A Crowd Analysis Framework for Detecting Violence Scenes

Konstantinos Gkountakos
Information Technologies Institute
CERTH
Thermi-Thessaloniki, Greece
gountakos@iti.gr

Konstantinos Ioannidis
Information Technologies Institute
CERTH
Thermi-Thessaloniki, Greece
kioannid@iti.gr

Theodora Tsikrika
Information Technologies Institute
CERTH
Thermi-Thessaloniki, Greece
theodora.tsikrika@iti.gr

Stefanos Vrochidis
Information Technologies Institute
CERTH
Thermi-Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute
CERTH
Thermi-Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

This work examines violence detection in video scenes of crowds and proposes a crowd violence detection framework based on a 3D convolutional deep learning architecture, the 3D-ResNet model with 50 layers. The proposed framework is evaluated on the Violent Flows dataset against several state-of-the-art approaches and achieves higher accuracy values in almost all cases, while also performing the violence detection activities in (near) real-time.

CCS CONCEPTS

• Computing methodologies → Scene understanding; Scene anomaly detection.

KEYWORDS

Violence detection; crowd analysis; deep learning; 3D-convolutional neural networks; Violent Flows

ACM Reference Format:
Konstantinos Gkountakos, Konstantinos Ioannidis, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2020. A Crowd Analysis Framework for Detecting Violence Scenes. In Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR ’20), June 8–11, 2020, Dublin, Ireland. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390725

1 INTRODUCTION

Monitoring visual streams from events, such as football matches and protests, for automatically detecting signs of violence is particularly valuable for law enforcement and security practitioners. Recent studies have focused on the detection of violence or fighting among multiple individuals in a crowd, with particular emphasis on violent scenes that cannot be effectively detected by the security personnel in the field. Towards this objective, the latest advances in deep learning have been exploited, whereby the temporal analysis of visual information is almost always performed using Convolutional Neural Networks (CNNs) [4], Recurrent Neural Networks (RNNs) [8], or 3D Convolutional Neural Networks (3D-CNNs) [22].

In this context, this work proposes a crowd violence detection framework that aims at analysing crowd-centered video footage and detecting scenes that contain indications of violence. Initially, the input streams (received from CCTVs and surveillance cameras) are encoded and, subsequently, specific key-frames are extracted and processed in order for the framework to provide the confidence score of the violence prediction (Figure 1). In particular, the proposed crowd analysis framework consists of four sub-modules: (i) the sampler which is responsible for balancing the information that should be processed by the neural network, (ii) the feature extractor that is exploited for encoding the frames to visual features, (iii) the main neural network that is built during the training phase and used during testing to evaluate in (near) real-time the presence of violence, and (iv) a graphical user interface for demonstrating the monitored streams and the prediction provided by the framework.

The proposed framework for detecting violent scenes of crowd-centered video footage relies on 3D-CNNs, and in particular on the 3D-ResNet [9], and is evaluated on the Violent Flows dataset [10]. To the best of our knowledge, this is the first time a 3D-ResNet architecture is applied to crowd violence detection, and it is only one of the very few deep learning approaches applied to this problem; several deep learning approaches have been previously applied to violence detection (e.g., [6, 21]), but only a few (namely [8] and [22]) to crowd violence detection. Additional contributions of this work
include the comprehensive experimental evaluation and extensive comparison to the state-of-the-art, the efficiency of the proposed framework that can process visual streams with no more than 1 second of delay, and a built-in demo to visualise the its predictions.

2 RELATED WORK

Methods based on hand-crafted features and on trajectory analysis were the first to be employed for the detection of violence in video scenes. In particular, Datta et al. [3] proposed a trajectory motion-based approach that considers the limb orientation of each person. Similarly, Nguyen et al. [18] used a hierarchical Hidden Markov model to enhance violence recognition, while [25, 29] took into account the motion modality of the SIFT features to generate robust descriptors. Other approaches incorporate additional modalities, e.g., audio [15], in order to improve violence detection.

Hassner et al. [10] proposed a framework based on the optical flow information; specifically, they proposed Violent Flow (ViF) descriptors followed by Support Vector Machines (SVMs), while Mabrouk et al. [14] generated a spatio-temporal feature extractor based on optical flow features. Zhou et al. [31] generated low level descriptors by extracting features from regions characterised by higher values of optical flow. Furthermore, Huang et al. [11] performed violent crowd behaviour analysis by considering only the statistical properties of the optical flow field in video data and performed classification using SVMs. Zhang et al. [30] presented a violence detection framework from surveillance video streams based on a Gaussian model of optical flow; they extracted violence flow vectors and also used SVMs for the classification. Gao et al. [7] proposed an oriented ViF descriptor that utilises the orientation of the optical flow information, which was not considered by the ViF. Recently, Mahmoodi et al. [16] proposed a method that computes the optical flow between sequential frames and compares the magnitude and orientation of each pixel in each frame to the global optical flow to obtain the changes in orientation and magnitude.

Works such as [1, 23, 27, 28], on the other hand, proposed algorithms that learnt discriminative dictionaries for semi-supervised classification. Bilinski et al. [2] reformulated the Improved Fisher Vectors in order to increase the accuracy of and speed up violence recognition. Yeffet et al. [26] proposed a fast method for detecting actions by encoding every pixel in every frame as a short string of ternary digits using a process which compares each frame to the previous and to the next frame. Laptev et al. [12] and Mohammadi et al. [17] took into account spatio-temporal features to generalise spatial pyramids across time and exploit the characteristics of substantial derivatives, respectively. Nievas et al. [19] constructed a versatile and accurate fight detector using a local descriptors approach. Finally, Lloyd et al. [13] proposed visual descriptors referred to as grey level co-occurrence texture measures (GLCM) to encode crowd scenes in a spatiotemporal manner in order to detect violence.

The breakthrough of Deep Learning (DL) techniques in computer vision has also affected the crowd violence detection field, by replacing the hand-crafted and trajectory analysis descriptors with learnable features extracted directly from deep neural networks, typically CNNs. The corresponding methods learn end-to-end representations from the images to feature vectors, with the goal to effectively detect violent scenes in videos. In particular, D. Xu et al. [24] proposed a novel unsupervised deep learning framework for anomalous event detection in complex video scenes. Sudhakaran et al. [21] proposed a method that encodes spatiotemporally the visual information and solves the classification problem of violence detection exploiting Long Short-Term-Memory (LSTM) units, while work in [8] improves the model using a Bidirectional Convolutional LSTM network. Fenil et al. [5] proposed a violence recognition framework applied to footages of football matches by extracting Histogram of oriented Gradient (HoG) features and feeding the vectors into a bidirectional LSTM network. Finally, Ullah et al. in [22] proposed a 3D-CNN architecture that first detects persons and considers only frames that contain persons for the final prediction.

3 CROWD VIOLENCE DETECTION FRAMEWORK

The proposed framework follows the supervised learning paradigm for crowd violence detection and employs a deep neural network architecture, namely the 3D-ResNet, a 3D CNN-based architecture, that was selected to fulfill the (near) real-time processing requirement. Based on the work of [9], we select the 3D-ResNet architecture with 50 layers depth, since it achieves close to state-of-the-art accuracy without the need for excessive computational resources.

The 3D-ResNet-50 consists of four bottleneck blocks, with each block consisting of three convolutional layers with filter sizes 1x1x1, 3x3x3, and 1x1x1, respectively. The shortcut pass connects the top of each block to the layer before the last activation layer of the block. The ReLU (Rectified Linear Unit) activation function was applied, while batch normalisation layers are also included (Figure 2). For demonstration purposes, the 2D-ResNet architecture is depicted with the only difference being the third dimension of the convolutional layers. The input layer was set to 112x112x3 and the kernel

Figure 2: ResNet-50 Model. (https://towardsdatascience.com/understanding-and-coding-a-resnet-in-keras-446d7ff84d33).
size of the third dimension of convolution to 16. Random cropping, flipping, and different scales were used for data augmentation for the model to generalise better and avoid overfitting.

First, all the frames of the videos are extracted and saved in a valid format, so that they could be fed into the neural network. For training our architecture, a learning rate equal to $10^{-4}$ was initially selected and was subsequently decreased following the reduce-on-plateau strategy with max patience set to 10 epochs. A negative log-likelihood criterion was used during training, along with Stochastic Gradient Descent (SGD) for implementing back propagation with momentum equal to $0.9$. The total number of epochs and the applied batch size were 200 and 1, respectively. All implementation activities were performed using the PyTorch 1.0 framework and an NVidia RTX 2080ti GPU with 11GB memory.

The requirement for (near) real-time processing is one of the main challenges. As mentioned, the selected neural network is based on 3D convolution processing. Specifically, the parallel processing of multiple frames provides the functionality of (near) real-time processing without generating video flickering. More specifically, our framework processes simultaneously 16 frames per batch. Hence, the processing of one second in a video stream requires at most two iterations when the model is set to inference state and for videos with frame rates lower than 30. Our implementation, using the aforementioned configuration and hardware, generates predictions (for batch size equal to 1) in no more than 150ms (300ms are needed for processing 32 frames, i.e., 2 batches). It should also be noted that the extraction of frames (which though does not take place in the case of video streams) is not a time-consuming process and can be covered within the $1000 - 300 = 700$ms timeframe.

4 EXPERIMENTAL EVALUATION & RESULTS

4.1 Datasets
To evaluate the performance of relevant methods, several evaluation datasets have been developed for research purposes. The most commonly used datasets are presented in Table 1; these include Violent Flows [10], Hockey Fights [19], and Action Movies [19].

Violent Flows is a widely used dataset that was introduced in 2012 and consists of 246 videos, with half the videos depicting violent crowd scenes and half non-violent scenes. The videos’ resolution is 320x240 pixels and the dataset is divided into 5 subsets, typically used for 5-fold cross-validation. In addition, the Hockey Fights dataset was introduced one year earlier and consists of 1000 videos that are divided into two categories, videos describing violent or non-violent scenes of ice hockey matches, while their resolution varies. Finally, the Action Movies dataset consists of 200 videos of resolution 720x576 pixels and involves violence scenes in movies focusing on fights between two persons. Hence, it is not so relevant to “crowd violence detection”, but to “violence detection”.

As the above discussion indicates, Violent Flows is the more relevant dataset to real-life violence scenes, and in particular to violence crowd-centered scenes, and will thus be used in this work.

4.2 State-of-the-Art Performance
Table 2 presents the evaluation performance of state-of-the-art methods for both hand-crafted (HC) and deep learning (DL) approaches on the Violent Flows (VF) dataset, as these have been reported in the respective publications. Specifically, the Mean Accuracy (MA), over the five predefined folds of the dataset, and Standard Deviation (SD), where available, are presented; in one case the max accuracy (i.e., the best accuracy value across the five folds) is reported. In all cases, the best values are depicted in bold.

The best performance has been reported by a 3D-CNN based model in [22], but this value corresponds to the max accuracy. When the mean accuracy is considered and the standard deviation is reported, the method in [31] performs best, while the method in [2] has reported a higher mean accuracy, without reporting though the standard deviation. Both DL and HC methods seem to achieve satisfying performance, without though being able to conclude which type of methods perform better given the reported results.

4.3 Experimental Results
For assessing the performance of the proposed framework, the Violent Flows dataset was used. To achieve a justifiable comparison with the methods presented in Table 2, we follow recent literature and perform both the training and the testing of our framework in the five predefined folds of the Violent Flows dataset, using as training data the four subsets, and the remaining one for testing.

For each of the five folds, the accuracy and the loss during training are presented in Figure 3, which shows that our framework performs accurately in each fold. Specifically, the accuracy stabilises after 100 epochs and gradually increases above 95% for the majority of experiments, while the loss, starting from ln(n) = ~ 0.69, n=2 ((violence, non-violence)), decreases significantly in the first epoch, and gradually converges to 0.1. Overall, the training process of the proposed model was carried out in approximately 4 hours using a GPU-based architecture as a processing unit.

Table 3 presents the performance of the proposed framework for each of the five folds, while the mean performance and standard deviation over these five folds are provided in the first row. Executing the model every 16 frames in inference mode, the execution time is estimated 120 ms. The proposed method outperforms state-of-the-art approaches both when mean and max accuracy are reported. Specifically, our method reports max accuracy 99.3% (Fold-4) and outperforms the reported state-of-the-art max accuracy [22]. Furthermore, our method reports mean accuracy 94.54% and is beyond all state-of-the-art methods, except the method proposed by Bilinski et al. [2] where the standard deviation is not reported.

### Table 1: Commonly used datasets in violence detection research.

| Dataset Name     | Description                                      | Year | Data Specification                           | Video Resolution     |
|------------------|--------------------------------------------------|------|----------------------------------------------|----------------------|
| Violent Flows    | Videos collected from YouTube                    | 2012 | 246 videos, 123 Violence/123 Non-violence   | 320x240 pixels, 25fps |
| Hockey Fights    | Videos showing fights in ice hockey rink         | 2011 | 1000 videos, 500 Violence/500 Non-violence  | Varied, ~30fps       |
| Action Movies    | One-to-one fights extracted from movies          | 2011 | 200 videos, 100 Violence/100 Non-violence   | 720x576 pixels, 25fps |
Table 2: Performance of state-of-the-art approaches.

| Acronym/abbreviation | VF dataset (MA+SD) | Type |
|----------------------|--------------------|------|
| HNF [12]             | 56.52 + 0.33       | HC   |
| HNF + BoW [19]       | 57.05 + 0.32       | HC   |
| MoSIFT + BoW [19]    | 57.09 + 0.37       | HC   |
| HOG [12]             | 57.43 + 0.37       | HC   |
| HOG + BoW [19]       | 57.98 + 0.37       | HC   |
| HOF [12]             | 58.53 + 0.32       | HC   |
| HOF + BoW [19]       | 58.71 + 0.12       | HC   |
| LTP [26]             | 71.53 + 0.17       | HC   |
| OViF [10]            | 76.80 + 3.90       | HC   |
| ViF [10]             | 81.30 + 0.21       | HC   |
| GOMOF [30]           | 82.79              | HC   |
| Substantial Derivative [17] | 85.53 + 0.21   | HC   |
| DiMOLIF [14]         | 85.83              | HC   |
| SCOF [11]            | 86.37              | HC   |
| ViF + OViF [7]       | 88.00 + 2.45       | HC   |
| MoWLD + BoW [19]     | 88.16 + 0.19       | HC   |
| MoSIFT + KDE + Sparse Coding [25] | 89.05 + 3.26   | HC   |
| MoWLD + Sparse Coding [29] | 89.38 + 0.13   | HC   |
| PSS [1]              | 89.50 + 0.13       | HC   |
| SSS [23]             | 91.90 + 0.12       | HC   |
| Spatiotemporal Encoder [8] | 92.18 + 3.29  | DL   |
| SSDLSC [27]          | 92.25 + 0.12       | HC   |
| MoWID [28]           | 93.19 + 0.12       | HC   |
| LHOG + LHOF + BoW [31] | 94.31 + 1.65   | HC   |
| STIF V [2]           | 96.40              | HC   |
| Ullah et al. [22]    | 98.00 (max accuracy) | DL   |

Table 3: Performance of the proposed 3D-ResNet-50 method.

| Acronym/abbreviation | Violent Flows (MA+SD) |
|----------------------|-----------------------|
| 3D-ResNet-50 (Proposed) | 94.54 + 4.13         |
| Fold-1                | 90.62                 |
| Fold-2                | 98.63                 |
| Fold-3                | 91.36                 |
| Fold-4                | 99.31                 |
| Fold-5                | 92.76                 |

the proposed framework reports high (4.13) standard deviation, a problem that might be related to the parameters during training.

Some sample results obtained from the Violence Flow dataset are provided in Figure 4. For each frame, we present (where available) the annotation (denoted by “Crowd Analysis”), as Violence or Non-Violence as well as the “Prediction” score for crowd violence detection as estimated by our framework. The “Crowd Analysis” values of “Violence” and “Non-Violence” are colourised as red and green, respectively, whereas the “Prediction” values and the bounding box are colourised gradually using a colour bar (from red to green) where the red colour indicates scenes predicted as being violent with a 100% confidence score. The bottom row depicts an example where the crowd violence in an event gradually increases.

5 CONCLUSIONS

This work presented a crowd analysis framework to detect violence in video streams. The proposed framework relies on a 3D-Convolutional architecture that is trained on the visual cues associated with violent scenes. The framework was evaluated against several state-of-the-art methods using the challenging Violent Flows dataset. Experimental results showed that the proposed framework can recognise violent crowd scenes in (near) real-time and with higher accuracy compared to the current baselines.

ACKNOWLEDGMENTS

This work was supported by the project CONNEXIONs (H2020-786731) funded by the European Commission.
REFERENCES

[1] Belnab Ababaholami-Mohammadabadhi, Ali Zarghami, Mohammadreza Zolfaghari, and Mahdieh Soleymani Baghshah. 2013. Pseudob for probabilistic semi-supervised dictionary learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 192–207.

[2] Piotr Bilinski and Francois Bremond. 2016. Human violence recognition and detection in surveillance videos. In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 30–36.

[3] Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. 2002. Person-on-person violence detection in video data. In Object recognition supported by user interaction for service robots, Vol. 1. IEEE, 433–438.

[4] Zhizhong Dong, Jie Qin, and Yunhong Wang. 2016. Multi-stream deep networks for person to person violence detection in videos. In Chinese Conference on Pattern Recognition. Springer, 517–531.

[5] E Femia, Gunasekaran Manogaran, GN Vivekananda, T Thanjavudivel, S Jeeva, A Ahilan, et al. 2019. Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Computer Networks 151 (2019), 191–200.

[6] Eugene Yujun Fu, Hong Va Leong, Grace Ngai, and Stephen CF Chan. 2017. Automatic fight detection in surveillance videos. International Journal of Pervasive Computing and Communications 13, 2 (2017), 130–156.

[7] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. 2016. Violence detection using oriented violent flows. Image and vision computing 48 (2016), 37–41.

[8] Alex Hanson, Koutilya Prv, Sanjukta Krishnagopal, and Larry Davis. 2018. Bidirectional Convolutional LSTM for the Detection of Violence in Videos. In Proceedings of the European Conference on Computer Vision (ECCV). 0–8.

[9] Kenso Har, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagem? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 6546–6555.

[10] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 1–6.

[11] Jian-Feng Huang and Shui-Li Chen. 2014. Detection of violent crowd behavior based on statistical characteristics of the optical flow. In 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE, 565–569.

[12] Ivan Laptev, Marcel Marxel, Cordelia Schmid, and Benjamin Rosenfeld. 2008. Learning realistic human actions from movies.

[13] Kaelon Lloyd, Paul L Rosin, David Marshall, and Simon C Moore. 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 (2017), 361–371.

[14] Amir Ben Mabrouk and Ezzeddine Zagrouba. 2017. Spatial-temporal feature using optical flow based distribution for violence detection. Pattern Recognition Letters 92 (2017), 62–67.

[15] Vijay Mahadevan, Wenzin Li, Viral Bhalodia, and Nuno Vasonceles. 2018. Anomaly detection in crowded scenes. In 2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 1975–1981.

[16] Javad Mahmoodi and Afkane Salajeghe. 2019. A classification method based on optical flow for violence detection. Expert Systems with Applications 127 (2019), 121–127.

[17] SadeghMohammad, Hamed Kian, Alessandro Perina, and Vittorio Murino. 2015. Violence detection in crowded scenes using substantial derivative. In 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1–6.

[18] Nam Thanh Nguyen, Dinh Q Phung, Svera Venkatesh, and Hung Bui. 2005. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), Vol. 2. IEEE, 955–960.

[19] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno Garcia, and Rahul Sukthankar. 2011. Violence detection in video using computer vision techniques. In International conference on Computer Analysis of Images and Patterns. Springer, 332–339.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[21] Swathikiran Sudhakaran and Oswald Lanz. 2017. Learning to detect violent videos using convolutional long short-term memory. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1–6.

[22] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. 2019. Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors 19, 11 (2019), 2472.

[23] Di Wang, Xiaojin Zhang, Mingyu Fan, and Xuzi Ye. 2016. Semi-supervised dictionary learning via structural sparse preserving. In Thirtieth AAAI Conference on Artificial Intelligence.

[24] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553 (2015).

[25] Long Xu, Chen Gong, Jie Yang, Qiang Wu, and Lixiu Yao. 2014. Violent video detection based on MoSIFT feature and sparse coding. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 3538–3542.

[26] Lahav Yeftet and Lior Wolf. 2009. Local trinary patterns for human action recognition. In 2009 IEEE 12th international conference on computer Vision. IEEE, 492–497.

[27] Tao Zhang, Wenhing Jia, Chen Gong, Jun Sun, and Xiaoning Song. 2018. Semi-supervised dictionary learning via local sparse constraints for violence detection. Pattern recognition letters 107 (2018), 98–104.

[28] Tao Zhang, Wenheng Jia, Xiangjian He, and Jie Yang. 2016. Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Transactions on Circuits and Systems for Video Technology 27, 3 (2016), 696–709.

[29] Tao Zhang, Wenhing Jia, Baogang Yang, Jie Yang, Xiangjian He, and Zhonglong Zheng. 2017. MoWLD: a robust motion image descriptor for violence detection. Multimedia Tools and Applications 76, 1 (2017), 1419–1438.

[30] Tao Zhang, Zhijie Yang, Wenhing Jia, Baogang Yang, Jie Yang, and Xiangjian He. 2016. A new method for violence detection in surveillance scenes. Multimedia Tools and Applications 75, 12 (2016), 7327–7349.

[31] Peipei Zhou, Qinghai Ding, Haibo Luo, and Xinglin Hou. 2018. Violence detection in surveillance video using low-level features. PLoS one 13, 10 (2018), e0203668.