and Lanz 2017)	94.57 ± 2.34%	–
Two-stream + IDT (Meng, Yuan, and Li 2017)	92.5%	–
MoSIFT + KDE + SC (Xu et al. 2014)	89.05 ± 3.26%	0.9357
DIMOLIF (Ben Mabrouk and Zagrouba 2017)	85.83 ± 4.26%	0.8925
3D Conv Net (Ding et al. 2014)	–	–
C3D + SVM (Proposed)	99.29 ± 0.59%	0.9900

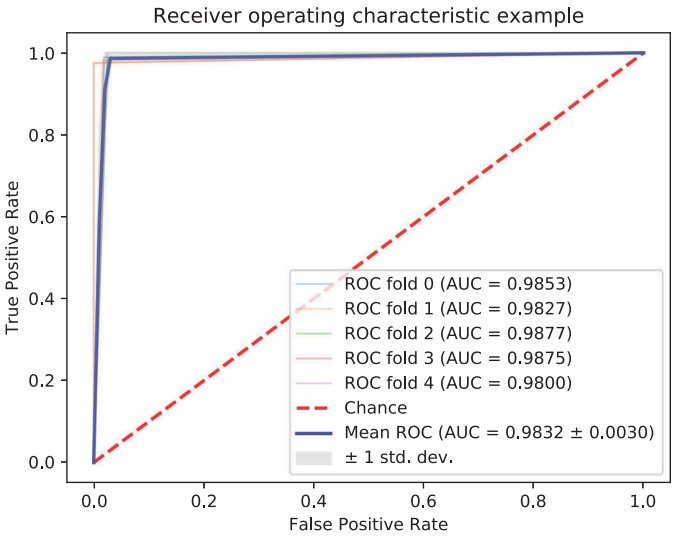

Figure 3. ROC and AUC for each of the five folds in the tests with the Hockey Fight dataset.

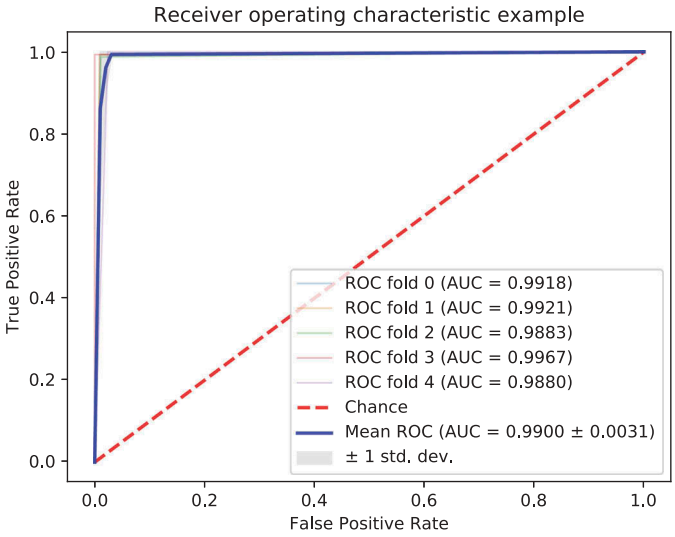

Figure 4. ROC and AUC for each of the five folds in the tests with the Crowd Violence dataset.

Furthermore, the AUC ROC is similar to the accuracy due to a combination of elements: a) each dataset is balanced, b) the accuracy is very high, and c) the number of false-positive and false-negative is similar. Moreover, we tested our approach on a new dataset which is the collection of the three-benchmark dataset with an addition of some videos taken from the UCF101 dataset (Soomro, Zamir, and Shah 2012) to enlarge the nonviolent behavior case. In this case we obtained an accuracy of 97,3%. We also analyzed the errors made by the network, considering as a positive case the

aggression class and negative case the non-aggression class, and observed that:

- Some of the false negatives are due to the fact that, among the 16 frames taken as input, there was not any presence of aggression. This is because in some videos labeled as violent there are portions of frames which do not contain a real aggression.
- Some of the false positives are friendly behaviors very similar to violent behaviors, such as small hit on the head of a person.

Finally, our results confirm the positive impact of deep learning-based techniques on automatic violence detection.

Limitations in Real-time Violence Detection

The proposed approach, based on the combination of C3D to compute the video descriptors and SVM as the classifier, completes the classification of each video in the datasets in 6.4 ± 0.15 s in average. Although six seconds could be better than the time necessary for a human operator watching the videos, such computation time might be considered too high for real-time violence detection. However, according to Tran et al. (2015), despite the convolution involves chunks of 16 frames of videos and might seem computationally expensive, C3D can compute video descriptors at 313 fps with a Tesla K40 GPU. In fact, in our implementation, a large portion of the computation time was due to the memory load of the C3D network, instead of the real output computation. Therefore, we can assert that, by improving the management of the memory in the proposed implementation, this approach can also be used for real-time violence detection, especially when coupled with dedicated software or co-designed hardware (Chen et al. 2018; Marques, Falcao, and Alexandre 2018).

Conclusions

This work presented a novel method to address violence detection in videos. We used C3D, a 3D Convolutional Neural Network architecture that allows to extract motion features without using any prior knowledge, to compute the feature descriptors of the videos. Then, we used such descriptors as an input for a linear SVM, to classify videos as violent or nonviolent. The proposed approach achieved better accuracy than other state-of-the-art techniques, and showed significant performances both in person-to-person and crowd fight datasets. The few errors are due to some misclassified friendly behaviors such as high fives, hugging or small hits. In this regard, the use of the acceleration information as a feature might improve accuracy. In general, future works can be summarized as:

- introducing additional information to improve accuracy;
- optimizing the implementation (for example by improving the loading of the model) to use the system for real-time violence detection;
- trying to categorize different violent behaviors, implementing a multi-class classifier.

Acknowledgments

The authors thank Site SpA for the support provided for the presented research.

Disclosure statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ben Mabrouk, A., and E. Zagrouba. 2017. Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognition Letters* 92:62–67. doi:10.1016/j.patrec.2017.04.015.

Bengio, Y. 2009. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning* 2 (1):1–127. doi:10.1561/2200000006.

Bermejo Nievas, E., O. Deniz Suarez, G. Bueno García, and R. Sukthankar. 2011. Violence detection in video using computer vision techniques. In *Computer analysis of images and patterns*, ed. P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, 332–39. Springer Berlin Heidelberg. doi:10.1007/978-3-642-23678-5_39.

Chen, A. T. Y., M. Biglari-Abhari, K. I. K. Wang, A. Bouzerdoum, and F. H. C. Tivive. 2018. Convolutional neural network acceleration with hardware/software co-design. *Applied Intelligence* 48 (5):1288–301. doi:10.1007/s10489-017-1007-z.

Chen, D., H. Wactlar, M. Chen, C. Gao, A. Bharucha, and A. Hauptmann. 2008. Recognition of aggressive human behavior using binary local motion descriptors. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 5238–41. doi:10.1109/IEMBS.2008.4650395.

Chen, L., H. Hsu, L. Wang, and C. Su. 2011. Violence detection in movies. 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, 119–24. doi:10.1109/CGIV.2011.14.

Chen, M. Y., and A. Hauptmann. 2009. MoSIFT: Recognizing human actions in surveillance videos. *Tech. Rep. CMU-CS-09-161*. Carnegie Mellon University.

De Souza, F. D. M., G. C. Chavez, E. A. Do Valle Jr, and A. de Araujo. 2010. Violence detection in video using spatio-temporal features. 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images, 224–30. doi:10.1109/SIBGRAPI.2010.38.

Deniz, O., I. Serrano, G. Bueno, and T. Kim. 2014. Fast violence detection in video. 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, vol. 2, 478–85.

Ding, C., S. Fan, M. Zhu, W. Feng, and B. Jia. 2014. Violence detection in video by using 3d convolutional neural networks. In *Advances in visual computing*, ed. G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. McMahan, J. Jerald, H. Zhang, S. M. Drucker, C. Kambhamettu, M. El Choubassi, et al., 551–58. Springer International Publishing. doi:10.1007/978-3-319-14364-4_53.

Dong, Z., J. Qin, and Y. Wang. 2016. Multi-stream deep networks for person to person violence detection in videos. In *Pattern recognition*, ed. T. Tan, X. Li, X. Chen, J. Zhou, J. Yang, and H. Cheng, 517–31. Singapore: Springer Singapore. doi:10.1007/978-981-10-3002-4_43.

Gao, Y., H. Liu, X. Sun, C. Wang, and Y. Liu. 2016. Violence detection using oriented violent flows. *Image and Vision Computing* 48-49:37–41. doi:10.1016/j.imavis.2016.01.006.

Giannakopoulos, T., D. Kosmopoulos, A. Aristidou, and S. Theodoridis. 2006. Violence content classification using audio features. In *Advances in artificial intelligence*, ed. G. Antoniou, G. Potamias, C. Spyropoulos, and D. Plexousakis, 502–07. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/11752912_55.

Guo, K., S. Wu, and Y. Xu. 2017. Face recognition using both visible light image and near-infrared image and a deep network. *CAAI Transactions on Intelligence Technology* 2 (1):39–47. doi:10.1016/j.trit.2017.03.001.

Hassner, T., Y. Itcher, and O. Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1–6. doi:10.1109/CVPRW.2012.6239348.

Ji, S., W. Xu, M. Yang, and K. Yu. 2013. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (1):221–31. doi:10.1109/TPAMI.2012.59.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, 675–78. ACM.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1725–32. doi:10.1109/CVPR.2014.223.

Lin, J., and W. Wang. 2009. Weakly-supervised violence detection in movies with audio and video based co-training. In *Advances in multimedia information processing - PCM 2009*, ed. P. Muneesawang, F. Wu, I. Kumazawa, A. Roeksabutr, M. Liao, and X. Tang, 930–35. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-10467-1_84.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60 (2):91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Marques, J., G. Falcao, and L. A. Alexandre. 2018. Distributed learning of cnns on heterogeneous cpu/gpu architectures. *Applied Artificial Intelligence* 32 (9–10):822–44. doi:10.1080/08839514.2018.1508814.

Meng, Z., J. Yuan, and Z. Li. 2017. Trajectory-pooled deep convolutional networks for violence detection in videos. In *Computer vision systems*, ed. M. Liu, H. Chen, and M. Vincze, 437–47. Springer International Publishing. doi:10.1007/978-3-319-68345-4_39.

Misra, I., C. L. Zitnick, and M. Hebert. 2016. Shuffle and learn: Unsupervised learning using temporal order verification. European Conference on Computer Vision, 527–44. Springer. doi:10.1177/1753193415618391.

Nam, J., M. Alghoniemy, and A. H. Tewfik. 1998. Audio-visual content-based violent scene characterization. Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), vol. 1, 353–57. doi:10.1109/ICIP.1998.723496

Neelakantan, A., L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens. 2015. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:151106807*.

Nissan, E. 2012. An overview of data mining for combating crime. *Applied Artificial Intelligence* 26 (8):760–86. doi:10.1080/08839514.2012.713309.

Niu, X. X., and C. Y. Suen. 2012. A novel hybrid cnn–svm classifier for recognizing handwritten digits. *Pattern Recognition* 45 (4):1318–25. doi:10.1016/j.patcog.2011.09.021.

Simonyan, K., and A. Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, Z. Ghahramani, eds. M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, vol 27, 568–76. Red Hook, New York, USA: Curran Associates, Inc.

Soomro, K., A. R. Zamir, and M. Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:12120402* doi:10.1094/PDIS-11-11-0999-PDN.

Sudhakaran, S., and O. Lanz. 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), 1–6. doi:10.1109/AVSS.2017.8078468.

Tang, Y. 2013. Deep learning using support vector machines. *CoRR abs/1306.0239*. http://arxiv.org/abs/1306.0239.

Tao, Q. Q., S. Zhan, X. H. Li, and T. Kurihara. 2016. Robust face detection using local cnn and svm based on kernel combination. *Neurocomputing* 211:98–105. doi:10.1016/j.neucom.2015.10.139.

Taylor, G. W., R. Fergus, Y. LeCun, and C. Bregler. 2010. Convolutional learning of spatio-temporal features. In *Computer vision – ECCV 2010*, ed. K. Daniilidis, P. Maragos, and N. Paragios, 140–53. Berlin, Heidelberg: Springer Berlin Heidelberg.

Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. 2015 IEEE International Conference on Computer Vision (ICCV), 4489–97. doi:10.1109/ICCV.2015.510.

United Nations Office on Drugs and Crime. 2019a. Intentional homicide victims. Accessed January 21, 2019 https://dataunodc.un.org/crime/intentional-homicide-victims.

United Nations Office on Drugs and Crime. 2019b. Official website. Accessed January 21, 2019. http://www.unodc.org/.

Xu, D., E. Ricci, Y. Yan, J. Song, and N. Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:151001553*

Xu, L., C. Gong, J. Yang, Q. Wu, and L. Yao. 2014. Violent video detection based on mosift feature and sparse coding. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3538–42. doi:10.1109/ICASSP.2014.6854259.

Xue, D. X., R. Zhang, H. Feng, and Y. L. Wang. 2016. Cnn-svm for microvascular morphological type recognition with data augmentation. *Journal of Medical and Biological Engineering* 36 (6):755–64. doi:10.1007/s40846-016-0182-4.

Yu, S., Y. Cheng, S. Su, G. Cai, and S. Li. 2017. Stratified pooling based deep convolutional neural networks for human action recognition. *Multimedia Tools and Applications* 76 (11):13367–82. doi:10.1007/s11042-016-3768-5.

Zajdel, W., J. D. Krijnders, T. Andringa, and D. M. Gavrila. 2007. CASSANDRA: Audio-video sensor fusion for aggression detection. 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, 200–05. doi:10.1109/AVSS.2007.4425310.

Zhou, P., Q. Ding, H. Luo, and X. Hou. 2017. Violent interaction detection in video based on deep learning. Journal of Physics: Conference Series, 6th conference on Advances in Optoelectronics and Micro/nano-optics, Najing, Jiangsu China, vol. 844, 1–9. IOP Publishing.

Zhou, P., Q. Ding, H. Luo, and X. Hou. 2018. Violence detection in surveillance video using low-level features. *PloS One* 13 (10):1–15. doi:10.1371/journal.pone.0203668.