MULTIMODAL VIOLENCE DETECTION IN VIDEOS

Bruno Peixoto[1]*, Bahram Lavi*[1]*, Paolo Bestagini*[2]*, Zanoni Dias*[1]*, and Anderson Rocha*[1]

[1]Institute of Computing, University of Campinas (Unicamp), Campinas, São Paulo, Brazil
[2]Politecnico di Milano, Milano, Italy

ABSTRACT

Effective tools for detection of violence are highly demanded, specially when dealing with video streams. Such tools have a wide range of applications, from forensics and law enforcement to parental control over the ever increasing amount of videos available online. Prior studies showed that deep learning has great potential in detecting violence, but focuses on detecting violence in general, or only specific cases of violent behavior. While the concept of violence is broad and highly subjective, simpler concepts such as fights, explosions, and gunshots, convey the idea of violence while being more objective. Even though different concepts relate to this same broader idea of violence, they differ widely in relation to whether or not they convey the idea of movement, the presence of a specific object, or even if they generate distinctive sounds. In this study, we propose to analyze different concepts related to violence and how to better describe these concepts exploring visual and auditory cues in order to reach a robust method to detect violence.

Index Terms— computer vision, violence classification, deep-learning, multimodal classification, forensic computing

1. INTRODUCTION

Violence detection is an essential application for the issue of video analysis in filtering sensitive media contents. It can provide a useful tool to protect users from being exposed to undesired media from various sources and, in conjunction with video surveillance systems, to detect inappropriate behavior and aid law-enforcement in forensic examination cases. Moreover, it can prevent content from being uploaded to social media, forums or educational platforms. In the same way, it can be used to avoid violent material being shown in specific places such as workplaces and schools. Indeed, early exposure to violent media contents might not be suited for children.

Currently, hundreds of hours of video are uploaded every minute through the Internet and different social media platforms. To handle and analyze them is heavily time consuming. Moreover, the concept of violence is considered very subjective to define and, as such, leads to different interpretations. This makes the development of violent detection methods even harder.

"Automatic" solutions in prior art have been developed to determine violence in videos (*if any*), and this task has definitely attracted much attention. Proof of interest is also clear from the competition "MediaEval Affect Task", which aims to identify violence in movies [1]. In this paper, we aim at addressing the violence detection task by breaking down the subjective concept of violence into more objective concepts: *Blood*, *Cold Arms*, *Explosions*, *Fights*, *Fire*, *Firearms*, *Gunshots*. Breaking down violence into different subjects is a proxy to achieving more accurate and robust performance [2]. This allows us to perform a better investigation on the behavior of different subjects, as each subject of violence has different characteristics over all others. We take into account the concept of *violence* as a single high-level concept to analyze the behavior of different integrating concepts, individually. We then perform a combination of the concepts of violence to identify general violence, and compare the performance on different setups.

The issue of violence detection in video scenes was firstly addressed for the task of action recognition. In this vein, before deep learning based methods, the Bag-of-Visual-Words (BoVW) approaches [3, 4] were a cornerstone in the area. In [3], low-level features obtained by an image descriptor such as Space-Time Interest Points (STIP) [5], were used to predict violence via Support Vector Machines (SVM). In [4], local spatial-temporal features for violence classification were investigated. Clarin et al. [6] addressed the local interest-point approach to detect fights as subjective violence. A novel descriptor was proposed in [7] for real-time crowd violence detection. After the first wave of methods exploiting spatial-temporal interest points methods, deep learning techniques paved the way for more complex solutions (and consequently also better results) for violence detection [8, 9, 10, 11, 12, 13]. In [14], a three-stream deep convolutional neural network (dCNN) approach was proposed for detecting violence for the specific case of person-to-person violence setup. To the best of our knowledge, most of the above mentioned works rely on a specific concept of violence (i.e., fights) without considering the myriad of possible different concepts for violence.

Following a different strategy from previous work in prior art, this work extends upon our former works [15, 16] in terms of the methodology on subjective violence detection. We aim at developing a fusion model over visual and audio feature representations. In a typical detection setup, some concepts might convey the idea of movement, while some have characteristic sounds associated to them. Therefore, we analyze ways of combining various concepts with different characteristics in order to detect the more complex (and subjective) concept of violence.

The most related work to our method proposed herein was proposed in [17], in which an approach based on visual and audio clues addresses the issue of cross-modality to tackle the violence scene detection in videos. The method consists on extracting audio features by collecting bag-of-audio-words, and utilizing a dCNN technique to obtain visual features. Both features are further concatenated in a late-fusion stage. Finally, a standard classifier is also applied to determine an occurred violence scene in video. Differently, our work presents a distinct methodology on the fusion step in the late stage to achieve more correlations between visual and auditory features on violence concepts and also exploits tailored deep description methods for each sub-concept of interest.

This work was funded by the São Paulo Research Foundation (FAPESP) under grants #2017/12646-3 and #2018/05668-3 and the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) under the DeepEyes grant.

This paper is organized as follows. In Sec. 2, we discuss our approach based on visual-auditory features for violence detection; we also explain our methodology of fusing two independent feature sets by utilizing different kinds of neural network techniques. In Sec. 3, we evaluate the effectiveness of our method on *MediaEval-2013-VSD* data set. Finally, we conclude the paper and discuss directions for future work in Sec. 4.

2. PROPOSED METHODOLOGY

The solution we propose decomposes violence detection in first detecting k more objective sub-concepts that convey the idea o violence. In our study, we use $k = 7$, more specifically, the concepts of *Blood*, *Cold Arms*, *Explosions*, *Fights*, *Fire*, *Firearms*, and *Gunshots*. For each sub-concept, we train specific neural networks: first to analyze its visual characteristics, then to analyze its auditory features. Then, we combine both features to obtain a better understanding of the sub-concept. We repeat this step for all k concepts. As a final step, we use a fusion network to combine all concepts (described through auditory and visual features) to detect the more general concept of violence. Fig 1 illustrates this methodology pipeline.

2.1. Visual-based violence detection

We sought to capture specific features for each sub-concept integrating the broader concept of violence. Some concepts related to violence convey the idea of movement, such as fights and explosions. In [15], we studied the difference between two convolutional networks that incorporate the concept of time in their formulation: C3D and LSTMs, and used them to analyze the best way to detect each sub-concept. In this work, we extend upon those early investigations and also consider using 2D networks with inputs that represent movement, such as optical flow and optical acceleration.

Pre-processing. For each video, we extract all frames individually. To represent the movement, we calculate the optical flow between frames, as well as the Farneback optical acceleration, defined in [18] as the difference between two consecutive optical flows between three adjacent frames. In this way we have three types of visual inputs: raw frames, optical flow, and optical acceleration.

Convolutional Neural Networks. We extend upon the work in [15] and used the architectures defined therein for the C3D and the LSTM combined with a CNN approach to receive the types of inputs discussed above. We also use Inception v4 [19] pre-trained with Imagenet [20] and finetune it for the target dataset with all types of inputs.

2.2. Audio-based violence detection

As the concept of violence is too subjective, some sub-concepts (e.g., fights) might benefit from an audio characterization other than just visual. Typically, audio clips contain noise, i.e., background sound and people talking. Therefore, we decided to extract features that are robust to noise and background clutter, rather than processing the raw audio waveform. In the following subsections, we describe the adopted audio feature representation and prediction model for the violence detection problem.

Feature extraction. For this task, we adopt a two-step approach. First, we generate feature vectors by leveraging four standard audio feature extractor methods. We then apply statistical methods on the features generated from the first step. This approach showed to be more robust than directly using the raw input waveform in a neural network because of the clutter and noise mentioned above.

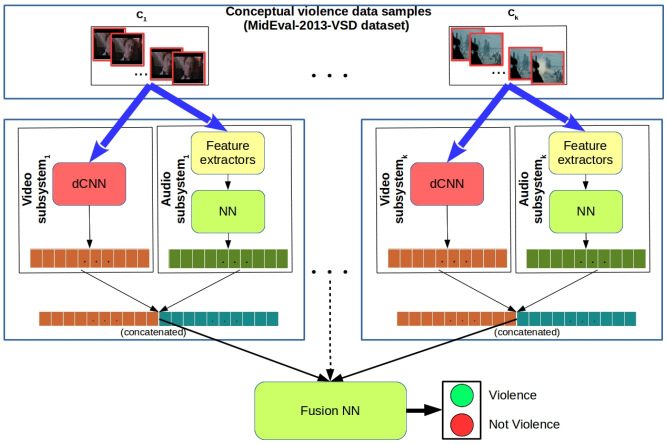

Fig. 1. Pipeline of the proposed visual-auditory feature fusion network. Videos are described using a dCNN, whereas audio features are processed with a shallow one in early-stage. In this pipeline, different sub-concepts $(C_1, \ldots, C_k)$ are treated in parallel. Finally, visual and audio features are combined into a feature vector and used to train a fusion network responsible for the final answer.

Formally, given an audio clip $x(t)$, we split it into a series of I temporal windows from which we extract a set of feature vectors by computing Mel-Frequency Cepstral Coefficients (MFCCs) [21], Chroma Short-Time Fourier Transform (C-STFT) [22], Mel- Spectrogram (MS) [21], and Spectral Contrast (SC) [23] features as defined in [24]. The feature set associated to the i-th time window is defined as

$$f_i = [f_i^{\text{mfcc}},\ f_i^{\text{cstft}},\ f_i^{\text{ms}},\ f_i^{\text{sc}}], \tag{1}$$

which is the concatenation of feature vectors obtained from different extractors.

According to [25], to extract a more discriminant feature vector that represents the overall audio excerpt $x(t)$, we apply four temporal statistics to the I extracted feature vectors f_i. This provides an additional set of information for the subsequent learning stage. Given a set of I feature vectors f_i, we compute the per-feature average, standard deviation, maximum, and minimum value as

$$f^{\mu} = \frac{1}{I}\sum_{i=1}^{J} f_i, \tag{2}$$

$$f^{\sigma} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(f_i - f^{\mu})^2}, \tag{3}$$

$$f^{\text{M}} = \max_{i \in I} f_i, \tag{4}$$

$$f^{\text{m}} = \min_{i \in I} f_i, \tag{5}$$

where all operations are applied element-wise. The final feature vector is presented by concatenating all the statistical features as

$$f^{\text{tot}} = [f_i^{\mu},\ f_i^{\sigma},\ f_i^{\text{M}},\ f_i^{\text{m}}], \tag{6}$$

where f^{tot} represents a feature vector of $4 \times 4 = 16$ elements, and is used as a compact feature representation over the four sets of audio features adopted. The obtained feature vector can be used to reduce both computational time and memory footprint.

Learning step. In order to learn violence concepts characteristics, we train a supervised classifier based on a shallow neural network fed with the extracted audio features.

Even though we experimented with different network designs (including deep ones), we decided to adopt a shallow neural network (NN) model to reduce complexity. Indeed, the designed NN has a single hidden layer, in which the number of neurons is equivalent to the length of the feature vector f^{tot}. As a matter of fact, with such a small feature vector (i.e., 16 elements), deeper networks did not provide much better results.

The network is trained to detect a specific kind of violence (i.e., *Blood*, *Cold Arms*, *Explosions*, *Fights*, *Fire*, *Firearms*, *Gunshots*), rather than general violence. In other words, we treat the audio violent detection problem as a 2-class classification problem by training a different binary classifier for each violence concept. A Softmax layer is deployed at the end of the network to determine whether violence occurred within the audio clip or not.

2.3. Visual-Auditory fusion network

In [15], we designed a solution to learn spatial-temporal information only on the subject of violence over different violence-related concepts. We showed that the method can independently learn the final decision from the output weights obtained from the binary classification networks. Basically, the networks are used in a late stage, and each network is trained independently on the presented feature vectors. This solution allowed us to achieve a better trade-off between efficiency and performance.

We use the same fusion network idea to leverage sub-concepts related to violence but now including raw-based motion explorations such as optical acceleration and optical flow as well as auditory features not explored before. The network takes a feature vector as input and outputs the probability for presence of violence within the audio clip. The final feature representation is obtained by concatenation of the different visual and audio features. In other words, we merge feature vectors obtained with audio-visual detectors trained on specific kind of violence, in order to detect a general presence of violence. We then pass the feature vector through a standard *Min-Max* normalization step.

3. EXPERIMENTS

For the experiments, we selected the MediaEval-2013-VSD dataset [26], a staple in the literature for this problem. This dataset consists of 25 Hollywood movies and provides shot segmentation from the movies, manually annotated on whether or not physical violence occurred within scenes. The definition of violence used by the competition is that a scene is violent if "one would not let an eight-year old child see". The training set includes 18 movies while the test set comprises 7 movies. Among all scenes, 20% of them have been annotated as violent. The data-set also provides annotations for sub-concepts related to violence (e.g., blood, fights, etc.), though these are only available for the training set.

For each concept, our experiments sought to find the best combination of input and architecture to classify scenes as containing or not the analyzed concept. For the visual aspect, we tested three neural network architectures: Inception v4, C3D, and a CNN-LSTM; as well as three different types of inputs: raw frames, optical flow, and optical acceleration. While the latter two architectures and inputs were chosen aiming at capturing movement through time for the concepts that convey this type of information, Inception v4 and the raw frames were chosen to both serve as a base-case and to classify scenes where movement is not intrinsically involved, such as blood and cold arms.

Implementation details. All networks were implemented using Keras DL library in Python and Tensorflow and ran on an NVIDIA GeForce GTX 1080 Ti GPU. Each network was trained for a binary classification of each individual concept, for both visual and audio concepts. The C3D and CNN-LSTM models details were kept the same as in [15]. For the Inception architecture, we used Root Mean Square Propagation (RMSProp) algorithm to train, with batches of 64 images of size 299×299. Due to the large number of negative samples on the training dataset, we randomly selected a fraction to match the number of positive samples for all our networks. The optical flow experiments were done analyzing consecutive frames, frames that are 5 frames apart, and 10 frames apart. The optical acceleration experiments were done using consecutive optical flows from consecutive frames and from frames that are 5 frames apart. The audio for each video clip was extracted separately and we followed the frame annotation to construct their respective labels for time intervals. For classification, we used two different approaches: a random forest classifier set up with 10 trees; and a shallow neural network described in [15] with a softmax layer deployed at the end of the network. For the loss function, we choose *Binary Cross-entropy* while for optimizer, we adopted Adadelta [27].

Visual features results. We found out that each violence-related sub-concept was better classified with varying types of inputs, whereas Inception v4 performed better than C3D and CNN-LSTM in all cases. In general, our results with optical flow were better when calculated between one frame and the fifth next frame, while the optical acceleration provided better results when comparing differences between two optical flows of three consecutive frames. Table 1 shows results for these inputs with all network architectures for comparison.

Optical acceleration was able to capture the concepts of firearms better than the raw images, even if it is essentially a static object. This concept close relation to gunshots can lead the network to better classify firearms when analyzing its optical acceleration.

Another interesting result is the better classification accuracy for explosions with raw frames rather than a motion-based input. Explosion patterns can vary widely in a Hollywood movie setting. They can be related not only to fire, but with dirt explosions, big and small ones. The significant difference in expansion and impact, and the visual cues for explosions can be better distinguished by the raw frames approach rather than its optical flow or acceleration.

The optical flow still was the better descriptor for fights. As optical acceleration describes sudden changes in pixels between frames, the relatively slower speed of fights could be better captured by the optical flow alone.

The general concept of violence, though, had a better classification accuracy when described by the optical acceleration. The better performance of this descriptor on the firearms and gunshots concepts may indicate that these concepts have a higher correlation with the general concept of violence considered in this dataset.

We combined the best results for each concept through the described fusion network to ultimately classify violence. As the best results were all from the same network architecture (Inception), we extracted the features from its last fully-connected layer, and used them as inputs for the fusion network. In Figure 2, we have the best results represented. The first bars indicate the best results with the visual experiments for each concept and for the fusion. With the proposed fusion method, we obtained a 6% increase in classification accuracy compared to classifying the general concept of violence (from 72.8% accuracy using only audio to 78.5%) when combining visual and audio features, the best result for the adopted dataset to date. The fusion of only visual concepts leads to 74.4%.

	Raw Frames			Optical Flow			Optical Acceleration		
	Inception v4	C3D	CNN-LSTM	Inception v4	C3D	CNN-LSTM	Inception v4	C3D	CNN-LSTM
Blood	**74.2**	58.0	57.2	68.3	59.2	60.2	58.2	60.2	58.2
Cold Arms	**81.6**	58.3	54.2	61.9	66.5	66.2	76.5	75.3	69.0
Explosions	**79.4**	77.1	61.4	77.8	66.4	64.8	70.6	73.0	68.1
Fights	73.1	70.5	53.7	**77.2**	68.0	66.9	74.3	65.4	61.7
Fire	70.1	60.2	55.6	68.1	60.3	61.3	**71.2**	64.9	61.9
Firearms	60.8	61.0	60.3	62.3	63.2	65.0	**66.8**	66.5	62.3
Gunshots	69.3	65.3	56.8	63.6	62.6	64.1	**73.1**	68.6	66.8
Violence	66.7	62.3	55.9	65.0	58.1	58.6	58.7	**68.3**	63.6
Fusion	**74.4**	67.3	63.3	68.2	67.2	64.8	72.8	69.2	64.2

Table 1. Classification accuracy for visual features. All values are indicated in accuracy percent. All seven concepts were trained and tested with the same subsets of movies. The 'violence' concept refers to the MediaEval VSD definition of violent scene. 'Fusion' does not include the 'violence' concept.

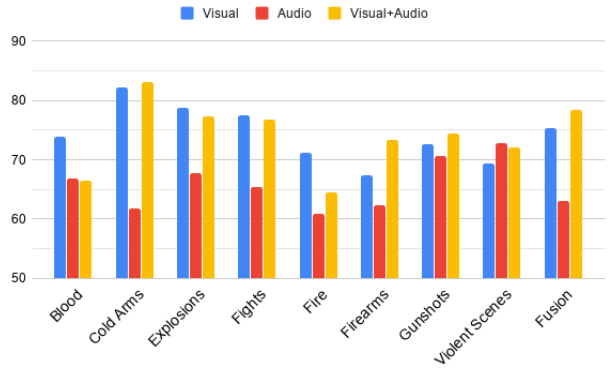

Fig. 2. Normalized accuracy for each of the best results in each concept considered. The first column of each concept is the result for the visual features, the second for the audio features, and the third for the fusion of both feature sets.

Audio features results. Table 2 shows results for audio. Our network achieved better results for all concepts, ranging from a 7.2% increase in classification accuracy for firearms to 12.8% in explosions. The baseline here is a random forest classifier (RF) receiving the audio features directly with a setup of 10 trees. Audio features by themselves are slightly inferior to visual features.

Computing statistics from each feature vector, to complement their information, yields a significant increase in accuracy, reaching a 5% increase with the concept of blood. This shows the importance of aggregating more audio information in the feature vector along with their extracted features.

We expected concepts related to loud and distinctive sounds to perform better than their visual counterparts. Though with gunshots we had close results, up to 2% lower accuracy, every concept performed worse on the audio alone, including explosions, which is almost 11% lower in classification accuracy. Classifying violent scenes in general, on the other hand, had 3.4% better accuracy with audio compared to the classifier for visual information, and 9.8% better accuracy than our concepts fusion. As we are working with Hollywood movies, the sound cues for more violent and action scenes are very characteristic and could explain why classifying the general concept via audio is a better approach for this scenario.

Even so, in a video, the visual part also plays an important role in determining a violent scene, so in order to explore how they complement each other, we used our same fusion network for the concepts to classify violence combining visual and audio features.

We used our best results for each concept, both in visual and audio. Best results for the visual part were obtained with the Inception v4 architecture, and the best results for the audio part were obtained

	Random Forest	NN	Statistics + NN
Blood	52.3	61.8	**66.9**
Cold Arms	51.7	59.2	**61.9**
Explosions	51.0	63.8	**67.8**
Fights	57.3	65.0	**65.3**
Fire	50.5	59.3	**61.0**
Firearms	53.3	60.5	**62.4**
Gunshots	58.0	70.0	**70.7**
Violence	57.4	71.6	**72.8**
Fusion	51.1	62.3	**63.0**

Table 2. Classification accuracy for audio. All values are indicated in accuracy percent. All seven concepts were trained and tested with the same subsets of movies. The 'violence' concept refers to the MediaEval VSD definition of violent scene. 'Fusion' does not include the 'violence' concept.

with the proposed shallow network classifier. We then extracted features for each concept in the visual part and concatenated with its respective feature in the audio part to feed as input for our fusion network. A standard *MinMax* normalization step was performed to keep both audio and visual features in the same range. With each concept containing information from both the visual and audio part of the video, we reached an accuracy of 78.46% with the fusion for violence, compared to our best result of 72.08% accuracy for the general concept – also achieved by joining visual and audio features. This result is better than those obtained by using only visual or only audio information, further suggesting that these are complementary to each other.

4. CONCLUSIONS AND FUTURE WORK

Classifying violence in videos is a challenging problem. Not only due to its own subjective nature, but also to the wide range of features to analyze and how visual and audio interact to convey the concept of violence. In some scenarios, audio may not be relevant (e.g., in closed-circuit security cameras where no sound is captured). There are some setups where not all kinds of concepts analyzed in this work are relevant. However our method is robust and modular enough to adapt to each case and try to better define what is a violent scene not by defining violence itself, but capturing more objective concepts related to violence and fusing them to better classify a scene.

In this vein, one important future work would be collecting specific datasets to train each individual concept. For example, if we can train a solution to classify fights or any other violent concept with a specialized dataset, we can plug the feature vectors from each of these specific networks into the fusion network and have a more general method to detect violence, that is even more robust and independent of datasets. Another important way forward is to explore additional concepts related to violence.

5. REFERENCES

[1] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani, "The mediaeval 2012 affect task: violent scenes detection," in *Working Notes Proceedings of the MediaEval 2012 Workshop*, 2012.

[2] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, "Semantic context detection based on hierarchical audio models," in *in ACM SIGMM International workshop on Multimedia information retrieval*. ACM, 2003, pp. 109–115.

[3] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*. Springer, 2011, pp. 332–339.

[4] Fillipe Souza, Eduardo Valle, Guillermo Chávez, and Arnaldo de A Araújo, "Color-aware local spatiotemporal features for action recognition," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2011, pp. 248–255.

[5] Ivan Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[6] C Clarin, J Dionisio, M Echavez, and P Naval, "Dove: Detection of movie violence using motion intensity analysis on skin and blood," *Phillipine Comput. Sci. Congr.*, vol. 6, pp. 150–156, 2005.

[7] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2012, pp. 1–6.

[8] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang, "Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning.," in *MediaEval*, 2015.

[9] Vu Lam, Sang Phan Le, Duy-Dinh Le, Shin'ichi Satoh, and Duc Anh Duong, "Nii-uit at mediaeval 2015 affective impact of movies task.," in *MediaEval*, 2015.

[10] Ionut Mironica, Bogdan Ionescu, Mats Sjöberg, Markus Schedl, and Marcin Skowron, "Rfa at mediaeval 2015 affective impact of movies task: A multimodal approach.," in *MediaEval*, 2015.

[11] P Marin Vlastelica, Sergey Hayrapetyan, Makarand Tapaswi, and Rainer Stiefelhagen, "Kit at mediaeval 2015-evaluating visual cues for affective impact of movies task.," in *MediaEval*, 2015.

[12] Yun Yi, Hanli Wang, Bowen Zhang, and Jian Yu, "Mic-tju in mediaeval 2015 affective impact of movies task.," in *MediaEval*, 2015.

[13] Qing Xia, Ping Zhang, JingJing Wang, Ming Tian, and Chun Fei, "Real time violence detection based on deep spatio-temporal features," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 157–165.

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

[15] Bruno Peixoto, Bahram Lavi, João Paulo Pereira Martin, Sandra Avila, Zanoni Dias, and Anderson Rocha, "Toward subjective violence detection in videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8276–8280.

[16] Bruno Malveira Peixoto, Sandra Avila, Zanoni Dias, and Anderson Rocha, "Breaking down violence: A deep-learning strategy to model and classify violence in videos," in *Proceedings of the International Conference on Availability, Reliability and Security*. ACM, 2018, p. 50.

[17] Qin Jin, Xirong Li, Haibing Cao, Yujia Huo, Shuai Liao, Gang Yang, and Jieping Xu, "Rucmm at mediaeval 2015 affective impact of movies task: Fusion of audio and visual cues.," in *MediaEval*, 2015.

[18] A. Edison and J. C. V., "Optical acceleration for motion description in videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1642–1650.

[19] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[21] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra, "Essentia: an open-source library for sound and music analysis," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 855–858.

[22] Dan Ellis, "Chroma feature analysis and synthesis," *Resources of Laboratory for the Recognition and Organization of Speech and Audio-LabROSA*, 2007.

[23] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*. IEEE, 2002, vol. 1, pp. 113–116.

[24] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[25] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Automatic reliability estimation for speech audio surveillance recordings," in *The IEEE International Workshop on Information Forensics and Security*. WIFS, 2019.

[26] Claire-Helene Demarty, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, and Cedric Penet, "Benchmarking violent scenes detection in movies," in *Content-Based Multimedia Indexing (CBMI), International Workshop*. IEEE, 2014, pp. 1–6.

[27] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.