UNIVERSITY OF CAMPINAS (UNICAMP)

INSTITUTE OF COMPUTING

SPECIFIC QUALIFYING EXAM

August 3, 2018

Violence Detection Through Deep Learning

Candidate: Bruno Malveira Peixoto

Advisor: Anderson Rocha

Co-Advisor: Zanoni Dias

Contents

1 Introduction **1**

2 State of the Art **2**

2.1 MediaEval Initiative 3

2.2 MediaEval 2015 violent scenes detection 4

3 Related Concepts **6**

3.1 Deep Learning 6

3.2 Convolutional Neural Networks 6

4 Objectives and Contributions **7**

5 Proposed Methodology **9**

5.1 Representing Violence 9

5.1.1 Defining Violence Concepts 9

5.1.2 Relative Violence 11

5.2 Incorporating Temporal Information 12

5.2.1 Temporal Features 12

5.2.2 Neural Networks and Finite State Machines 16

5.3 Localizing Violence 17

5.4 Datasets and Validation Metrics 19

6 Cronogram **21**

7 Current Status **22**

7.1 Classifying Violence 22

7.2 Relevance of individual concepts 25

Abstract

Detecting violence in videos through automatic means is significant for law enforcement and analysis of surveillance cameras with the intent of maintaining public safety. Moreover, it may be a great tool for protecting children from accessing inappropriate content and help parents make a better informed decision about what their kids should watch. However, this is a challenging problem since the very definition of violence is broad and highly subjective. Hence, detecting such nuances from videos with no human supervision is not only technical, but also a conceptual problem.

With this in mind, in this work we will explore how to better describe the idea of violence for a convolutional neural network. Initially by breaking it into more objective and concrete related concepts, such as fights, explosions, blood, etc, for later fusing them in a meta-classification to describe violence. We will also explore ways to represent time-based events for the network, since many violent acts are described in term of movement. And finally we will explore how to localize violent events, since many video streams do not contain only violence, but is a mixture of violent and non-violent scenes.

1 Introduction

Proper content filtering of sensitive media is an important issue nowadays, for its many applications: it can be used in conjunction with surveillance cameras to detect inappropriate behavior; aiding parental control by rating videos of streaming services; protecting users from receiving undesired media via messaging applications; blocking content from being uploaded to websites such as social networks, forums or educational platforms; or preventing it from being shown in specific places such as schools and workplaces.

One kind of sensitive media of particular interest is violence. With hundreds of hours of video uploaded every minute through the Internet and becoming a part of everyday life, comes many violent scenes not suited for people, especially for children. Hence, the demand for automated tagging and rating systems that detect these violent scenes is increasing. Violence is also a public health problem, that demands constant effort from authorities to provide the population with safer public places. One such effort is the investigation of automatic violence detection on surveillance camera feeds, aiming to support faster and more accurate official responses in cases of crime occurrences and dangerous situations.

Although there has been extensive work regarding action recognition in general [6, 50, 56–58], violence detection in videos is still a growing field, with some works using descriptors developed for general actions and tweaking them in order to observe sudden changes in motion or make use of distinctive sounds that represent specific kinds of violence [7, 39]. Some works combine several descriptors in a bag of visual words (BoVW) approach [10, 15, 40]. Although they show the importance of using spatio-temporal information to detect violence, fusing audio features with motion from the video stream, they still depend on manually constructed descriptors to tackle a highly subjective concept.

Machine-learning techniques, specially convolutional neural networks, have been demonstrating excellent results in image and video classification [27, 52, 53, 55]. In many different challenges and datasets, these networks greatly outperformed the previous approaches, including the BoVW-based solutions. As such, it is natural that these networks have been studied and adapted in the action recognition field [20, 24, 32, 52], combining the spatio-temporal information. Different architectures are used, each of which with its own way of conveying the motion information, resulting in various features for the classification task.

Recently, these convolutional network have been used in the specific task of violence detection, showing promising results [9, 23, 35, 41, 51, 60], but still far from the groundbreaking results in the image classifications tasks. Usually these methods combine features extracted from the neural network with several hand-crafted ones in hope to solve the problem without knowing its impacts.

This work aims to develop deep-learning techniques to detect violent scenes in videos, using spatio-temporal information to classify and localize such scenes. We attack this problem in three fronts: The first is to grasp a better understanding of the definition of violence, and use this to select relevant features to a more robust representation of the problem. The second is to incorporate the time element in this representation, as motion and the context of an event are important to determine if a specific video is violent. The third is to expand the classification problem and localize violent scenes in a video stream, identifying exactly when the violent scene starts and ends.

2 State of the Art

Research on violence detection in videos is scarce, compared with other video analysis problems. Wang et al. [57] developed an action recognition method using dense trajectories. They introduced a descriptor that computes motion boundaries from the optical flow information in order to find a trajectory for the motion. This approach has been further improved, by correcting camera motion [58] and using Fisher Vector encoding [46] (in [43]), becoming state-of-the-art in the field. This approach however, is dependent on numerous hand-crafted descriptors being put together in a bag of features, leaving the classifier to decide how they interact to describe each kind of action. In order to automatically extract features from videos, Ji et al. [21] proposed a 3D convolutional neural network for action recognition, stacking multiple contiguous frames of video and using as input for the network, capturing the motion information. This method achieved similar results to those using dense trajectories, but computing over frames with a 4-times lower resolution.

In the specific area of violence detection, however, most of the work is based on low-level features. The usual approach involves the extraction of features around interest points, such as optical flows, gradients, intensities or other local features. One of the earlier works is by Nam et al. [39], which proposes threshold values for auditory and visual features. For the auditory features, they considered the amplitude and energy of the audio signal, as well as sudden changes in the overall entropy. As visual features, they calculate the dynamic activity in order to identify quick movements as well as pixel color thresholds for blood detection.

Cheng et al. [7] proposed an auditory approach to detecting basic audio events such as gunshots, explosions, engines, car breakings, etc. They trained Hidden Markov Models (HMM) to recognize and target sound events and then model the correlations among several events with Gaussian mixture models in order to extract more complex semantic contexts.

These earlier methods though, relied on specific events and looked for each one individually. One method that generalizes and tries to identify violence through motion is proposed by Bermejo et al. [40]. They exploited a Bag of Visual Word (BoVW) approach, using low-level features such as Space-Time Interest Points (STIP) and Motion SIFT (MoSIFT) [6], which is an extension of the SIFT [34] image descriptor for video, adding a histogram of optical flows representing local motion. These features were then used to estabilish a bag of words for each video that was classified via Support Vector Machine (SVM). Souza et al. [10] also used a BoVW-based approach with local spatio-temporal features to classify video shots as violent or not. Several STIP-detected descriptions were hard-coded to make use of the spatio-temporal information and compose a bag of features for each shot, and a linear SVM was then trained to classify the videos, achieving comparatively better results. These approaches highlight the importance of using motion and space-temporal features in violence detection.

These works, however, reported results on different datasets, with different metrics. Moreover, the different concepts of violence prevent us from directly compare the existing mehods. Such problems further sparked the MediaEval initiative as a form of standardizing validation in the field.

2.1 MediaEval Initiative

The MediaEval Benchmarking Initiative for Multimedia Evaluation [31] provided the scientific community with a unified violence dataset, with a common groundtruth, which reflected a clear understanding of the concept of violence and standardized evaluation protocols. Since then, a gamut of works have been proposed in the literature, aiming at attending the Violent Scenes Detection (VSD) task.

In its first years, the task challenged participants to classify pre-segmented video shots as violent or not. A common trend among the VSD task attendants was to combine of visual and auditory features, similar to previous works in the related literature. Only as recently as 2015, various teams have started venturing with deep convolutional neural networks, automating the process of extracting features and achieving promising results.

2.2 MediaEval 2015 violent scenes detection

Vlastelica P. et al. [41] proposed a method that used multiple visual features and linear SVM classifiers. In their work, the BVLC Reference CaffeNet model provided with the Caffe framework [22] was used to extract CNN features, using the output of the last fully-connected (FC) layer and training a linear SVM on the 4096 dimensional features for the images from video clips. Another feature used was the Improved Dense Trajectory (IDT), which is a descriptor used in action recognition [58]. To represent the motion information of video content, the IDT approach combines several descriptors for each trajectory, mainly the Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) an Motion Boundary Histogram (MBH). These features are then projected via Principal Component Analysis (PCA) to reduce their dimensionality and encoded using a Fisher Vector model [46].

Yi et al. [60] combined CNN features with various additional features, such as Dense SIFT, Hue-Saturation Histogram and an IDT approach with the aid of a new proposed Trajectory Based Covariance descriptor [56], using also audio features, extracting the Mel-Frequency Cepstral Coefficients (MFCC). The CNN-based features were extracted using the architecture of the CNN-M-2048 in [5], using the frames of the videos in the violence detection task to fine-tune the first five layers and retrain the last three.

Dai et al. [9] trained a CNN model based on AlexNet [27] with a subset of ImageNet classes manually picked to be related to violence and extracted features for both static frames and motion optical flows. For the static frames, a pre-trained CNN model on the ImageNet Challenge dataset was used and the last three fully-connected layers were used as features. For the motion information, a CNN model was trained to take staked optical flows as input and the last FC layer used as features. After the feature extraction, a Long Short-Term Memory (LSTM) model was applied to further model the long-term dynamic information. Conventional features, such as IDT, Space-Time Interest Points (STIP) and MFCC were also used and the classification was done via

SVM.

Table 1 shows all teams that used CNN to some extent, either for the violence detection problem by itself or in conjunction with other conventional features, such as IDT, STIP, SIFT and MFCC. The results were compiled using the percentage of the reported mean average precision (MAP), which was the official performance measure in the 2015 MediaEval Violent Scenes Detection Task.

Table 1: Results for the violent scenes detection of teams that used CNNs in the 2015 MediaEval Violent Scenes Detection Task. The official performance measure is the mean average precision (MAP), shown here in percentages.

Team	**CNN**	**Non-CNN Features**	**CNN+Others**
Fudan-Huawei [9]	23.5	16.5	27.0
MIC-TJU [60]	17.4	21.8	28.5
RFA [35]	14.2	7.7	8.2
RUCMM [23]	11.8	10.6	21.6
KIT [41]	10.2	8.6	12.9
NII-UIT [30]	-	20.8	26.8
UMons [51]	9.67	9.56	-
TCS-ILAB [4]	-	6.4	-
ICL-TUM-PASSAU [4]	-	14.9	-
RECOD [38]	-	11.4	-

These initial deep-learning approaches usually combine the features extracted from the neural network with other hand-crafted features, in an attempt to improve their results without knowing exactly how each one of them is contributing. The NII-UIT team [30] achieved their best results combining auditory, image and motion features with various different layers extracted from the CNN, plus non-CNN features from the past year [28] as external features, but no explanation was provided as to why fusing so many different features worked better. In our research we will investigate the problem and its nuances, finding which features are relevant to reach an understanding of the definition of violence.

3 Related Concepts

To extract meaning from a set of images, computers need to recognize patterns and use them to identify different concepts. In this section we explore some of the tools we use in this project to make this possible.

3.1 Deep Learning

Deep Learning is a method of machine learning that uses multiple processing layers to learn non-linear representations of data in high levels of abstraction. Various architectures such as Deep Neural Networks, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied in computer vision and natural language processing fields showing state-of-the-art results in various tasks.

The main advantage of these models is the ability to generate features automatically from the available data, allowing pattern recognition systems to rely less on manually-built heuristics [33]. Even though Convolutional Neural Networks have been showing excellent performance in handwritten digit classification and face recognition tasks since the late 1990s, in recent years CNNs have shown outstanding performance on more challenging visual classification tasks, most notably with Krizhevsky et al. [27] winning the 2012 ImageNet classificaton benchmark with their CNN model achieving an error rate of 16.4% compared to the second place result of 26.1%.

Zeiler et al. [61] noted some factors responsible for this improvement as the availability of much larger training sets, increasing considerably the number of labeled data; more powerful GPU implementations, making the training of very large models practical; and better model regularization strategies, such as Dropout [19], to prevent overfitting.

3.2 Convolutional Neural Networks

A CNN is a type of Neural Network inspired by the animal visual cortex [33]. Individual neurons respond to stimuli in a restricted region known as receptive field, and the receptive fields of different neurons partially overlap with one another, creating a visual field. The response of an individual neuron can be approximated mathematically by a convolution operation. This way, by stacking multiple layers of neurons, each of which responsible to capture information on a small region and feed it forward, leads to filters that become increasingly global. This is what is called a Feed Forward Neural Network. In the end, a classification layer outputs the results and, in the case of training, it can change slightly the weights of the functions of the last layers, trying to minimize the error of the classification. This process is called backpropagation, as it pulls the error backwards through the network, instead of retraining it from the start with these new adjustments.

In 2015, Szegedy et al. [55] presented GoogLeNet, a deep CNN in which all filters in the architecture are learned and the layers repeated several times, leading to a 22-layer deep network that obtained excellet performance in the 2014 ImageNet classification benchmark [48]. This challenge involved the classification of images into one of 1000 possible categories, and GoogLeNet achieved an error rate of 6.67%, compared to the second place result of 7.32%.

4 Objectives and Contributions

There are various different approaches to detect violence in videos, and although deep learning techniques show promising results, they are not thoroughly explored and their performance can be greatly increased. Many use deep learning as a complementary tool to extract features for the general concept of violence, fusing them with several classic hand-crafted ones, each one trying to do the same by itself, but there is no investigation of how each one is contributing to the solution or if it is really necessary.

Thus, this project has the following objectives:

- Explore and find a robust representation of the concept of violence. Being such a subjective concept, there are many different interpretations and definitions of what is a violent scene. While sensitive media as a whole is bound to subjectivity, violence is a very broad concept with no definitive boundaries. In the 2015 MediaEval challenge, violent videos were annotated using the definition that violence is content that “one would not let an 8-year old child see in a movie because it contains physical violence” [31]. Just by this definition we can already note the subjectivity it implies, as one person might find a specific scene not suited for an 8-year old, but another person might find otherwise. One possible approach is to break down specific types of violence, such as fights, explosions, gunshots, knife cuts, etc. and then bring all these concepts together to define violence as a whole. By using deep learning formulations and techniques, we can extract different kinds of features for each violence concept, since each one has its own audio and visual characteristics, and detect not only if a video contains violence, but which kinds, in a more granular approach to the problem.

- Incorporate temporal information in the Neural Network. A violent scene often happens within a pre-existing non-violent context. A still frame of a ball of fire might be part of an artistic close-up of the sun or the peak of the explosion of a bomb in the middle of a city. To reliably detect violence, one must consider the sequence of frames as a whole. Recent approaches use many different static features in conjunction with motion information but each working separately to describe a general violence event and finding the best way to fuse these features within each violence concept representation is another objective of this project, incorporating temporal features into a robust and discriminative representation of violence.

- Not only detecting violence is important, but also localizing it in a video stream, specially in forensic setups in which, for example, a surveillance video has violent scenes but a precise timestamp is not available. In this case, only a specific interval of time is of interest, and being able to automatically localize when potential violence snnipets occur can significantly help human experts. Many different difficulties arise from this problem, such as how to properly segment the video, checking whether a key frame is relevant, defining the starting and end points of the scene, and so on. All of these explorations need to take into consideration the different types of violence and how they define which parts of the video could be considered violent or not.

From these objectives, the main contributions we expect to achieve with this project are:

- A clear and robust representation of violence, whether it is needed to specify each different kind with its own features and descriptions or being possible to unify the concept and translate it to distinctive features making it possible to detect any type of violence.

- A realible way to detect violence in videos with data-driven techniques incorporating spatio-temporal features. The most recent solutions are far from being as good as image classifiers that showed the potential of deep neural networks for other visual-related problems, and another contribution is the adaptation of these networks to use temporal features in violence detection.

- Localizing violence excerpts in videos is still uncharted territory with very limited works in the literature. Therefore, as a final contribution of this research, we expect to design and develop methods to pinpoint the timestamp where there might be a violent scene in a video stream.

5 Proposed Methodology

Our proposal for this research is that using deep learning, we can automatically identify features that represent violence and use them to detect and localize violence in videos. Our Methodology is divided into three main fronts: i) Finding a reliable representation for detecting violence; ii) Incorporating temporal features into this representation; and iii) Localizing violence in a video stream.

5.1 Representing Violence

The best deep-learning solutions work with very clear and concrete concepts, such as well-defined objects, facial expressions and specific actions.

The concept of violence, on the other hand, is subjective and complex, raising the challenge on how to represent it reliably to a neural network.

Thus, our proposal for this research is that the concept of violence can be broken down into more clear and concrete concepts. With these, it is possible to aggregate specific features to detect a broader concept such as violence.

5.1.1 Defining Violence Concepts

The idea to break down violence into more specific concepts is not entirely new. Back in 2003, Cheng et al. [7] identified audio signatures of various types of events that could signal different kinds of violence, such as explosions, gunshots and car crashes. In line with this, we can train different detectors to find more specific kinds of violence. Instead of fusing several general features that try to encapsulate the whole concept of violence, we can break down into smaller ones and find features for each one, combining them later for a more robust system.

This is not without its challenges, as data for specific events is scarce. Much of the labeled data available for study treats violence as a general concept, and it is highly subjective, since one scene

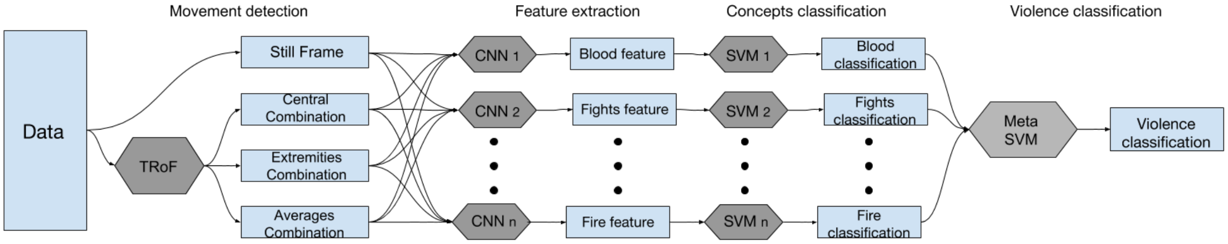


Figure 1: Overview of the pipeline explored. Still frames are extracted directly from the data. After the movement detection by TRoF, the three types of combined images are formed. All four inputs are then feeded to independent CNNs to learn features relative to each concept. These features are used as inputs to SVM classifiers for each concept. At last, the independent classification of concepts are fused in a meta-SVM to classify violence in general.

can be characterized as violent for one person but not so for another. By using specific violence concepts as a starting point, we can find a better definition of violence itself, recognizing common characteristics between them. Grasping the nuances of the concept of violence and understanding its definitions is the first step towards a more robust representation of this problem.

Analyzing the data for the MediaEval benchmark [54], we used seven of the defined concepts of violence: fights, gunshots, explosions, presence of blood, fire, firearms, and cold arms. They were chosen mainly for being the ones that are more represented in the dataset.

We then trained individual CNNs for each concept in order to learn their specific features. We used a variation of the LeNet architecture proposed by Perez et al. [44] that maps the last layer into two filter outcomes (violence vs. non-violence) instead of the original 1000. Also, the network is finetuned for each concept. The features were extracted from the last fully-connected (FC) layer of the network, which is the last layer before classification. The dimensionality of this output is 1024-d.

With these features, distinct SVM classifiers were trained for each concept of violence. Here is important to note that a single sample would have seven different features, one for each concept, and each classifier would have classified these distinct features of the same sample.

Combining the now seven specific classifications for each sample into a single array, we train a meta-classifier for the broader concept of violence. Figure 1 shows a pipeline of the process.

Each of these specific concepts has its own characteristics, thus have to be independently learned by the neural network. For example, the presence of firearms, cold arms and blood can be detected by still images and confidently classified. Firearms have a well-defined shape and color range, while blood can be confidently characterized by its color and texture.

Each of these specific concepts has its own characteristics and have to be independently learned. For example, the presence of firearms, cold arms and blood can be detected by still images and confidently classified. However, fights and explosions are time-based events and can benefit from a network that uses this information to extract features related to these concepts. Gunshots, on the other hand, can be identified by audio and a network that is capable of analyzing this signal from the video can give us a better set of features for the classification. So for each concept, we train a specialized detector and send its features to independent classifiers and only then we make the fusion to classify the scene as violent or not.

5.1.2 Relative Violence

One of the challenges that the subjectivity of violence rises is that the same scene can be classified as violent by one person and not violent by another. This ambiguity is lessened if violence is treated as a relative attribute [42]. Instead of asking if a scene is violent, we can show a pair of scenes and ask which one is more violent. Kovashka et al. [26] showed that humans agree more when they make relative statements vs absolute statements.

With this in mind, we can label our data with violence as a relative attribute and use statistical models such as the Bradley-Terry [3] to estimate and assign real-number scores for each scene of the perceived violence. Now we have a global ranking of perceived violence and can use it to define better thresholds of what is considered as a violent scene.

This method can also be used in conjunction with violence concepts to help determine which ones are perceived as more violent and thus more important for the final classification. For example, if scenes ranked higher in violence contain fights more frequently than fire, we can adjust the weight of these concepts in a later classification fusion. It is also a step in making the definition of violence more robust to study.

Even then, finding specific events may not be enough, as they might not always characterize violence by themselves, for example, a building collapsing can be an act of violence or a planned demolition. The context in which the event is a part of is important in determining its nature.

5.2 Incorporating Temporal Information

For the second front, we will take advantage of the temporal information that exists in videos. For violent scenes, the context in which they occur is also important and this temporal information is crucial to distinguish it from non-violence.

5.2.1 Temporal Features

We will explore different motion descriptors and other forms to represent time. Most of the recent approaches use a combination of features, being them audio ones such as MFCC [9, 23, 30, 60]), or visual features associated with time to describe motion, like STIP [9], Dense Trajectories [9, 30, 41, 60], MoSIFT [6], and Temporal Robust Features a.k.a. TRoF [37]). The most prominent features though, are those extracted by a CNN [9, 23, 30, 35, 41, 51, 60]. The ability to process huge amounts of data and automatically find a set of features that can be used to classify this information is a great advantage, and even though early results are far from ideal, the improvement and potential is significant, specially when temporal information is fed to the neural network.

Our first approach to this front is using TRoF [37] - Temporal Robust Features - which is capable of identifying a sequence of frames that represent movement in a scene. Then we combine specific frames from this sequence to try to capture the event as a whole.

In order to do that, we first run the TRoF detector through each movie of our database and select the most relevant sequences of frames detected. These sequences are comprised of a center frame and a diameter, that represents the amount of frames that encapsulate a specific kind of movement.

Since we are trying to represent fights, explosions, fire and gunshots, we are searching mainly for sudden movements, being it short like gunshots, an explosion or a punch, or a long sequence of shooting or a fight. We can expect the sequences of frames detected by TRoF that actually represent a violent concept to have a wider range of movement throughout their individual frames, compared to a sequence of frames that does not represent any physical violence, such as walking or talking.

With this in mind, we assembled three types of images to use as inputs to the neural network, along with a single, still frame extracted directly from the movies: One to capture the climax of the movement; one to capture the difference between the start and end of the movement; and one to capture the flow of movement throughout the sequence.

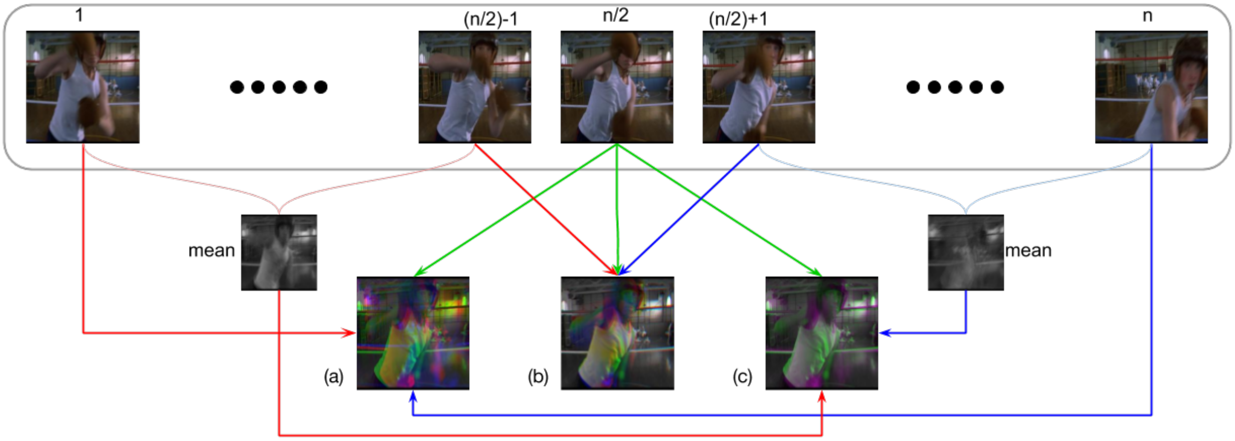


Figure 2: Overview of how the combinations were made from a single sequence of frames detected by TRoF. (a) The first and last frames of the sequence are combined with the center one to form the Extremities combination. (b) The 3 central frames are put together as each of the color channels of the Central Combination. (c) To form the Average Combination, the center frame is joined with the average of the first half frames in the red channel, and the average of the last half of frames in the blue channel.

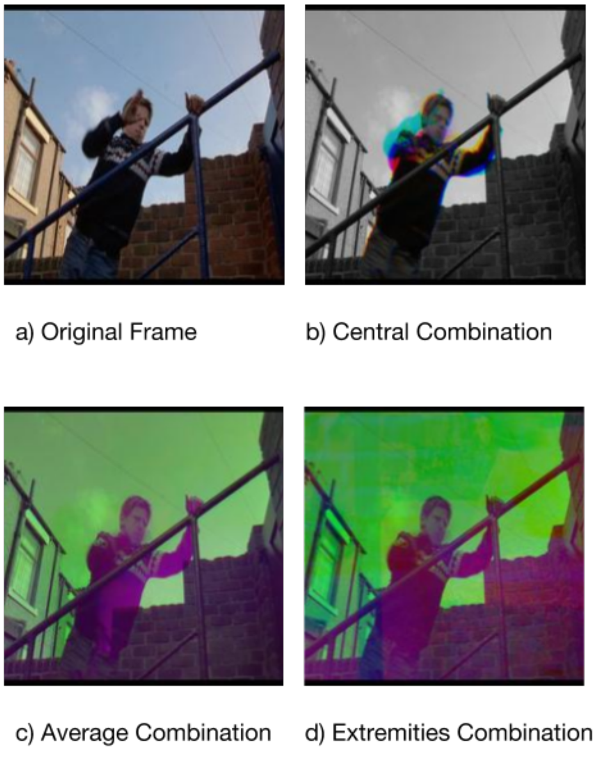
- **Central Combination**: From the center of the sequence detected by TRoF, we combined it with the frame immediately before and immediately after, assigning them to the color channels green, red and blue, respectively and obtaining a single image representing this movement. This type of combination focuses on the climax of the movement, using only the 3 central frames of the sequence.

- **Extremities Combination**: For each sequence of frames detected by TRoF, we combined its center with its start and end frames in the color channels green, red and blue, respectively, to represent the whole movement detected in a single image. This type of combination captures the changes between the start of the movement and its end, using the central frame as a bridge.

- **Averages Combination**: For each sequence of frames detected by TRoF, we combined its center with the average of the frames from the start up to the frame immediately before the center and with the average of the frames from the frame immediately after the center up to the end. Again, assigning each resulting image to the green, red an blue color channels respectively. This type of combination was made to capture the flow of movement throughout the whole sequence, using all frames of the sequence to represent how the movement occurred.

Figure 2 show a scheme of a sequence detected by TRoF represented as an interval of frames and how they were selected to compose each of the combined images. Furthermore, Figure 3 and Figure 4 show, respectively, examples of a non-violent scene and a violent scene from the same movie, including the center frame of the sequence of movement detected by TRoF and the different types of combined images that were used as inputs for the network.

It is important to note that for the TRoF detector, a single center frame can spawn many different sequences, since one frame can contain various kinds of movement, and each one can last for a different amount of frames. Therefore, in our Extremities and Averages Combinations, more than one diameter can be associated with the same center frame.



a) Original Frame

b) Central Combination

c) Average Combination

d) Extremities Combination

Figure 3: Example of a non-violent scene, and the four types of inputs for the neural network, including the Original Frame and the three different types of combinations.

Another approach that is being considered to incorporate motion information to the neural

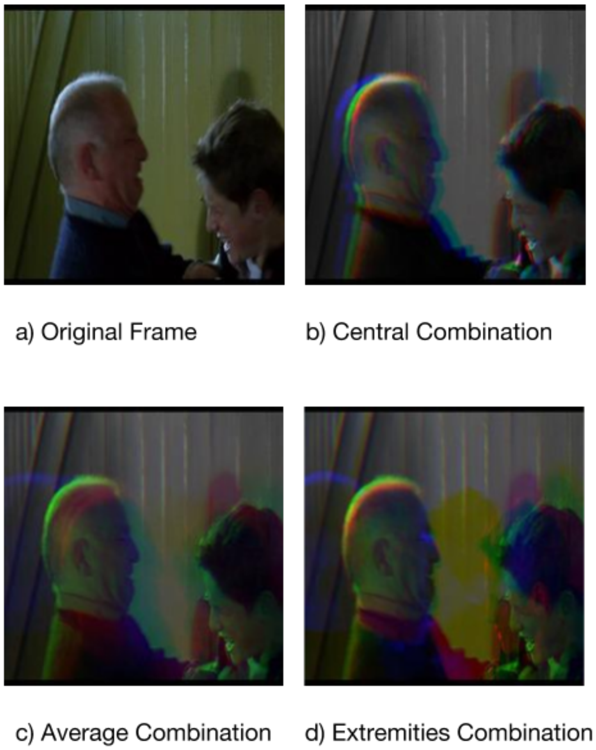


Figure 4: Example of a violent scene, and the four types of inputs for the neural network, including the Original Frame and the three different types of combinations.

network is the one used by Perez et al. [44], that developed a technique that uses the LeNet architecture [55] to detect pornographic scenes in videos. While it uses static information, feeding the raw frames of video to the network, it also calculates the optical flow, commonly used to represent motion in a sequence of frames, and MPEG motion vectors, which are encoded within the MPEG codec of the video itself. To use this information, though, a pre-processing step is performed to generate image representations that are then used as the input the CNN expects. Both the Optical Flow and the MPEG motion vectors are represented as two motion maps, one for the horizontal component and one for the vertical component, containing in each position a measure of motion in the respective direction. These measurements are then rescaled linearly and stored as gray-scale images, one for each component of the motion. The gray-scale maps are used as input to the CNN and a fine-tune step of the parameters is performed through backpropagation, with the weights learned from ImageNet as a starting point. Each of these motion representations are processed separately, generating different CNN models and classifications for each kind of motion information. This technique had very good results and since it worked for pornography, it is worth investigating whether it is effective with other types of sensitive media, such as violence.

5.2.2 Neural Networks and Finite State Machines

Convolutional neural networks process information as distinct data points, but other types of networks are also promising, such as Recurrent Neural Networks (RNN) or Long Short Term Memory (LSTM) Networks, both storing information from previous frames and using it to compute features for the next frames, incorporating some temporal information within its own architecture.

The main appeal of RNNs is that they take as input not just the current input they see, but also what they perceived one step back in time and combine them to determine how they respond to new data. This means that the network can take advantage of the sequential information, and this has a clear purpose of using the information of the sequence itself to perform tasks that feedforward networks cannot. The sequential information is preserved in the recurrent network's hidden state, that span many steps as it cascades forward, affecting the processing of each new input. This architecture can be very useful for detecting violence, since it is a sequence of events that happens over time.

An LSTM network functions similarly, but instead of only using the information passed along by one step in the past, they learn over many time steps, allowing it to link causes and effects that happen over an extended period of time. They contain information outside the normal flow of the recurrent network in a gated cell. This cell makes decisions about what to store, and when to allow reads, writes and erasures, via gates that open and close. These types of networks open up the possibilities for violence detection, since we can effectively use the temporal information of the video as input.

Another type of network, 3D convolutional neural network [20], makes convolutional operations using multiple frames, capturing both temporal and spatial information from video streams. Although more complex, this kind of network can provide valuable temporal information without the need to pre-processing features and use them as inputs.

All of these other types of neural networks point to a hidden transitional state between a non-violent and a violent state, and identifying this hidden state can be a valuable information, not

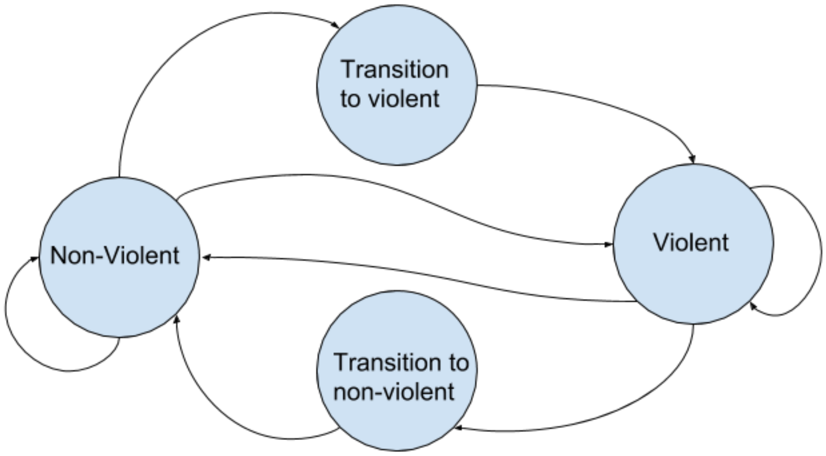


Figure 5: Overview of a Markov decision process for violence and its transitional states

only to help identify when the violent scene start and end, but to be able to predict if a violent scene will happen.

Design a Markovian decision process from a neural network was used for optimization problems [25,49], but can be adapted to identify the boundaries of a violent scene and carry context information to the classification process.

A violent scene is full of subjective contexts, and depending on the nature of the video, a fight scene, for example, can be interleaved with close-ups of spectators, which by themselves do not indicate violence, but are inserted in a violent context. Identifying these transitional states can help determine if a non-violent scene belongs in the context of a violent one or not, illustrated by figure 5.

5.3 Localizing Violence

When detecting violence in a well-defined unit of video, being it a shot, a scene or an entire file, it is possible to label, extract features and classify it as a whole. But for the third research front we aim at in this project, localizing violence excerpts in a video stream, various different challenges arise. For instance, how to segment the video in order to label it, aggregate its features, or even classify it. We could segment the video in frames, shots, scenes, intervals of fixed length, and all of these will potentially generate different results. Since the solutions for this problem are supposed to find when the violent content starts and ends, discovering the best way to segment the video becomes part of the problem.

One option, as used by Moreira [36], is to divide the video into *snippets*, that are fixed-length segments in such a fashion that the overlap between them help pinpoint the time-frames that the content we are interested in occurs, in our case, violence. Each snippet can be individually classified, using the methods we find more efficient for detecting violence. After this, late-fusion techniques are used to combine the results of the snippets classifiers and find out which segments of videos are relevant.

This approach is flexible and lets us take advantage of the various types of features we can extract, being it from static frames, the audio signal or time-based. Its success also makes it a good starting point for tackling this problem. It was second place in the 2014 MediaEval Violent Scenes Detection Task [2]. Further exploration is needed though, since the dataset used was mainly composed of Hollywood movies [54], and even though the test set featured 86 web videos, their total time of 157 minutes is similar to each of the 7 movies reserved for testing. The problem of this dataset is the implicit bias of movie language, where violent scenes are often accompanied by distinct soundtrack and exaggerated sound and visual effects. The same task in the 2015 benchmark featured a more varied dataset, being composed of both professional and amateur videos beng pulled from the web or screened in film festivals, of various genres. From the results it is clear this is a harder dataset to explore.

With deep-learning techniques, this snippet classification approach can be improved upon, for instance, by expanding the fixed-length segments to a variable length one, since this can greatly increase the amount of data we can work with, and neural networks thrive in this scenario. We can also explore the idea of semantic segmentation similar to Günsel et al. [17], where they use video-editing cues to find cuts and scene changes and use these as snippets. Another option is to include key frames detection using the motion information [59] and train a specialized classifier for these frames. All of these can be combined with different fusion strategies, incorporating different meta-learning techniques for classification of the fused vectors.

5.4 Datasets and Validation Metrics

Due to its novelty and subjectivity, the field of violence detection does not have a single main dataset available for study. The first database this project explores is the modified LIRIS-ACCEDE for the 2015 MediaEval competition, comprising 10900 short video clips extracted from 199 movies of various genres [31]. Until 2014, though, the competition used another dataset, containing 2 hollywood movies plus 86 short web videos [13].

There are other datasets available that we intend to study, and a brief overview and description of them is provided by Table 2.

Most of these datasets suggest the use of Accuracy (ACC) for evaluation, which is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Some of the works also calculate the Area Under Curve (AUC), that can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example (assuming positive examples ranks higher than negative) than to a randomly chosen negative example. Precision, Recall and F1 measures usually are calculated together. Precision is the proportion of true positives among all the cases classsified as positives; Recall is the proporton of true positives among all the positive cases examined; and F1 is the harmonic mean between precision and recall.

The MediaEval Violent Scenes Detection (VSD) task motivation was the development of systems that could help users choose suitable titles for their children, by previewing parts of the movie that include the most violent moments. This means that the best performing systems are the ones that return the largest number of violent shots at the first positions of the top-k retrieved shots. To achieve that, the competition suggests using the Mean Average Precision (MAP) at the 100 top ranked violent shots (MAP@100), as the official evaluation metric. Therefore, a solution is considered better than another if it presents a higher MAP@100 value, since it indicates that such solution returns less false positive shots in the first positions of a 100-violent-shot ranked answer.

To be comparable to the MediaEval datasets, which are the ones that have more works associated with, our results must be reported with the Mean Average Precision. The other validation metrics used across the various datasets will be properly calculated and analyzed to give our work a broader range of comparisons.

Table 2: Violent Scenes Datasets available with a brief description of the data each one contains and its respective suggested evaluation metrics.

Name	Year	Evaluation Metrics	Data Description
RE-DID (Real Life Events - Dyadic Interactions Dataset) [47]	2015	ACC	Real Life Scenarios - urban fights 30 videos Length: 0:20 to 4:20
LIRIS-ACCEDE (MediaEval 2015) [31]	2015	MAP@100 AUC	10900 short video clips (8-12 seconds) -Train:100 movies, 6144 shots - Test:99 movies, 4756 shots
Violent Scenes Dataset (VSD) [13]	2014	MAP@100 AUC	-Train: 25 Hollywood movies, 32678 shots -Test: 7 Hollywood movies, 11245 shots 86 web videos
Violent Flows (Crowd Violence) [18]	2012	ACC + AUC	246 videos from YouTube (1-7 seconds)
Hockey Fight [40]	2011	ACC + AUC	1000 clips - 50 frames per clip
Movies [40]	2011	ACC + AUC	200 video clips from action movies
Perperis [45]	2011	ACC Recall Precision F1	25 movie segments (aprox. 1 min each)
Giannakopoulos [16]	2010	Recall Precision F1	50 clips from 10 movies (2,5 hours total)
Souza [11]	2010	ACC	400 videos (50% fight scenes) Easy: STIP + BoVW 99% ACC
BEHAVE [1]	2006	ACC + AUC	Surveillance cameras Only clip 1 annotated (52:11 minutes) Fixed point of view Simulated actions

6 Cronogram

This project has the following programmed activities:

1. Attainment of discipline credits;
2. Literature review;
3. Exploration of available datasets and deep-learning methods for feature extraction;
4. Exploration of violence concepts and proposal of new features for their representation;
5. Teaching internship program;
6. Qualification exam;
7. Exploration of temporal information and proposal for its incorporation into the violence concepts representations;
8. Exploration and proposal of new methods for localizing violence concepts in video streams;
9. Publication of achieved results;
10. Writing of the thesis;
11. Defense of PhD thesis.

Table 3 presents the steps of the project in bimonthly durations.

Year		2016						2017						2018						2019					
Bim.		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Steps	1	●	●	●	●	●	●																		
	2	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●						
	3				●	●	●	●	●	●	●														
	4						●	●	●	●	●	●	●	●	●	●									
	5																●	●	●						
	6																●								
	7											●	●	●	●	●	●	●	●	●	●				
	8													●	●	●	●	●	●	●	●	●	●	●	●
	9															●				●					●
	10																			●	●	●	●	●	●
	11																								●

Table 3: Cronogram of the project.

7 Current Status

The results presented in this section are a summary of the article “Breaking down violence: A deep-learning strategy to model and classify violence in videos” (Bruno Peixoto, Sandra Avila, Zanoni Dias, Anderson Rocha) that will be presented in in the ARES/WSDF conference between the day 27 and 30 of August 2018 in Hamburg, Germany.

Our first experiments were from breaking the concept of violence in seven sub-concepts: Blood, Cold Arms, Explosions, Fights, Fire, Firearms and Gunshots.

Thus, the first major step of our research was the definition of the pipeline used for detecting these concepts and then join them together to classify violence.

We conducted two kinds of experiments, the first is a direct comparison of how our method classifies violence in relation to the literature. The second is an in-depth analysis of each concept and how it succeeds in classifying violence individually. The next section details our results for each one.

7.1 Classifying Violence

To validate and compare our results with existing methods in literature, we used the MediaEval 2013 VSD dataset [12]. It is comprised of 25 Hollywood movies of diverse genres. The dataset provides shot segmentation for all movies, and the resulting segments are individually annotated as containing or lacking physical violent scenes, which “one would not let an eight-year old child see” [12]. This definition is referred to as “Subjective violence”. The segments are also annotated with the definition used in prior years, being that a scene is violent if it contains “physical violence or accident resulting in human injury or pain” [12], which in turn is referred to as “Objective violence”. The annotations process was carried out by seven human assessors, with varied ages and cultural backgrounds, and the shot segmentations were obtained through a proprietary software.

The dataset comes separated into a training set, with 18 movies distributed among 32,678 shots, and a test set comprising 7 movies divided into 11,245 shots. Approximately, 20% of all shots are violent.

The dataset also provides annotations for individual concepts, such as presence of blood, fights, explosions, etc. However, these annotations are only available for the training set of the competition.

As the experimental setup, we extracted all frames from the shots and employed four different types of inputs for separate neural networks: (i) the Original Frame, and the three combined images from the TRoF output discussed in section 5.2.1, (ii) Central Combination, (iii) Extremities Combination, and (iv) Average Combination. These neural networks provided features that were then used to train SVM classifiers for each individual violence concept. Finally, the results were aggregated in a meta SVM that classifies the shots as violent or not.

TRoF details: We used the recommended empirical settings [37], which split the video stream and compute the integral video at every 250 frames. From the output, one blob of each center was used to generate the Central Combination, while the other combinations used all blobs detected.

CNN details: We used the LeNet [55] CNN architecture, implemente finetuning it to each individual concept of violence for each type of input used. Each finetuning ran for 200 epochs and the best one was selected in order to extract the features in the last pooling layer of the network.

SVM details: For the training, we employed three different kernels: (i) linear, used with the liblinear tool [**?**], (ii) χ^2, using power-mean SVM (PmSVM) [**?**] (iii) rbf, applying a grid search to find the best C parameter with $C \in \{2^c : c \in [-8, -6, ..., 8]\}$.

We trained three SVM kernels: Linear, RBF and chi-squared, the latter used by PmSVM [**?**]. The results of the AUC of each SVM for each violence concept is in Tables 4–6.

Linear	**Fire**	**Blood**	**Explosions**	**Gunshots**	**Firearms**	**Coldarms**	**Fights**
Original	**0.739**	**0.590**	0.654	0.566	**0.601**	**0.725**	0.544
Central	0.627	0.576	0.721	0.524	0.573	0.619	0.694
Extremities	0.703	0.585	**0.732**	**0.619**	0.566	0.650	**0.709**
Averages	0.688	0.547	0.715	0.532	0.593	0.670	0.678

Table 4: Area Under Curve of the SVM classifier with the linear kernel of each violence concept.

RBF	**Fire**	**Blood**	**Explosions**	**Gunshots**	**Firearms**	**Coldarms**	**Fights**
Original	**0.767**	0.543	0.731	0.611	**0.594**	**0.711**	**0.742**
Central	0.724	0.587	0.695	0.639	0.587	0.637	0.707
Extremities	0.724	**0.611**	**0.743**	**0.654**	0.586	0.658	0.723
Averages	0.737	0.600	0.731	0.570	0.588	0.661	0.736

Table 5: Area Under Curve of the SVM classifier with the RBF kernel of each violence concept.

From these results we can see that the Original Frames tend to be better classified in the concepts that do not depend so heavily on movement to be identified, such as fire, blood, firearms

Chi-square	**Fire**	**Blood**	**Explosions**	**Gunshots**	**Firearms**	**Coldarms**	**Fights**
Original	0.577	0.618	**0.710**	0.739	**0.698**	0.687	0.723
Central	0.552	0.635	0.687	**0.784**	0.644	0.605	0.721
Extremities	**0.617**	**0.696**	0.679	0.726	0.659	**0.689**	0.735
Averages	0.559	0.628	0.660	0.719	0.638	0.635	**0.743**

Table 6: Area Under Curve of the SVM classifier with the chi-square kernel of each violence concept.

and cold arms. Our combined images using the information from the movement blobs are better classified in the explosions, gunshots and fights concepts.

Also, the overall performance of the chi-square kernel from PmSVM rendered better results, with every concept achieving an AUC of at least 0.617.

From the classification of each concept, we used a meta-SVM classifier to aggregate these into what represents violence in the terms of the competition to better compare our results. Table 7 show our meta-clasification AUC for each type of image input and SVM kernel as well as the MAP@100 measure, used by the competition on the VSD task.

	Linear		**RBF**		**Chi-square**	
	MAP@100	**AUC**	**MAP@100**	**AUC**	**MAP@100**	**AUC**
Original Frames	0.612	0.744	**0.677**	**0.764**	0.669	0.758
Central Combination	0.577	0.730	0.678	0.768	**0.682**	**0.772**
Extremities Combination	0.618	0.762	0.653	0.759	**0.701**	**0.783**
Averages Combination	0.623	0.742	0.688	0.761	**0.696**	**0.779**

Table 7: Results for the classification of violence for each type of input. MAP@100 is reported to provide comparisons with the literature and AUC is reported to compare with the results of each individual concept

.

Table 8 shows our best results on the MediaEval 2013 test dataset. In the top of the table we report the MAP@100 and AUC of our four different types of inputs for the CNN: the three types of combinations and the Original Frames extracted directly from the movies. The bottom of the table shows the MAP@100 of the top three competitors of the VSD task. Since this was the only official metric of the competition, they did not report their AUC. In the middle is the result of Moreira et al. [37] that uses TRoF as a direct feature to classify violence in the same dataset.

The most successful official competitors used a combination of various auditory and visual classical descriptors, including still-image approaches such as HOG, spatio-temporal descriptors such as STIP and auditory descriptors such as MFCC. Even so, our still-image approach reach a competitive MAP@100 (0.677 compared with the 0.690 of the best competitor). The only difference was the features used in the classification that were extracted from the neural network.

The combined inputs based on temporal features got better MAP@100 results, with our Extremities (0.701) and Averages (0.696) types of combinations surpassing the best competitor. Not only this metric, but the AUC of all our methods indicate a better result than the one reported by only using TRoF as a descriptor, with the Extremities Combination reaching an AUC of 0.783 compared to the 0.722 reported by TRoF.

Even though our results were better, it was still close to the best competitor. It is worth noting, however, that our approach relies solely on features extracted from a CNN, whereas the participants in the competition used various types of features combined to reach their results. This shows the power and promise of this deep learning solution, that can still be expanded upon.

Solution	**MAP@100**	**AUC**
Original Frames	0.677	0.764
Central Combination	0.682	0.772
Extremities Combination	0.701	0.783
Averages Combination	0.696	0.779
TRoF [37]	0.508	0.722
LIG-Multimodal [14]	0.690	-
Fudan-Multimodal [8]	0.682	-
NII-UIT-Multimodal [29]	0.596	-

Table 8: Results on the MediaEval 2013 dataset. MAP@100 values were obtained with the same evaluation tool of the competition. All multimodal competitors' solutions employed five or more description modalities. Competitors did not report AUC.

7.2 Relevance of individual concepts

Our second experiment was conducted in order to further investigate the breaking down of violence into concepts. Our aim was to classify violence using only one of the seven concepts we previously used in conjunction, to invstigate the importance of each concept apart from each other. Unfortunetely, the VSD dataset does not contain annotations for each individual concept of violence in the test set, only on the training set.

Thus, for this experiment, we split the VSD training set into our own training and test set, using two movies as the test set: *Fight Club* and *Leon*. For each violence concept, Table 9 shows the area under the receiver operating characteristic curve (AUC) in two distinct situations:

- **Concept x Concept:** Using the violence concept to classify itself, for example, using our blood classifier to classify blood in our test set.

- **Concept x Violence:** Using the violence concept to classify violence in general, for example, using our blood classifier to classify violence in our test set.

The chosen classifier for each violence concept for this experiment was the best one among the four different types of inputs, that include the Original Frame, and the three types of combinations of frames.

Our results show that whilst the aggregation of concepts achieved an AUC of 0.764 in the worst case, and 0.783 in the best as shown in Table 8, if we use only one concept to classify violence, the AUC drops considerably, with the biggest difference being with the firearms concept, to 0.501. This is despite the classifier reporting a better AUC when classifying its own concept, that is, the firearms concept had an AUC of 0.736 in classifying firearms.

	concept x concept	concept x Violence
Blood	0.724	0.513
Cold Arms	0.740	0.504
Explosions	0.748	0.634
Fights	0.778	0.686
Fire	0.631	0.522
Firearms	0.736	0.501
Gunshots	0.809	0.617

Table 9: Results for the AUC of each concept when classifying shots by its presence and classifying violence by itself.

It is noteworthy that even though the classifier for individual violence concepts report poor AUC results when classifying violence, when aggregated we achieve a better result than any one individually. The closest result was with the fights concept, that had an AUC of 0.686 when classifying violence. This can be explained by the correlation that exists with the ground-truth annotation of the concept of fights and the one of violence in general.

In order to have a better understanding of how the definition of violence affects the annotation and the correlations between each individual concept and this definition, Table 10 shows an analysis of the ground-truth annotation of the VSD training set. It shows the percentages of frames that contain each individual concept that are actually annotated as violent.

The dataset provides annotations for two distinct definitions of violence, stated by the competition [12] as:

- **Objective Violence:** physical violence or accident resulting in human injury or pain.
- **Subjective Violence:** events which one would not let an 8 years old child see, because they contain physical violence.

Table 10 then, shows how each concept is viewed by annotators under these two different definitions of violence. Each shot that contains a specific concept can be annotated as having violence by the Objective definition, the Subjective definition, or both. Also, the shot can be classified as non violent by both definitions.

	Non violent	Objective	Subjective	Both
Blood	50.94	1.18	37.83	10.05
Cold Arms	76.06	0.35	15.89	7.70
Explosions	44.48	0.41	28.25	26.86
Fights	16.42	1.45	36.99	45.14
Fire	71.18	1.35	12.93	14.54
Firearms	66.63	0.35	24.51	8.51
Gunshots	44.57	0.19	35.72	19.52

Table 10: Analysis of the annotations of the concepts of violence and how each concept is present in violent scenes, according to the different definitions used by the MediaEval 2013 VSD task.

For example, only 16.42% of the shots that contain fights are not annotated as violent under any of the two definitions. This number is significantly lower than any other concept considered. This indicates that for human annotators, fights are more correlated to violence than any other concept.

It is also shown that 36.99% of the shots that contain fights are annotated as violent when we consider only the Subjective definition of violence provided by the competition. Of all the shots that contain fights, 45.14% are annotated as violent by both definitions. And when considered only the Objective definition, 1.45% of the shots that contain fights are annotated as violent.

These numbers indicate how challenging it is to define violence. Changing the definition from being “physical violence or accident resulting in human injury or pain” to “a scene in which one would not let an 8 years old child see”, made a significant difference. Using the Firearms concept as an example, we can add the percentage of shots annotated as violent by both definitions with the percentage of shots annotated as violent only by the Objective definition to find that 8.51% + 0.35% = 8.86% of the shots that contain firearms are considered violent by the Objective definition, whereas 8.51% + 24.51% = 33.02% of them are considered violent by the Subjective definition.

References

[1] Ernesto L. Andrade, Scott Blunsden, and Robert B. Fisher. Modelling crowd scenes for event detection. In *18th International Conference on Pattern Recognition, 2006. ICPR 2006.*, volume 1, pages 175–178. IEEE.

[2] Sandra Avila, Daniel Moreira, Mauricio Perez, Daniel Moraes, Isabela Cota, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, Anderson Rocha, et al. RECOD at MediaEval 2014: Violent scenes detection task. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, October 16-17*, CEUR Workshop Proceedings. CEUR-WS.org, 2014.

[3] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] Rupayan Chakraborty, Avinash Kumar Maurya, Meghna Pandharipande, Ehtesham Hassan, Hiranmay Ghosh, and Sunil Kumar Kopparapu. TCS-ILAB-MediaEval 2015: Affective Impact of Movies and Violent Scene Detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[5] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference, Nottingham, UK, September 1-5, 2014*, 2014.

[6] Ming-yu Chen and Alexander Hauptmann. MoSIFT: Recognizing human actions in surveillance videos. *In CMU-CS-09-161, Carnegie Mellon University*, 2009.

[7] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu. Semantic Context Detection Based on Hierarchical Audio Models. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 109–115, New York, NY, USA, 2003.

[8] Qi Dai, Jian Tu, Ziqiang Shi, Yu-Gang Jiang, and Xiangyang Xue. Fudan at MediaEval 2013: Violent Scenes Detection Using Motion Features and Part-Level Attributes. In *MediaEval*, 2013.

[9] 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 *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[10] Fillipe Dias Moreira de Souza, Eduardo Valle, Guillermo Cámara Chávez, and Arnaldo de Albuquerque Araújo. Color-Aware Local Spatiotemporal Features for Action Recognition. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 16th Iberoamerican Congress, CIARP 2011, Pucón, Chile, November 15-18, 2011. Proceedings*, pages 248–255, 2011.

[11] Fillipe D.M. De Souza, Guillermo C. Chavez, Eduardo A. do Valle Jr., and Arnaldo de A. Araújo. Violence detection in video using spatio-temporal features. In *23rd SIBGRAPI: Conference on Graphics, Patterns and Images, 2010*, pages 224–230. IEEE.

[12] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Bogdan Ionescu, Vu Lam Quang, and Yu-Gang Jiang. Benchmarking Violent Scenes Detection in movies. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June.

[13] Claire-Hélène Demarty, Cédric Penet, Mohammad Soleymani, and Guillaume Gravier. VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation. *Multimedia Tools and Applications*, 74(17):7379–7404, 2015.

[14] Nadia Derbas, Bahjat Safadi, and Georges Quénot. LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words. In *MediaEval*, 2013.

[15] Nadia Derbas, Bahjat Safadi, Georges Quénot, et al. LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words. In *MediaEval*, 2013.

[16] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In *Hellenic Conference on Artificial Intelligence*, pages 91–100. Springer, 2010.

[17] Bilge Günsel, A. Müfit Ferman, and A. Murat Tekalp. Temporal video segmentation using unsupervised clustering and semantic object tracking. volume 7, pages 592–604, 1998.

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

[19] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *Computing Research Repository (CoRR)*, abs/1207.0580, 2012.

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 35, pages 221–231, 2013.

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *ACM Multimedia*, pages 675–678. ACM, 2014.

[23] 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 *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[24] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[25] M. Kovačič. Markovian neural networks. *Biological Cybernetics*, 64(4):337–342, Feb 1991.

[26] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image Search with Relative Attribute Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] Vu Lam, Duy-Dinh Le, Sang Phan, Shin'ichi Satoh, and Duc Anh Duong. NII-UIT at MediaEval 2014 Violent Scenes Detection Affect Task. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, October 16-17*, CEUR Workshop Proceedings. CEUR-WS.org, 2014.

[29] Vu Lam, Duy-Dinh Le, Sang Phan, Shinichi Satoh, and Duc Anh Duong. NII-UIT at MediaEval 2013 Violent Scenes Detection Affect Task. In *MediaEval*, 2013.

[30] 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 *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[31] Martha A. Larson, Bogdan Ionescu, Mats Sjöberg, Xavier Anguera, Johann Poignant, Michael Riegler, Maria Eskevich, Claudia Hauff, Richard F. E. Sutcliffe, Gareth J. F. Jones, Yi-Hsuan Yang, Mohammad Soleymani, and Symeon Papadopoulos, editors. *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[32] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*, pages 3361–3368, 2011.

[33] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[34] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[35] Ionut Mironica, Bogdan Ionescu, Mats Sjöberg, Markus Schedl, and Marcin Skowron. RFA at MediaEval 2015 Affective Impact of Movies Task: A Multimodal Approach. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[36] Daniel Moreira. *Sensitive Video Analysis*. PhD thesis, Institute of Computing, University of Campinas, 2016.

[37] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Pornography classification: The hidden clues in video space–time. *Forensic Science International*, 268:46–61, 2016.

[38] Daniel Moreira, Sandra Eliza Fontes de Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. RECOD at MediaEval 2015: Affective Impact of Movies Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[39] Jeho Nam, Masoud Alghoniemy, and Ahmed H. Tewfik. Audio-Visual Content-based Violent Scene Characterization. In *IEEE International Conference on Image Processing (ICIP)*, pages 353–357, 1998.

[40] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence Detection in Video Using Computer Vision Techniques. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part II*, pages 332–339. Springer-Verlag, 2011.

[41] Marin Vlastelica P., Sergey Hayrapetyan, Makarand Tapaswi, and Rainer Stiefelhagen. KIT at MediaEval 2015 - Evaluating Visual Cues for Affective Impact of Movies Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[42] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 503–510, Washington, DC, USA, 2011. IEEE Computer Society.

[43] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *European Conference on Computer Vision*, pages 581–595. Springer, 2014.

[44] Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230:279–293, 2017.

[45] Thanassis Perperis, Theodoros Giannakopoulos, Alexandros Makris, Dimitrios I. Kosmopoulos, Sofia Tsekeridou, Stavros J. Perantonis, and Sergios Theodoridis. Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. *Expert Systems with Applications*, 38(11):14102–14116, 2011.

[46] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher Kernel for Large-scale Image Classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

[47] Paolo Rota, Nicola Conci, Nicu Sebe, and James M. Rehg. Real-life violent social interaction detection. In *IEEE International Conference on Image Processing (ICIP), 2015*, pages 3456–3460, 2015.

[48] 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)*, 115(3):211–252, 2015.

[49] T Sastri, J English, and M Krishnamurthi. Modeling a markovian decision process by neural network. 17:464468, 11 1989.

[50] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional Sift Descriptor and Its Application to Action Recognition. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 357–360. ACM, 2007.

[51] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, and Thierry Dutoit. UMons at MediaEval 2015 Affective Impact of Movies Task including Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[52] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[53] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computing Research Repository (CoRR)*, abs/1409.1556, 2014.

[54] Mats Sjöberg, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, and Claire-Hélène Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, October 16-17*, CEUR Workshop Proceedings. CEUR-WS.org, 2014.

[55] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[56] Hanli Wang, Yun Yi, and Jun Wu. Human Action Recognition With Trajectory Based Covariance Descriptor In Unconstrained Videos. In *ACM Multimedia*, pages 1175–1178. ACM, 2015.

[57] Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action Recognition by Dense Trajectories. In *CVPR 2011 - IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011. IEEE.

[58] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[59] Wayne Wolf. Key frame selection by motion analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP)*, volume 2, pages 1228–1231. IEEE, 1996.

[60] Yun Yi, Hanli Wang, Bowen Zhang, and Jian Yu. MIC-TJU in MediaEval 2015 Affective Impact of Movies Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[61] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.