Universidade Estadual de Campinas
Instituto de Computação

Bruno Malveira Peixoto

Harnessing high-level concepts, visual, and auditory features for violence detection in videos

Utilizando conceitos de alto-nível, e características visuais e auditivas para detecção de violência em vídeos

CAMPINAS
2021

Bruno Malveira Peixoto

Harnessing high-level concepts, visual, and auditory features for violence detection in videos

Utilizando conceitos de alto-nível, e características visuais e auditivas para detecção de violência em vídeos

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Anderson de Rezende Rocha
Co-supervisor/Coorientador: Prof. Dr. Zanoni Dias

Este exemplar corresponde à versão final da Tese defendida por Bruno Malveira Peixoto e orientada pelo Prof. Dr. Anderson de Rezende Rocha.

CAMPINAS
2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

P359h Peixoto, Bruno Malveira, 1987-

Harnessing high-level concepts, visual, and auditory features for violence detection in videos / Bruno Malveira Peixoto. – Campinas, SP : [s.n.], 2021.

Orientador: Anderson de Rezende Rocha.
Coorientador: Zanoni Dias.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Aprendizado de máquina. 2. Videovigilância. 3. Extração de características (Inteligência artificial). 4. Visão por computador. 5. Processamento de sinal de vídeo. 6. Descrição de eventos (Computação). 7. Análise de imagem. I. Rocha, Anderson de Rezende, 1980-. II. Dias, Zanoni, 1975-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Utilizando conceitos de alto-nível, e características visuais e auditivas para detecção de violência em vídeos
Palavras-chave em inglês:
Machine learning
Video surveillance
Feature extraction (Artificial intelligence)
Computer vision
Video signal processing
Events description (Computer science)
Image analysis
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Anderson de Rezende Rocha [Orientador]
Luca Bondi
Flavio Elias Gomes de Deus
Tiago Fernandes Tavares
Jacques Wainer
Data de defesa: 14-07-2021
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-4994-8973
- Currículo Lattes do autor: http://lattes.cnpq.br/8782940489004393

Universidade Estadual de Campinas
Instituto de Computação

Bruno Malveira Peixoto

Harnessing high-level concepts, visual, and auditory features for violence detection in videos

Utilizando conceitos de alto-nível, e características visuais e auditivas para detecção de violência em vídeos

Banca Examinadora:

- Prof. Dr. Anderson de Rezende Rocha (Supervisor/*Orientador*)
  IC - Universidade de Campinas
- Dr. Luca Bondi
  Bosch
- Prof. Dr. Flavio Elias Gomes de Deus
  Universidade de Brasília
- Prof. Dr. Tiago Fernandes Tavares
  FEEC - Universidade de Campinas
- Prof. Dr. Jacques Wainer
  IC - Universidade de Campinas

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 14 de julho de 2021

No escuro, o muro se arma
Quanto mais alto,
Mais silêncio,
Mais nos afasta

(Alexandre Kumpinski)

Acknowledgements

For this work to be done, I relied on the help and support of many people. My family, for the unconditional love and support through such long distances. My advisors, Anderson and Zanoni, for the patience and the help throughout this journey. My friend Lavi, who believed so much in me and was a joy to discuss ideas with. My partner Tiago, who dealt with so much emotional blows but kept by my side, lifting me up. My friends from 7P, for being my refuge, my safe place. My friend Karina, for inspiring me to follow a path that is meaningful to me. And so many more people that dealt with my doubts and hardships through this time. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. This study was done with the support from CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

Resumo

Ao detectar mídias sensíveis, violência é uma das mais difíceis de se definir objetivamente, e por isso, um desafio significante quando se trata de detectar automaticamente. Enquanto muitos estudos foram conduzidos para detectar aspectos de violência, poucos tentam solucionar o conceito de forma mais geral. Neste trabalho, é proposto um método que tem por objetivo habilitar uma máquina a entender o conceito de alto-nível de violência. Isso é feito a princípio dividindo o conceito em outros mais simples e objetivos, como lutas, explosões, sangue e tiros para depois combiná-los, levando a um melhor entendimento da cena. Para isso, as características individuais de cada sub-conceito são levadas em consideração para guiar a forma como elas devem ser descritas, usando redes neurais convolucionais específicas para obter tais características. Por exemplo, uma cena de luta deve incorporar características temporais que uma cena com sangue não precisa. Uma cena com explosões ou tiros deve levar características auditivas mais em consideração. Com essa solução multimodal, detectores de características visuais e auditivas são treinados separadamente e depois combinados em uma rede neural de decisão que compõe um detector de violência que considera diferentes aspectos do problema. Essa solução robusta e modular permite que diferentes pessoas e culturas adaptem o detector para suas necessidades específicas. Resultados experimentais obtidos em datasets padrões mostram importantes avanços em relação ao estado da arte.

Abstract

When detecting sensitive media, violence is one of the hardest to define objectively, and thus, a significant challenge to detect automatically. While many studies were conducted in detecting aspects of violence, very few try to approach the general concept. In this work, a method is proposed that aims to enable machines to understand a high-level concept of violence. This is achieved by first breaking it down into smaller, more objective ones, such as fights, explosions, blood, and gunshots, to combine them later, leading to a better understanding of the scene. For this, we leverage characteristics of each individual sub-concept of violence to guide how they should be described, relying upon custom-tailored convolutional neural networks. As an example, a fight scene should incorporate temporal features that a scene with blood does not need to describe. A scene with explosions or gunshots should weigh more on its audio features. With this multimodal approach, we trained visual and auditory feature detectors and later combined them into a decision neural network to give us a violence detector that considers several different aspects of the problem. This robust and modular approach allows different cultures and users to adapt the detector to their specific needs. The obtained results on standard datasets in the literature show important advances over prior art.

List of Figures

2.1 Types of violence according to the WHO 21

2.2 Comparison of scenes that could be classified as violent or not 22

3.1 Pipeline of the multimodal features fusion solution 32

3.2 Representative frames of each violent sub-concept considered 33

3.3 Input and output of the inception v4 architecture 34

3.4 Architecture of the C3D network 34

3.5 Architecture of the CNN-LSTM network 35

3.6 Representation of Optical Flow and Optical Acceleration 36

3.7 Overview of the construction of TRoF combinations 37

3.8 Raw-audio waveforms and their corresponding spectrograms 40

3.9 Fusion neural network structure 42

4.1 Examples of frames classified correctly and not 50

List of Tables

2.1 Summary of Violence Detection datasets 24

2.2 Results of the 2015 MediaEval VSD Task 27

2.3 Recent studies on violence detection 29

3.1 Features extracted from each raw audio clip. 39

4.1 Details of the Mediaeval 2013 VSD task dataset 44

4.2 Accuracy for TRoF Combinations with a linear kernel SVM 47

4.3 Accuracy for TRoF Combinations with a RBF kernel SVM 47

4.4 Accuracy for TRoF Combinations with a Power Mean kernel SVM 47

4.5 Accuracy for TRoF Combinations with a neural network classifier 48

4.6 Accuracy for all types of input manipulation visual dynamic features . . . 49

4.7 Accuracy for all time-based architectures and input transformations 51

4.8 Accuracy for the best C3D and audio features 51

4.9 Accuracy for the best visual and audio features 52

4.10 Mean Average Precision in the Mediaeval VSD task in 2013 and 2015 . . . 53

4.11 Mean Average Precision in the NTU-CCTV dataset 53

4.12 Accuracy and Mean Average Precision with a specialized dataset training . 53

Contents

1 Introduction 13

1.1 Challenges 15

1.2 Contributions 15

1.3 Publications 16

1.4 Thesis Organization 16

2 Related Work 17

2.1 Related Concepts 17

2.1.1 Deep Learning 17

2.1.2 Convolutional Neural Networks 18

2.1.3 Extracting Temporal Features 18

2.1.4 Using Audio Features 19

2.2 Definitions of Violence 19

2.3 Annotation Challenges 22

2.4 Prior studies in violence detection 24

2.4.1 Action Recognition and Early Studies 25

2.4.2 MediaEval Violence Scenes Detection Task 26

2.4.3 Recent Studies 27

3 Breaking Down Violence 31

3.1 Concepts of Violence 31

3.2 Visual Static Features 33

3.3 Visual Dynamic Features 33

3.3.1 Time-based Architecture 34

3.3.2 Input Transformations to Capture Motion Information 35

3.4 Audio Processing 37

3.4.1 Standard Acoustic Features 38

3.4.2 Temporal Statistics Calculation 41

3.4.3 Learning step 41

3.5 Fusion Neural Network (NN) 42

4 Experiments and Results 43

4.1 Dataset and Metrics 43

4.2 Validation and Results 45

4.2.1 Visual and Motion Detection 46

4.2.2 Combined Visual and Audio Features per Concept 50

4.2.3 Specialized Dataset 53

5 Conclusion and Future Work 55

Bibliography 56

Chapter 1

Introduction

Violence detection is a crucial aspect of sensitive media filtering, be it a tool to protect users from undesired media or to detect inappropriate behavior in surveillance systems. Currently, countless hours of video are uploaded every minute through the Internet and different social media platforms. Analyzing such a significant amount of footage is heavily time-consuming. Thus, this process's automation is a desirable application for a large portion of society, from schools that want to avoid violent material being shown to children to law enforcement in forensic examination cases.

The issue of violence detection in video scenes was firstly addressed for the task of action recognition. In that case, before deep-learning-based methods, the Bag-of-Visual-Words (BoVW) approaches [43, 14] were a cornerstone in the area. Some examples of this are the work of Bermejo et al. [43], that used low-level features obtained by an image descriptor such as Space-Time Interest Points (STIP) [31] to predict violence via Support Vector Machines (SVM). Souza et al. [14] investigated local spatial-temporal features for violence classification. Clarin et al. [9] addressed the local interest-point approach to detect fights as subjective violence. Even a novel descriptor was proposed by Hassner et al. [22] in 2012 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 better results) for violence detection. Most of the first deep-learning solutions come from the "MediaEval Affect Task" competition, which aims to identify violence in movies [32]. The works that resulted from this competition [12, 30, 37, 65, 72, 71] were the first that incorporated deep-learning features into their BoVW approaches or relied solely on these new features to classify violence.

Most of the works mentioned above, using or not deep-learning, rely either on a specific concept of violence (i.e., fights) or on a generalized definition of violence that tries to capture all its manifestations.

The success of automatic action and object detection methods relies upon how reliable is the studied concept description. This description, be it by hand-crafted features or data-driven derived characteristics such as those produced by neural networks, is backed up by observable and commonly identifiable features. For example, to describe a car, we can locate the wheels, the doors, glasses, and bodywork. To describe a person walking, we have the physical description of a human (head, torso, arms, legs, etc.) and can use the body's movement (one leg in front of another), to detect a walking pattern.

For violence, this kind of description is a challenge in itself and make up for one of our key questions and contributions of this work:

- How to define violence in terms that a computer can understand, i.e., in terms of a series of more well-defined concepts with specialized and complementary detectors?
- How different kinds of violence interact to enable a system to understand the meaning of violence in general that is scalable and amenable to additional concepts added in the future?
- How to build a system that relies upon visual and sound signals characteristics with different violence sub-concepts to detect violence in the general sense?

In this work, we aim at addressing the definition of violence problem by breaking down the subjective concept of "violence" into more objective concepts:

- **Blood:** The presence of blood, even in a still image, is a strong indicative of violence. In a video, blood is likely a consequence of a violent action.
- **Cold Arms:** Knives, swords, clubs, all of these make up this concept. The presence of these weapons in a scene can be considered non-violent, but if it is wielded by a human it can be an indication of an incoming fight scene.
- **Explosions:** While this concept can not be considered inherently violent, explosions rarely make up a non-violent scene. It can be big or small, with or without fire involved.
- **Fights:** Most of the studies on violent scenes consider fights as the definition of violence. Many datasets focuses only on fight scenes. They can be one-vs-one, one-vs-many, or a crowded scene in a many-vs-many fight.
- **Fire:** This is another concept that is not inherently violent, but can be a strong indication of a violent scene that is happening. It can appear in conjunction with explosions, fights or gunshots.
- **Firearms:** The presence of firearms can be considered violent or not, depending on the subjective definition of the viewer. A strong indication, though, is a scene where a human is holding a weapon. Usually gunshots are also involved.
- **Gunshots:** Even if the shot does not hit another person in-scene, a gunshot can be considered a clear indication of violence.

Breaking down violence into different subjects is a proxy to achieving more accurate and robust performance [8]. The breakdown allows us to perform a better investigation of different subjects' behavior, as each subject of violence has different characteristics.

To study how these different concepts interact with each other, we consider the concept of *violence* as a single high-level concept to analyze the behavior of different integrating concepts individually. We then combine violence concepts to identify the more general concept of violence and compare different setups' performances.

Lastly, we discuss the particularities of each distinct concept of violence, mainly focused on how some concepts have more pronounced sound signatures, such as explosions and gunshots, and how combining features from both the visual and audio signals contribute to the violence detection system.

This work aims at developing a combined model of visual and audio feature representations. The sub-concepts analyzed convey different information. Blood and fire, for example, can be easily detected with still frames, with very distinct textures and colors. Fights and explosions, on the other hand, convey the idea of movement, so it is important to use this information to better detect the action through time. Gunshots have very characteristic sounds, thus taking this into consideration will help aggregate more information. The idea is to analyze various sub-concepts with different characteristics to detect the more complex (and subjective) concept of violence.

The rest of this work is organized as follows. Chapter 2 reviews the related works on the task of violence detection by pointing out the definitions of the concept of violence and its challenges. Chapter 3 introduces our proposed method, detailing how the visual and audio features are used and combined. Chapter 4 presents our experimental evaluation and a discussion of the results obtained. Finally, a conclusion is drawn in Chapter 5.

1.1 Challenges

The main challenges of this work rely on agreeing on a satisfactory definition of violence, finding a diverse dataset with annotated data, and using a data-driven approach to classify violent scenes. Additional challenges include:

- Designing a method that can describe violence in a way that suits different cultures;
- Finding relevant databases to support the development of a data-driven approach; and
- Combining visual and sound features in a complementary way to extract the best of both worlds.

1.2 Contributions

The main contribution of this work is to provide a robust methodology for violence detection that can be adapted to the culture and subjective definition of the end-user. In order to get to this point, this work contributes with:

- A modularized definition of violence, using more objective concepts as building blocks for a highly subjective one.
- A plug-and-play method of violence detection. With a final fusion network that is independent of the method used to extract features from sub-concepts, allowing for the addition or removal of relevant components.

- A complementary way of combining visual and audio features to have a more complete information extraction that enhance each of its solo counterparts.

1.3 Publications

During this research, the following papers were produced:

- B. Peixoto, S. Avila, Z. Dias, and A. Rocha. 2018. “Breaking down violence: A deep-learning strategy to model and classify violence in videos.” In Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018). Association for Computing Machinery, New York, NY, USA, Article 50, 1–7. [47]

- B. Peixoto, B. Lavi, J. P. Pereira Martin, S. Avila, Z. Dias and A. Rocha, “Toward Subjective Violence Detection in Videos.” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), Brighton, UK, 2019, pp. 8276-8280. [45]

- B. Peixoto, B. Lavi, P. Bestagini, Z. Dias and A. Rocha, “Multimodal Violence Detection in Videos.” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 2020, pp. 2957-2961. [44]

- B. Peixoto, B. Lavi, Z. Dias and A. Rocha, “Harnessing high-level concepts, visual, and auditory features for violence detection in videos.” Journal of Visual Communication and Image Representation (JVCI), 78. 10.1016/j.jvcir.2021.103174 [46]

1.4 Thesis Organization

This thesis is organized as follows:

Chapter 2 introduces some of the concepts used as well as some related works and challenges of the violence detection problem.

Chapter 3 details the methodology used as well as going deeper into the concepts introduced earlier.

Chapter 4 shows the experiments, datasets and metrics used to evaluate the method, as well as a discussion of the results obtained.

Chapter 5 offers a conclusion to this work and possible future work.

Chapter 2

Related Work

This chapter explores some of the related work that helped our research on each of these challenges.

2.1 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 briefly explore some of the tools we use in this project to make this possible and later we discuss them in more detail.

2.1.1 Deep Learning

Deep Learning is a family of methods 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 hand-written 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. [28] winning the 2012 ImageNet classification benchmark with their CNN model achieving an error rate of 16.4% compared to the second-place result of 26.1%.

Zeiler et al. [74] 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 [23], to prevent overfitting.

2.1.2 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 the 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 is 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 back through the network, instead of retraining it from the start with these new adjustments.

One of the most influential deep networks is called AlexNet [28], an 8-layer network with the first five being convolutional layers, some followed by max-pooling layers, and the last three fully connected layers. This network won several image competitions and made deep learning the main method of classifying images.

In 2015, Szegedy et al. [61] 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, but with 12 times fewer parameters than AlexNet [28], making it an efficient architecture for computer vision. This network obtained excellent performance in the 2014 ImageNet classification benchmark [52]. 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%.

2.1.3 Extracting Temporal Features

Convolutional neural networks typically are applied to 2D images. In order to use them to detect movements in videos, a simple approach is to treat video frames as still images and apply CNNs to recognize movements at the frame level. This approach, though, does not consider the motion information encoded in multiple contiguous frames. To incorporate this kind of information in video analysis, some solutions are proposed.

Long Short Term Memory (LSTM) network is a variant of recurrent neural networks, proposed in 1997 and designed to learn information from sequential data. An LSTM unit incorporates a memory unit cell that can choose to store the predecessor state or clear its content. The controller has three gates, called the input gate, output gate, and forget gate. Stacking LSTM units layer after layer will form an LSTM model. In short, this kind of network recursively maps the input representations to the output gates while every memory cell may contain the current content inside, depending on the state of the controller. These memory cells store information over time to use with subsequent inputs and learn long-range features.

3D Convolutions - C3D was proposed in 2010 by Ji et al. [24] as an approach to solve the action recognition problem. Using several sequential frames as input, they applied multiple distinct convolutional operations at the same location on the various frames of the input. They developed a 3D CNN architecture that generates multiple channels of information from adjacent video frames and makes convolutional operations separately in each channel. In short, they used a 3D kernel to convolve a cube formed by stacking contiguous frames together. With this, the temporal information correlated between these frames can be learned by the network.

Optical Flow is another way of representing temporal information from videos. Instead of doing this with the network architecture, optical flow is a descriptor that can be used as input to a regular CNN. This descriptor is the pattern of apparent motion of objects in a sequence of frames. A differential equation is used to calculate the distance and velocity that a pixel moved from one frame to the next. A visual representation of the movement between two frames can be used as an input that has embedded motion information.

Another way of representing temporal information within the input of the network is using **Temporal Robust Features (TRoF)**. These features, proposed by Moreira et al. [38], aims to analyze a sequence of frames and locate the center of a movement cube, that represents the spatial center of a moving object in the central frame that this movement occurs. These features can then be used as an efficient starting point to the action detection solution, by reducing the scope of the problem.

2.1.4 Using Audio Features

Another important aspect of violence detection in videos is the ability to use sound as a complementary feature. Many audio features can be used to aid this detection, the main one being

Mel-Frequency Cepstral Coefficients (MFCCs), which is a common feature in speech recognition problems. This feature is a representation of the power spectrum of sound based on a linear cosine-transformation on a nonlinear Mel scale frequency.

Inspired by the work of Borreli et al. [4], we can also extract **Temporal statistics** from audio signals, computing simple features such as the average frequency, standard deviation, maximum and minimum in a given time frame and associate them with the corresponding video frames.

This way, we can incorporate more features and reduce the impact of noise and background sounds.

2.2 Definitions of Violence

In order to study violence, our first challenge is to reach a definition of what is violence. Several studies sought to reach a conclusion for this:

In 1978, Giuliano Pontara published an article analyzing different conditions for an action to be considered violent [51]. It is discussed if violence is characterized by illegal action, if it only involves physical injury, if there is unintentional or even consensual violence. This study arrives at a definition that reads: “An action performed by an agent is an act of violence if it causes at least one human being to die, suffer or get physically or psychologically injured against his will while the agent believes that harm has been done.”

While this definition seems to cover many aspects of what can be considered violence, one can argue that the context in which the action takes place is an important factor to account for. In 2007, Morrison and Millwood [40] performed studies with groups of adults and children in order to understand how they define screen violence.

Interesting points brought to attention were that some adults familiar with violence did not consider a fight violent unless some kind of accepted codes of behavior are broken. For example, during a fistfight, one of the involved breaks a glass in the face of the other, or if one pulls a knife out of their pocket.

For the children, a scene with orcs marching from the movie *Lord of the Rings: The Fellowship of the Ring* is considered violent because orcs are like monsters, thus, more violent or "scary".

In 1996, the World Health Organization (WHO) global consultation on violence and health defined violence as "the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either result in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation." [64].

On top of this definition, a typology of violence is also presented and can help understand the context in which violence occurs, even though these classifications are not uniformly accepted. Namely, there are four types of violence: physical, sexual, psychological, and deprivation or neglect. Here we can already sense how broad this concept truly is. All these four kinds of violence have drastically different visual cues, and each one has its consequences that allow for any kind of late identification.

Besides these, there is yet another classification according to the victim-perpetrator relationship: self-directed, community, and interpersonal, with their own sub-categories. Figure 2.1 shows how all these types of violence are organized. Self-directed is violence that the victim inflicts upon themselves. Collective violence refers to violence committed by a large group of individuals, such as terrorist acts, urban riots, and police interventions. Interpersonal is the violence committed between individuals, and has its own sub-categories, whether the aggressor is part of the family or the community.

All these different approaches to even define what violence is makes it a challenging task to automate the identification of such an act. The success of automatic action and object detection methods depends upon how reliable is the studied concept description. This description, be it by hand-crafted features or data-driven derived characteristics such as those produced by neural networks, is backed up by observable and commonly identifiable features. Hence, in the case of detecting a physical object like a car, we can describe concrete, tangible parts of it, such as wheels, doors, metal. To describe a person we also use physical descriptions, even if we are dealing with a walking movement, we can use the way the body moves to detect different patterns.

Violence, on the other hand, is a subjective action with a contested description, meaning that there is no single set of descriptions for a violent act. Since there is no single set of observable characteristics that define violence, we can narrow our research to only tackle physical violence and still have a broad set of actions. For example, to describe a violent fight, we use different visual cues, it could be a fistfight, or a fight with knives, even the presence of guns and shooting, or it could be a fight involving running actions, or

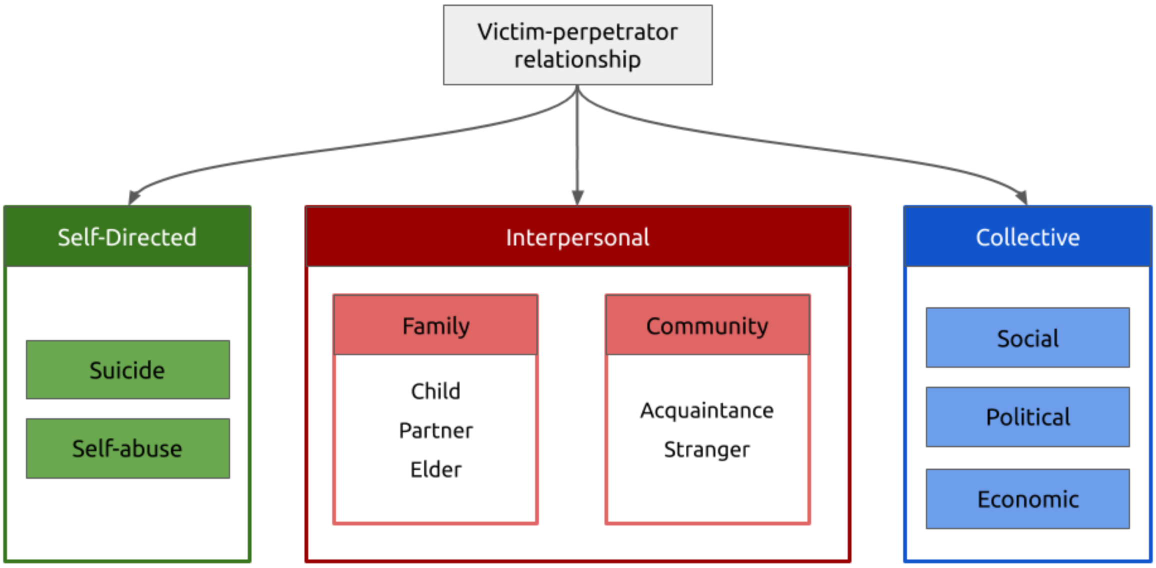

Figure 2.1: Classification of the WHO for the types of violence according to the victim-perpetrator relationship.

more standing still or rolling on the floor fight. Reflection of this difficulty can be found in early works on violence detection, that instead of defining what violence is, focused on different categories or smaller concepts and features that convey violence.

Nam et al. [42] proposed a method that detects flames, blood, the degree of motion in the scene, and characteristic sounds to detect a violent event. Cheng et al. [8] proposed an auditory approach to detecting critical audio events such as gunshots, explosions, engines, and car breakings. Datta et al. [13] proposed a hierarchical approach for detecting violent events involving two people, such as fist fighting, kicking, and hitting each other with objects. Clarin et al. [9] developed Dove, a detection method that uses skin color, blood, and motion activity to classify violence.

One method that seeks to identify violence through motion is proposed by Bermejo et al. [43]. They used a BoVW approach, with low-level features such as STIP and Motion SIFT (MoSIFT) [7], adding a histogram of optical flows representing local motion. These features were then used to establish a bag of words for each video, classified via SVM. However, their work was tested in a dataset containing scenes from hockey games and labeled as either 'fight' or 'non-fight' classes.

These works are examples of the difficulty of defining a single concept for violence and the lack of a unified benchmark for the problem itself. Later in 2013, the Mediaeval initiative proposed a Violence Detection Task (VSD) competition [15] that spawned many works on the area and proposed a widely accepted database and definitions of violence — both of which evolved in the following years of the competition.

The first definition for the MediaEval benchmark is to classify violence as "physical violence or accident resulting in human injury or pain". This is a restrictive and objective definition. While it includes many violent scenarios, it requires human participation and a

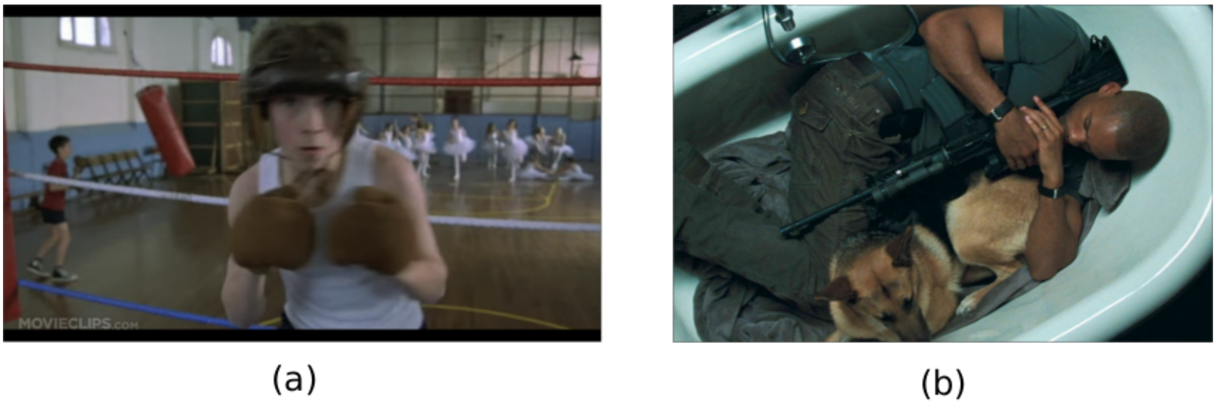

Figure 2.2: Comparison between two scenes that could potentially be classified as violent or not, following the MediaEval VSD subjective definition. (a) From the 'Billy Elliot' movie, a scene of a kid training for a boxing match - labeled as violent. (b) From the 'I Am Legend' movie, a scene of a man sleeping while holding a firearm - labeled as non-violent.

human actually being the victim of violence. For example, a scene where a bomb explodes in a desert or a simple theft would not be included in this definition. Consequently, another more subjective definition included: A scene is violent if "one would not let an eight-year-old child see it, because they contain physical violence". This definition has its problems due to its highly subjective nature, and an example can be seen in Figure 2.2. This definition still comprises a broader range of violent scenes and is the one we adopt in this work.

2.3 Annotation Challenges

The difficulty in defining violence reflects the lack of annotated data samples. Even though we can find several violence-related datasets, they have few samples, or the specific violence they represent varies significantly. This section provides some examples of existing datasets we considered for our work and briefly explains why we deemed those datasets fit or not for our research.

Older datasets consist of detecting various human interactions, not specifically violent, focusing on the human action recognition problem.

The **BEHAVE** dataset [2] comprises clips of groups of people meeting, chasing one another, fighting, following, or just walking together. The **Two-person** dataset, created by Yun et al. [73], also aimed at general human action recognition but explicitly based on two-person interactions as approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. While being widely used, both of these datasets do not contain a significant amount of violent samples.

In contrast, some datasets focused on violent behavior. Nievas et al. [43] introduced the **Hockey Fights** dataset, containing 1000 short clips from the National Hockey League games depicting violence in a dynamic environment while retaining similar non-violent scenes. Although this dataset has its difficulties, especially the motion blur present in most scenes, its nature does not generalize violence, neither the concept of fights, capturing only a specific scenario.

The **Violence in Movies** dataset has also been presented by Nievas et al. [43] and comprises 200 video clips with 100 fight scenes in a more varied scenario. However, the violent scenes have a similar structure and differ significantly from the non-violent ones. This kind of dataset can lead to a biased network that not necessarily detects the violent action but surrounding clues of the scenes specific to this dataset. This behavior is further impacted by the very short length of the dataset, that total 6 minutes of videos.

Another explored aspect of the problem is crowd violence, aimed at detecting violence in scenarios where the specific action is not explicit. We considered two datasets for this kind of problem, but both of them have small sample sizes:

The **Violent Flows** dataset published by Hassner et al. [22] comprises 246 short-video sequences and focuses on the scenario of violent crowds, including football stadiums, bars, and public demonstrations, both indoors and in open areas.

In short, we are faced with a challenge where the available datasets' size tends to be small, reducing the capabilities of data-driven approaches such as deep learning, but not only that.

While these datasets can detect some violent behavior, they do not address the general purpose of detecting violence. We are presented with either *tangential solutions* (i.e., human action recognition that also have some violent actions) or *specific solutions–* such as detecting fights, accidents, or crowd riots.

This is the reason, in this work, why we mainly use the **Violent Scenes Detection (VSD)** dataset, published by Demarty et al. [16], which is the most prominent dataset used for violence detection to date. It was proposed as part of the Mediaeval competition and evolved over the years. Initially, the dataset comprised 25 Hollywood movies of diverse genres, but later on, it also added 86 YouTube clips. This dataset provides fully annotated movies and annotation of more specific violence triggers in each scene, such as fights, explosions, and blood presence. It is important to note that even though this is an important dataset, the annotation of clips as violent or non-violent is still mainly a subjective matter. Our work herein adopts this dataset but always acknowledges the subjective aspect of the final violence classification.

Since the start of this study, other datasets were made available, we will briefly discuss them here and provide a summary table of all the datasets mentioned.

In 2018, Sultani et al. [59] designed a **Crime Anomaly** dataset extracted from CCTV cameras with 1900 videos, totaling 128 hours of video. The training set does not have time-based annotation, though, limiting the violent label to video-level. They proposed a Multi-Instance Learning method where each video is segmented into a fixed number of segments and videos labeled as positive are assumed to have at least one violent segment while negative labeled videos contain no violent segments. They use a ranking loss function to find which segment in a positive video is most likely to contain a violent scene and compare it to the highest-ranked segment of a negative video, aiming to further the distance between them. This dataset, while very large, does not contain audio information, and their annotated labels limit the solution methods.

In 2019, Perez et al. [49] proposed a **CCTV-Fights** dataset containing 1000 videos, in a total of 18 hours, mainly focused on CCTV and mobile videos of fights. This dataset contains 280 CCTV videos, ranging from 5 seconds to 12 minutes in length. The additional 720 videos were uploaded to YouTube from multiple sources, mainly mobile cameras. These videos were annotated at frame level. They proposed three different approaches to tackle this new dataset: i) A two-stream 2D-CNN architecture, a spatial stream, using the RGB data of video frames, and a temporal stream, using a stack of optical flows. ii) a 3D-CNN approach that enables convolution on three dimensions directly from a stack of frames. iii) a local interest points solution based on Temporal Robust Features (TRoF) [38].

Finally, in 2020, Wu et al. [70] proposed yet another dataset, **XD-Violence**, with more than 4700 videos, totaling 217 hours, aiming to design a neural network on large-scale data. This dataset has its audio signals available to allow for a more streamlined multimodal approach and has videos collected from movies and in-the-wild scenarios. This dataset contains frame-level annotation and is currently the largest one we are aware of. Their proposed solution to tackle this dataset involves a multimodal fusion, using both audio and video information and Holistic and Localized Networks, inspired by Graph neural networks (GNNs) such as [77, 75] to exploit relationships between video snippets. Since this dataset was proposed late in our research, we could not do any testing with it.

Dataset	# Videos	Length	Source of scenarios	Audio
Violence in Movies [43]	200	6 min	Movies and sports	No
Hockey Fights [43]	1000	27 min	Ice hockey	No
Violent Flows [22]	246	15 min	Streets, school and sports	Yes*
VSD 2013 [15]	25	48 hours	Movies	Yes
VSD 2015 [16]	111	51 hours	Movies, mobile cameras and sports	Yes
CCTV-Fights [49]	1000	18 hours	CCTV and mobile cameras	Yes*
Crime Anomaly [59]	1900	128 hours	CCTV Camera	No
XD-Violence [70]	4754	217 hours	Movies, sports, mobile cameras and CCTV	Yes

Table 2.1: Summary of Violence Detection datasets. (*means that many videos are silent or only contain background music.)

2.4 Prior studies in violence detection

Early studies on violence detection were largely based on action recognition ones. The idea behind classifying different actions is to find a representation of motions that could uniquely identify them. These early studies were conducted in the limited datasets discussed in the annotations challenges, which increased the difficulty of directly comparing them since there was not a single dataset that was used for most of them.

In 2013, with the MediaEval VSD dataset, we had not only a dataset to compare a large number of studies, but also a more unified definition of the concept of violence. Since 2015, the competitors started using deep neural networks in this competition and many more works in the field also adopted this approach. In the following sections, we explore a little more in-depth some of these studies and the history of violence detection.

2.4.1 Action Recognition and Early Studies

Wang et al. [67] developed an action recognition method using dense trajectories. They introduced a descriptor that computes motion boundaries from the optical flow information to find a trajectory for the motion. This approach has been further improved [48] by correcting camera motion [68] and using Fisher Vector encoding [50], 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. [24] proposed a 3D convolutional neural network for action recognition, stacking multiple contiguous frames of video and using them 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.

Simonyan and Zisserman [56] proposed a convolutional network for action recognition that separates the spatial information from the temporal one and later combines them. Both of them are implemented in distinct networks. The spatial stream uses still video frames and the temporal stream uses optical flow as inputs. While the temporal information still relied on a manipulated input, the extracted features showed that a standard 2D convolutional network could be used for this task.

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. [42], 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 to identify quick movements as well as pixel color thresholds for blood detection.

Cheng et al. [8] 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 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. [43]. They exploited a Bag of Visual Word (BoVW) approach, using low-level features such as Space-Time Interest Points (STIP) and Motion SIFT (MoSIFT) [7], which is an extension of the SIFT [35] image descriptor for video, adding a histogram of optical flows representing local motion. These features were then used to establish a bag of words for each video that was classified via Support Vector Machine (SVM). Souza et al. [14] 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, which were discussed earlier, with different metrics. Moreover, the different concepts of violence prevent us from directly compare the existing methods. Such problems further sparked the MediaEval initiative as a form of standardizing validation in the field.

2.4.2 MediaEval Violence Scenes Detection Task

The MediaEval Benchmarking Initiative for Multimedia Evaluation [16] 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, 2013 and 2014, the task challenged participants to classify pre-segmented video shots from Hollywood movies as violent or not. A common trend among the VSD task attendants was to combine visual and auditory features, similar to previous works in the related literature. In 2015, Youtube video clips were added to the dataset and various teams have started venturing with deep convolutional neural networks, automating the process of extracting features and achieving promising results.

Vlastelica et al. [65] 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 [25] 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 [68]. 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), and 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 [50].

Yi et al. [72] 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 [66], 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 [6], 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. [12] trained a CNN model based on AlexNet [28] 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 stacked optical flows as input, and the last FC layer used as feature. 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 2.2 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.

Team	**CNN**	**Non-CNN Features**	**CNN+Others**
Fudan-Huawei[12]	23.5	16.5	29.6
MIC-TJU[72]	17.4	21.8	28.5
RFA[37]	14.2	7.7	8.2
RUCMM[27]	11.8	10.6	21.6
KIT[65]	10.2	8.6	12.9
NII-UIT[30]	-	20.8	26.8
UMons[54]	9.7	9.6	-
TCS-ILAB[5]	-	6.4	-
ICL-TUM-PASSAU[5]	-	14.9	-
RECOD[39]	-	11.4	-

Table 2.2: 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.

As we can see 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 [29] as external features, but no explanation was provided as to why fusing so many different features worked better.

Jin et al. [27] also proposed a solution based on visual and audio clues. The method consists of extracting audio features by collecting bag-of-audio-words features and utilizing a deep convolutional neural network technique to obtain more robust visual features. Instead of using the CNN features as one part of a BoVW approach, though, they used these features as the only visual features of their solution. Both audio and visual features are further concatenated in a late-fusion stage. Finally, a standard classifier is applied to determine if a violent scene occurred in a video.

2.4.3 Recent Studies

Since the 2015 Mediaeval VSD task, we have had other improvements on the field. Table 2.3 shows a comparison of the recent studies mentioned below, with a summary of their methods and results.

In 2016, Bilinski et al. [1] proposed an extension of the Improved Fisher Vector techniques and a sliding window method to localize the timeframe where the violent scene starts and ends. By representing the position of local features in the video and reducing the influence of motionless regions at the boundaries, they outperformed previous works in three critical datasets (Hockey Fights, Violence in Movies and Violent Flows), achieving more than 93% accuracy in all datasets, and improving the accuracy in the violent flows dataset from 85% to 96%.

The use of neural network have since risen, and in 2016, a three-stream DNN framework was proposed for detecting violence under the subject of person-to-person violence setup by Dong et al. [18]. They used a CNN to extract features on three streams: Spatial, using raw pixels; and temporal, using optical flow and acceleration flow maps. These features are then used in an LSTM network to further encode temporal information for a late fusion.

In 2017, Meng et al. [36] proposed a method that integrates dense trajectory and deep networks to capture more action information as input for the network. They used VGG-19 to extract spatial features from the video frames. For the temporal information, they compute the optical flow and finetune their pre-trained spatial model. With these new features, they compute the improved trajectories [68] and use a final Fisher Vector with an SVM classifier. This method achieved high accuracy in both the Violent Flows (92.5%) and Hockey Fights (98.6%) datasets.

While many rely on optical flow to describe temporal features, another method for extracting temporal information was proposed by Moreira et al. [38] in 2016. The Temporal Robust Features (TRoF) were designed to quickly compute an optimized amount of spatio-temporal interest points inspired by the still-image counterpart, SURF. This detector localizes a 3D point in a given scene that is the center of the movement, not only pinpointing its location in the scene, but also the time duration of the movement.

Senst et al. [55] also proposed another way of describing features for violence detection, based on Lagrangian local features and the BOW model. Lagrangian methods are used to describe nonlinear dynamic systems by a series of time-dependent fields. They are characterized by motion vector fields that describe the physical motion of particles. This method quantifies the properties of each particle while moving and reveals motion patterns through time. They adopted these concepts to a sequence of optical flow fields to characterize the motion within an image over time and used interest point detection based on SIFT to pinpoint the location of the movement within the image.

Sudhakaran and Lanz [58], though, went deeper in the LSTM approach. Using a convLSTM, which replaces the fully-connected layers of the LSTM with convolutional layers, they were able to store spatio-temporal information in the LSTM memory cells. Hence, they applied the frames of the input video sequentially to the network. After all the frames are applied, the hidden state of the convLSTM contains the representation of the input video. Finally, this representation is applied to a series of fully-connected layers for classification. Inspired by the Simonyan and Zisserman model for action recognition [56], they used the difference between adjacent frames as input for the convLSTM. This way, the network could model the changes of adjacent frames, instead of in the frame itself.

This is used as a crude approximation of the optical flow, in an attempt to avoid the computational complexity of calculating it. This method achieved better results than any other previous works in the Hockey Fights and Violence in Movies dataset, while achieving 94.57% accuracy in the Violent Flows one, behind only to Bilinski and Bremond [1]. A big upside of this method is that it needs significantly fewer parameters to optimize than a regular LSTM, while achieving a better result. Unfortunately, the datasets used were too limited.

In 2018, Hanson et al. [21] presented an approach based on a Bidirectional convolutional LSTM (BiConvLSTM). First, they encode each video frame as a collection of feature maps extracted from a forward pass through a VGG13 network [57] and then pass the obtained feature maps to a BiConvLSTM, which is an extension of a ConvLSTM and contains two main cell states: one for a forward sequence and other for a backward sequence in time. This way, the network has access to a long range of context in both directions of the time sequence. They also had high accuracy results in the Hockey Fights (96.96%), Violence in Movies (100%), and the Violent Flows dataset (92.18%).

Also in 2018, these datasets were then studied with a much simpler approach by Mumtaz et al. [41]. They used a pre-trained GoogleNet [61] with learned features from the ImageNet dataset and used it for transfer learning experiments. They removed the last dense fully connected classification layer that classifies 1000 ImageNet classes and replaced it with 2 classes for the Hockey and Movies datasets, discriminating violent/fight scenes from non-fight ones. This method resulted in a 99.28% accuracy in the Hockey Fights dataset and a 99.97% accuracy in the Movies dataset, demonstrating how powerful transfer learning can be.

Year	Author	Features/Classifier	Datasets Accuracy (%)		
			Hockey	Movies	Flows
2011	Bermejo et al.	BoVW - MoSIFT + SVM(HIK)	90.90	89.50	-
2016	Bilinski et al.	IFV / Sliding Window + SVM (χ^2)	93.70	99.50	96.40
2016	Dong et al.	Three-streams + LSTM	93.90	-	-
2017	Senst et al.	Lagrangian SIFT + SVM (χ^2)	94.42	94.95	93.12
2017	Meng et al.	CNN + Optical Flow + IDT	98.60	-	92.50
2017	Sudhakaran et al.	Convolutional LSTM	97.10	100.00	94.57
2018	Hanson et al.	Biconvolutional LSTM	96.96	100.00	92.18
2018	Mumtaz et al.	Transfer Learning from Inception	99.28	99.97	-
2019	Ullah et al.	CNN + 3D CNN	96.00	99.99	98.00

Table 2.3: Summary of the recent studies on violence detection mentioned.

In 2019, Ullah et al. [63] proposed a solution based on 3D CNNs. They first detect people in the video stream with a CNN model and then pass a 16-frame sequence as input for the 3D CNN to extract temporal features. This method allows for the C3D to work with a more streamlined sequence of frames and achieved the best results in the violent flows dataset, with a 98% accuracy, while maintaining 96% accuracy in the hockey dataset and a 99.99% accuracy in the violence in movies one. They also cross-tested each of their models on the other two datasets, but only two achieved more than 60% accuracy.

It is important to note that, while many of these methods use neural networks to achieve high accuracy results, the datasets tested are relatively small and limited. While solutions based on hand-crafted features already achieved more than 90% accuracy, many solutions based solely on CNN features could surpass them. This work builds on these promising results and use a more data-driven approach to a more generalist problem, working on a bigger dataset.

Chapter 3

Breaking Down Violence

Our research hypothesis, in this work, is that violence can be broken down into more precise, concrete, and objective concepts. With such concepts, it would be possible to aggregate specific features to detect a broader, more complex concept such as violence. In other words, we will rely upon divide-and-conquer modeling to approach the problem of violence detection.

The idea of breaking down violence into more specific concepts is not entirely new. In 2003, Cheng et al. [8] 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 types of violence. However, rather than fusing several general-purpose features that try to encapsulate the whole concept of violence, we can break down violence into smaller ones and find tailored features for each one, combining them later on for a more robust detection system.

This divide-and-conquer modeling has its challenges, as data for specific events (or concepts) is scarce. Much of the labeled data available for study treats violence as a general concept, and it is highly subjective as one scene can be characterized as violent for one person but not so for another. 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 toward a more robust representation of this problem.

3.1 Concepts of Violence

Our initial modeling adopts seven defined sub-concepts of violence: fights, gunshots, explosions, blood, fire, firearms, and cold arms. These concepts were chosen mainly for those that are more represented in the MediaEval benchmark [15], as well as their naming for each concept.

We then trained individual neural networks for each concept to learn their specific features. Each concept, however, has different classes of features associated with it. The concept of gunshots has robust auditory features, while fights rely upon dynamic, motion-related features. With this in mind, we can adjust what kinds of features are more critical for each concept, extracting all of them and learning a tailored fusion manifold to decide

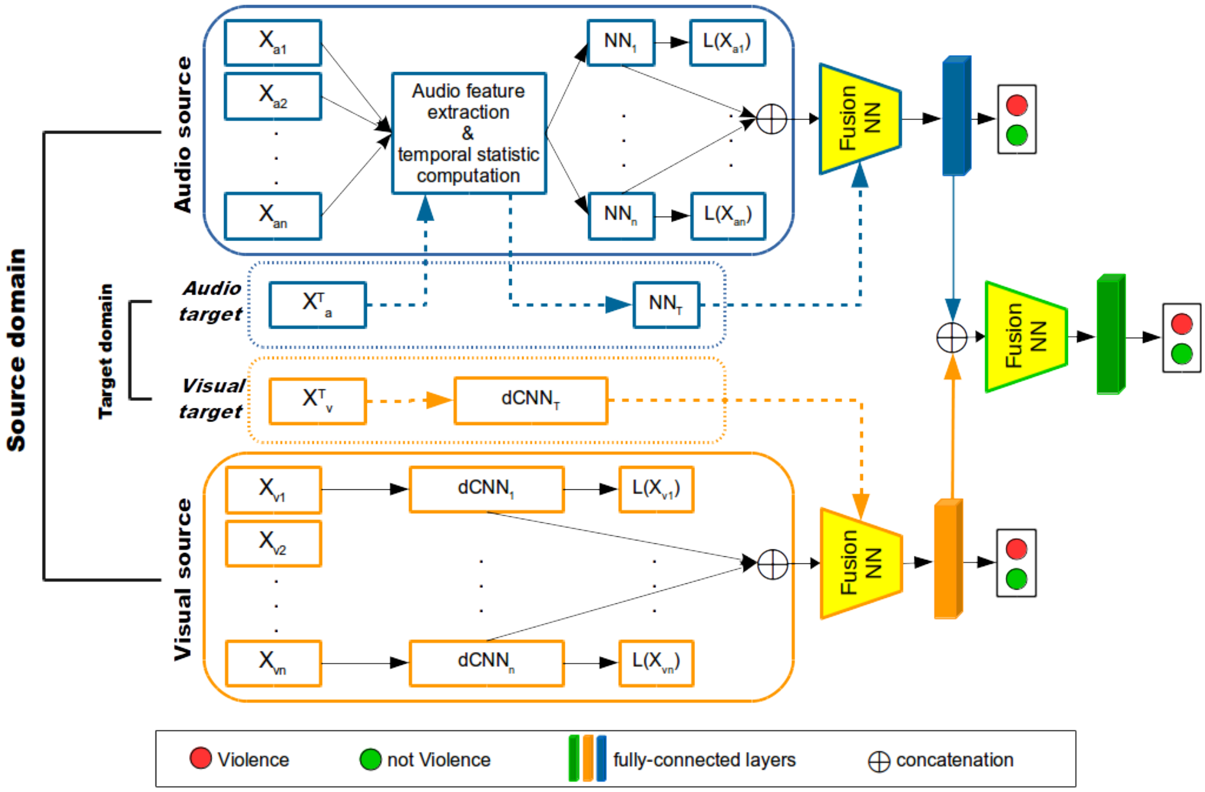

Figure 3.1: Pipeline of the multimodal feature fusion solution we devise in this work. Videos are described using a dCNN, whereas the extracted audio features are processed with a shallow neural network in the early-stage. In this pipeline, the source domain data for each audio and visual sources with different sub-concepts $(X_{a1}, X_{v1}, \ldots, X_{an}, X_{vn})$ is treated in parallel. All the features for each sub-concept are then combined and used as input for our fusion network, still maintaining the audio and visual pipeline separate. The features from both of these fusion networks are finally combined into a single feature representation and further used to train the final fusion network. The fusion network is trained in all its instances with the label for violence in general.

the weights to give to each class of features.

In order to classify violence within a movie, for example, we extract the frames for this movie and go through independent networks to classify them for each sub-concept. Audio and visual signals are also independent. This way, with our seven defined sub-concepts, each frame goes independently through 14 networks.

After this, we extract the features from each visual network and concatenate them into a single feature vector for our visual fusion network. We also do the same with the audio features.

Finally, we concatenate the features from each fusion network to create the input to our final fusion network, classify violence as a broader concept that is layered from smaller, more objective ones.

All fusion networks are trained with the annotation for the general violence, discussed as the one where a scene is violent if "one would not let an eight-year-old child see".

Figure 3.1 demonstrates the whole pipeline of our described solution on the violence detection problem.

The main takeaway is the idea of domain transfer learning from a broad, subjective concept into a collection of more specific, objective ones that help the machine understand the high-level target.

To achieve this, we need to extract features from each specific concept individually, hence our solution's first step consists of extracting the scene's static and dynamic visual features, followed by its auditory ones.

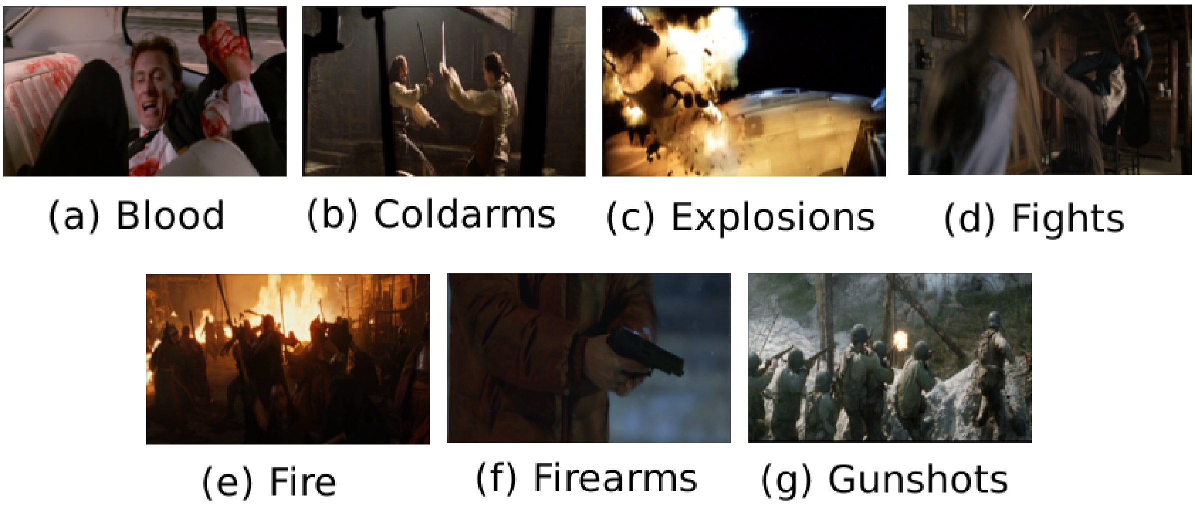

Figure 3.2: Representative frames of each violent sub-concept considered. Frames extracted from the movies in the VSD dataset: (a) Pulp Fiction; (b) Pirates of the Caribbean; (c) Armageddon; (d) The Wicker Man; (e) Eragon; (f) The Bourne Identity; (g) Saving Private Ryan.

3.2 Visual Static Features

The static visual features we used are mainly extracted from a dCNN such as inception v4 [60]. They represent our most basic and straightforward features, processing each training video frame by frame through a dCNN and extracting a pure data-driven feature set directly from the last average pooling layer. Though this method seems simple, our experiments later show that for sub-concepts represented by a purely visual aspect, such as the presence of blood or cold arms, these are the most accurate features to identify them. Figure 3.3 illustrates our input and output sizes for this network.

3.3 Visual Dynamic Features

The visual features' dynamic properties are an essential aspect of our work, as the most troublesome sub-concepts to detect convey the notion of movement and the passage of time, e.g., fights, explosions, and (indirectly) gunshots.

We approached this problem on two main fronts: (i) explicitly designing a time-based network architecture to consider motion; and (ii) modifying the input to capture the

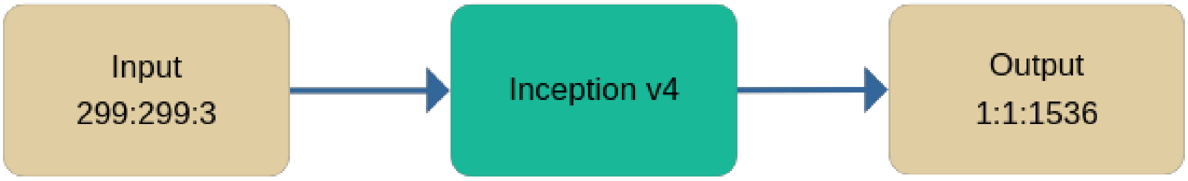

Figure 3.3: Our input and the output of the frames through the inception v4 architecture. No changes were made to the architecture, and we used the last fully connected layer as features when necessary.

notion of motion.

3.3.1 Time-based Architecture

In this approach, we studied the possibility of using different CNN architectures specializing in detecting time-based features.

3D convolutional networks are trained on frame sequences of video clips. This type of network can learn correlations directly in the 3D space, wherein the convolutional processing of a CNN can compute features from both spatial and temporal domains. This is achieved by constructing an input representing a fixed time-span, such as 30 frames (or one second, if we have a video frame rate of 30 frames per second). The 3D network then can process this block of frames as one and computes features from the temporal domain using the different frames in the same block.

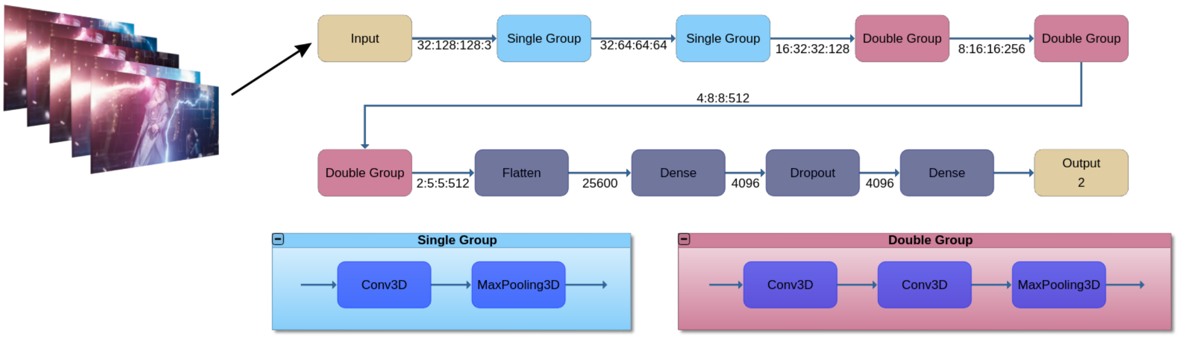

Figure 3.4: Architecture of the C3D network. The input is a full set of 32 sequential frames of size 128x128, forming an input of four-dimensional input of 32:128:128:3 (32 frames of 128 width by 128 height and 3 color channels). Each edge shows the size of the input for the following layer.

Figure 3.4 shows the C3D network architecture used, which comprises eight convolution layers divided into five groups. The first two groups are formed by a single convolution layer followed by a max-pooling layer, while the last three groups are formed by two consecutive convolution layers and a max-pooling layer.

Simultaneously processing a sequence of frames is crucial if we consider that a violent scene will not last for a long time and possibly contains a higher variance between the frames.

The CNN-LSTM model aims to learn spatial-temporal information jointly. It achieved promising results on the hand gesture recognition problem [62] and fitted our problem of detecting temporal features and their associated spatial features. This architecture receives as input a sequence of frames and acts mainly as a CNN network, with the last feature layer functioning as input to an LSTM layer, as illustrated by Figure 3.5.

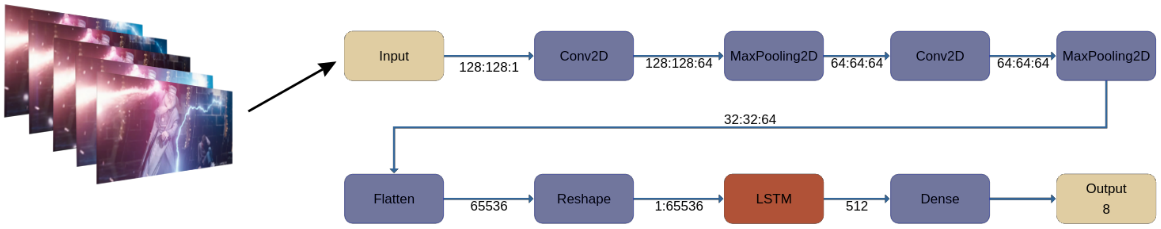

Figure 3.5: Architecture of the CNN-LSTM network. The input here is a sequence of consecutive frames that the LSTM layer uses to store information of previous frames.

3.3.2 Input Transformations to Capture Motion Information

In addition to processing temporal video cubes, we also studied different input adaptation types to convey movement information into a single image and use it as input to a 2D CNN. For this, we used three different approaches:

Optical Flow. This method comprises one of the most widely used features to represent motion. We use the Lucas-Kanade method of estimation for the optical flow, which assumes that the image's displacement between two consecutive frames is approximately constant within a given pixel neighborhood.

Optical Acceleration. As we aim to capture movement information from a wide variety of scenarios, it comes to reason that different types of violent actions will have different accelerations. One can expect that a gunshot scene will have a different acceleration than an explosion scene or a fight scene (considering the same video frequency sampling rate). We adopted the Farneback optical acceleration method [19] as the difference between two consecutive optical flows between three adjacent frames.

Figure 3.6 shows an example of the optical flow and optical acceleration inputs derived from a set of three consecutive frames.

TRoF and Input Transformations. Another way to represent movement in an image is to first detect a sequence of frames where the movement happens and identify a centralized point with a radius to make a cube of movement. That is what the temporal robust features detector (TRoF) [38] is capable of. With the detected cube, we can combine specific frames from this sequence to capture the event. We first run the TRoF detector through each movie sequence of a database and select the most relevant sequences of frames detected. These sequences comprise a center frame and a diameter representing the number of frames that encapsulate a specific kind of movement.

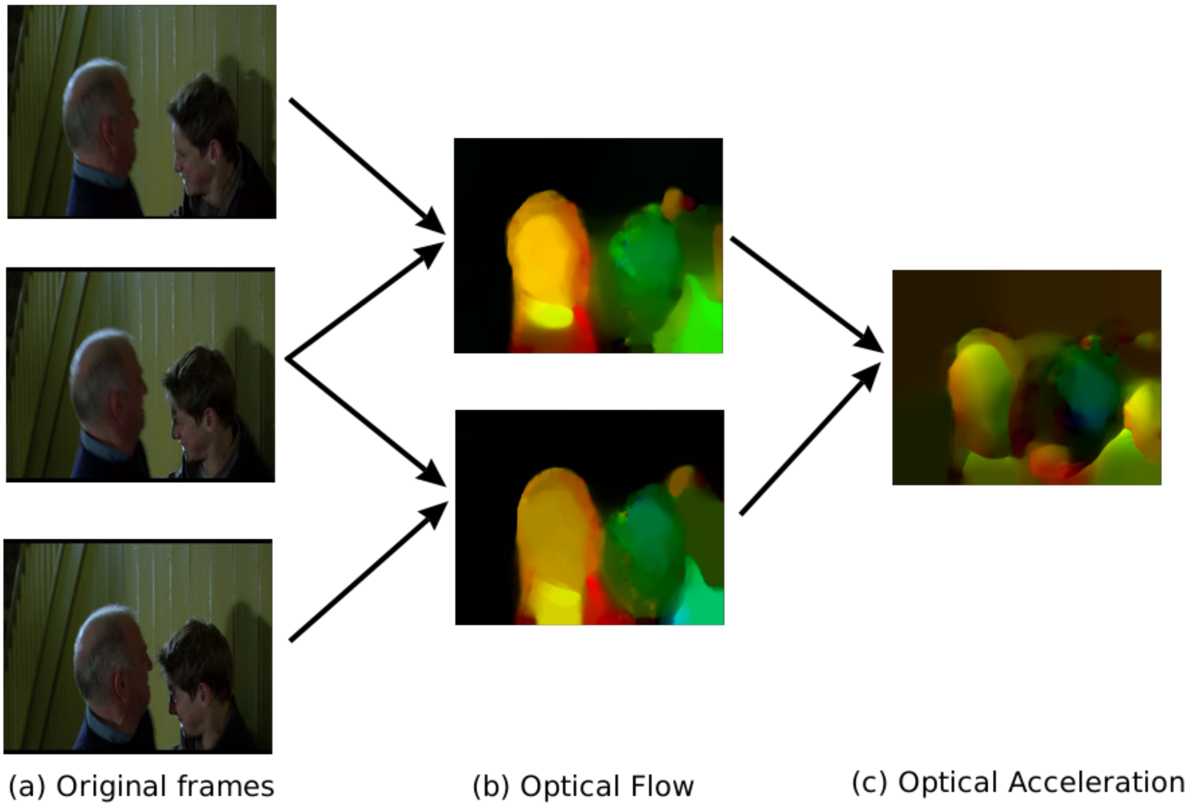

Figure 3.6: Visual representation of Optical Flow and Optical Acceleration. From a sequence of three consecutive frames we extract the two optical flows related, each combining two sequential frames. From the two consecutive optical flows, we then extract the optical acceleration.

As we seek to represent fights, explosions, fire, and gunshots, we are searching mainly for sudden movements, being it short-lived such as 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 represent a violent concept to have a broader range of movement throughout their frames than 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 designed neural network: One to capture the apex 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 3.7 illustrates how a sequence of frames detected by TRoF is used to form the combination of images discussed below.

- **Central Combination**: From the center of the sequence detected by TRoF, we combined it with the frame immediately before and immediately after. This forms a 3-tensor input (as a comparison, a typical natural image is also a 3-tensor input but with color channels, green, red, and blue, used instead). Thus, obtaining a single image representing this movement to feed the network. This type of combination focuses on the movement's climax, using only the sequence's three central frames.

- **Extremities Combination**: For each sequence of frames detected by TRoF, we combined its center with its start and end frames also in a 3-tensor input to represent the whole movement detected in a single input. This type of combination captures the changes between the start of the movement and its end, using the central frame as a bridge. Each channel in the tensor is a grayscale image representing the specific frame adopted.

- **Averages Combination**: For each sequence of n frames detected by TRoF, we combined its center with the average of the $n/2$ preceding frames and the average of $n/2$ frames after it. We again assigned each resulting grayscale-converted image to a 3-tensor input, respectively. This type of combination was made to capture the flow of movement throughout the whole sequence, using all the sequence frames to represent how the movement occurred.

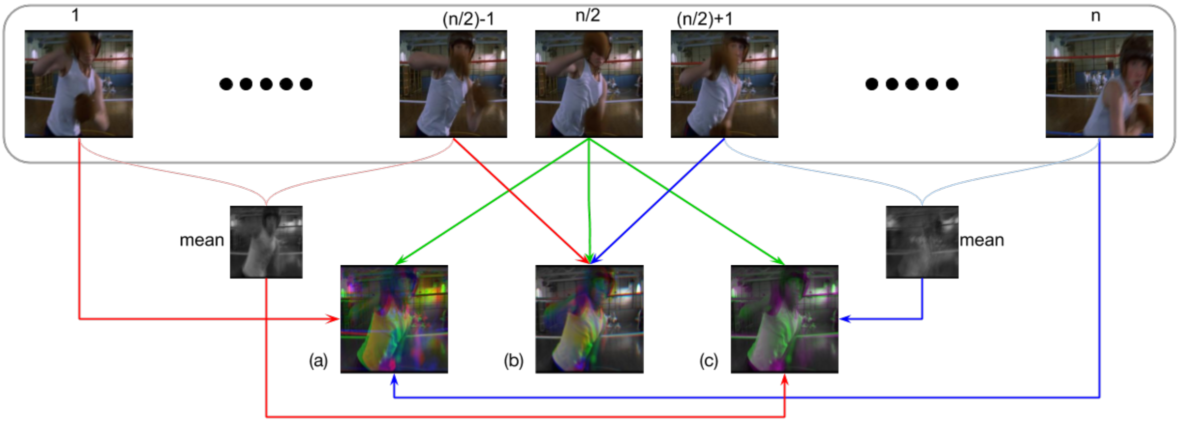

Figure 3.7: Overview of how the combinations are constructed from a single sequence detected by TRoF. (a) The first and last frames of the sequence are combined with the center to form the Extremities combination. (b) The three central frames are put together as each of the color channels of the Central 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 the blue channel frames.

3.4 Audio Processing

Typically, violence is subjective, thus for some sub-concepts (e.g., gunshot, explosion, fight), it might be possible that auditory sensation better captures such a notion. However, the visual features might be the only available cue for some of the other sub-concepts, such as *blood*. Audio features can help boost the performance accuracy of violence detection tasks as a complementary feature. Undeniably, for the available audio data in the violence detection problem, almost all sub-concepts contain noise such as background sound and people talking. In this task, we consider audio descriptors robust to noise and background clutter. The descriptors can be processed through raw audio waveforms in the frequency domain. This means the frequency domain can be directly calculated on the spectral distribution of a given audio waveform. We opt to rely upon expert knowledge and handcrafted features in this case because a data-driven solution trained on audio-related violence would be hard to train. The lack of available annotated data is different from visual-related features, which presents a better data availability scenario.

We describe the used audio feature representation and prediction model for the violence detection problem. We tackle the audio modality in two steps processing. We first extract the audio features by leveraging some standard audio feature extracting methods. For a given raw audio waveform, $x(t)$, we split it into a series of I temporal windows, and we extract the features for each window. We then apply statistical methods to the results generated in the first step. This approach can generate a robust and compact version of the extracted audio features and reduce clutter and background noise. In the following, we describe each step in detail.

3.4.1 Standard Acoustic Features

We consider four standard audio feature extractors for subjective violence detection. Figure 3.8 depicts the raw waveform of a given audio clip, including the spectrogram visualization for each audio feature on various sub-concepts. We can observe, for some sub-concepts, such as *blood* and *Fire*, the spectrogram has no semantic audio representation. However, some sub-concepts such as *Gunshot* and *Firearm* stand out, with similar acoustic characteristics appearing on the spectrograms.

Mel-Frequency Cepstral Coefficients (MFCCs) [3]: This is a commonly used audio feature in speech recognition problems. For a given audio clip, it first computes MFC as a power spectrum by applying a linear cosine-transformation on a nonlinear Mel-scale (*filterbank*) frequency, and it represents the coefficients obtained through the calculation of MFC.

Chroma Short-Time Fourier Transform (C-STFT) [20]: Chroma feature (a.k.a., chromagram) is mainly extracted to capture harmonic characteristics of an audio waveform in a short-time window. It first computes the magnitude spectrum through the short-time Fourier transform. This feature can handle the tone (pitch) of a sound that appeared within the audio clip.

Mel-Spectrogram (MS) [3]: This feature is represented as a more straightforward and lower-level form of frequency to imitate the function of the human ear. It filters the components of frequency through log-Mel filter banks in the spectrogram.

Spectral Contrast (SC) [26]: This feature is generated as the decibel difference between peak and valley frequencies in the spectrogram. It is widely considered when the ratio between signal and noise is relatively large and can significantly reduce the noise for

Feature	#coefficients
MFCC (f^{mfcc})	40
C-STFT (f^{c-stft})	12
MS (f^{ms})	128
SC (f^{sc})	7

Table 3.1: Features extracted from each raw audio clip.

a given audio waveform.

The audio features obtained from the above extractors are finally represented as the feature set associated to the i-th time window and defined as

$$f_i = \{f_i^{mfcc}; f_i^{cstft}; f_i^{ms}; f_i^{sc}\} \tag{3.1}$$

Table 3.1 reports the number of coefficients obtained for each audio feature.

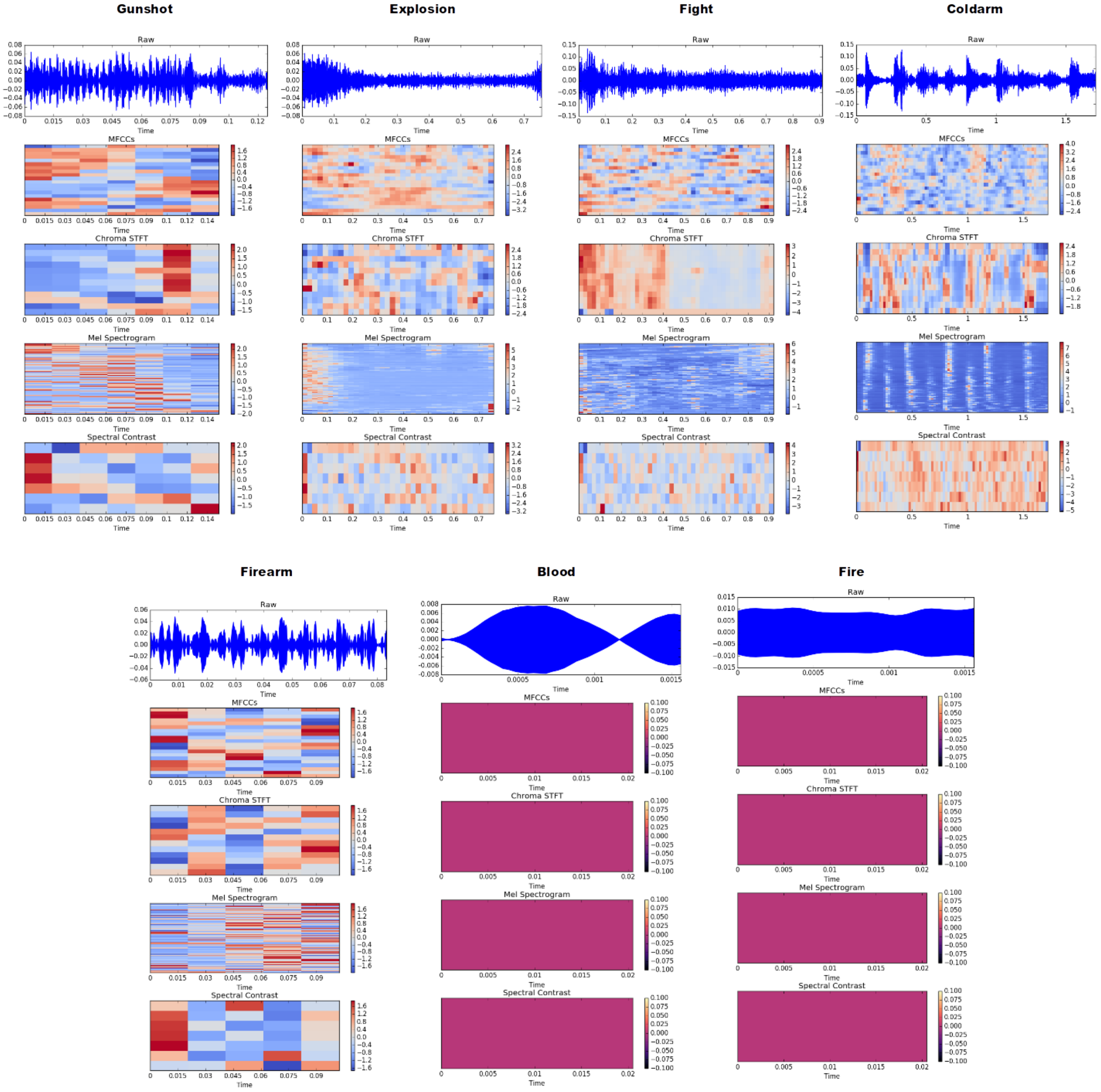

Figure 3.8: Visualization of raw-audio waveforms and their corresponding spectrograms for the set of extracted audio features in various violence sub-concepts.

3.4.2 Temporal Statistics Calculation

Inspired by the work of Borrelli et al. [4], to obtain a more discriminant feature vector over the set of extracted audio features, we apply four temporal statistics to I extracted feature vectors of f_i. This provides an additional set of information for the subsequent learning stage. This feature vector is subsequently projected onto a lower-dimensional space and represented as a compact form of the four extracted audio features in f_i. This can also significantly reduce the computational cost and memory footprint and eliminate features not relevant for a particular sub-concept.

Given a set of I feature vectors f_i, we compute the per-feature average, standard deviation, maximum, and minimum value as

$$f^{\mu} = \frac{1}{I} \sum_{i=1}^{J} f_i, \tag{3.2}$$

$$f^{\sigma} = \sqrt{\frac{1}{I} \sum_{i=1}^{I} (f_i - f^{\mu})^2}, \tag{3.3}$$

$$f^{M} = \max_{i \in I} f_i, \tag{3.4}$$

$$f^{m} = \min_{i \in I} f_i. \tag{3.5}$$

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

$$f^{tot} = [f_i^{\mu},\, f_i^{\sigma},\, f_i^{M},\, f_i^{m}], \tag{3.6}$$

where f^{tot} denotes a feature vector of $4 \times 4 = 16$ elements.

3.4.3 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.

3.5 Fusion Neural Network (NN)

We design a multilayer perceptron neural network to identify the ultimate decision on violence by itself — aiming to leverage and integrate the sub-concepts of violence. In this vein, we present a straightforward strategy that can independently learn an embedding of the feature maps obtained from earlier stages (i.e., audio and visual modalities). We treat the network as a binary classification problem to determine whether violence is perceived in a video. This solution allowed us to gain a better trade-off between efficiency and performance.

Our designed fusion NN solution is involved in three tasks of this work on the hand, and for each task, the network is trained independently. First, feature maps obtained from each modality (both audio and visual) are fed to an individual fusion network and learn the concept of violence. A third fusion network is then used to combine feature vectors obtained with audio and visual detectors trained on specific kinds of violence to detect the presence of violence. This technique enables us to transfer domain knowledge from sub-concepts to the main concept, expecting better generalization in *unknown* target domains. We pass the feature vector through a standard *MinMax* normalization step before using it as input for the network.

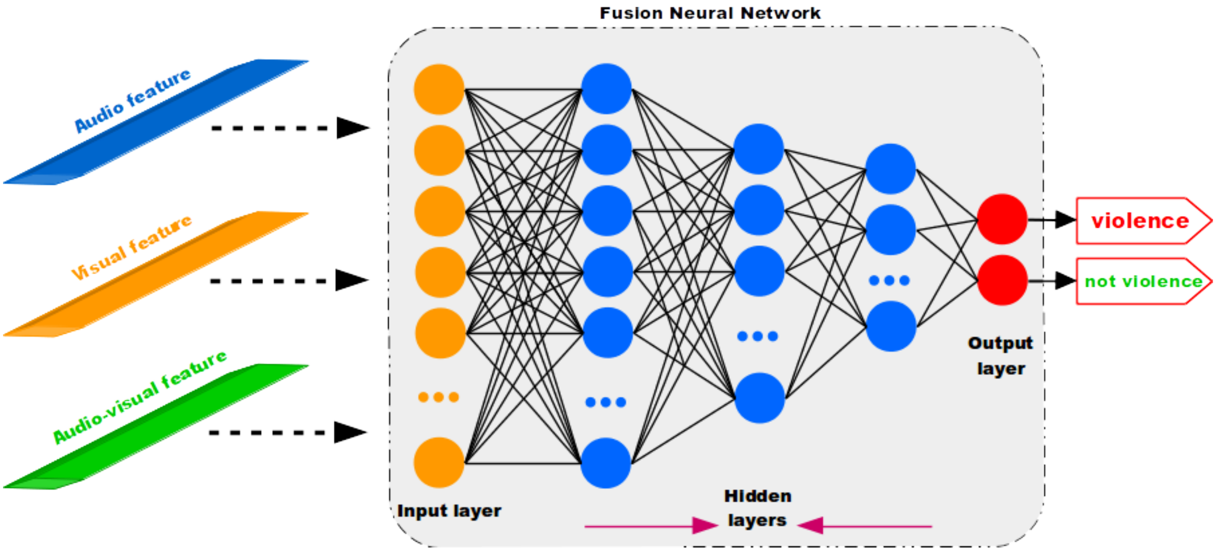

Figure 3.9: The structure of our fusion neural network.

Figure 3.9 shows the structure of our fusion neural network in detail. The designed network mainly consists of three hidden layers. We applied a grid search method on the number of hidden layers and, accordingly, the number of neurons for each hidden layer. The network's best performance is achieved by choosing 512, 128, 32 neurons, respectively, from the first to last hidden layer. The network is utilized as a feed-forward structure with a back-propagation algorithm for its training process.

Chapter 4

Experiments and Results

In this chapter, we detail our experiments and metrics used to validate them and a detailed view of the datasets used. First, we explore the main dataset and the metrics used to analyze our results. Then, we discuss these results and how our research iterate over them to build a better method to detect violence. We go from using still frames to find a robust representation of motion to incorporate audio information.

4.1 Dataset and Metrics

For our experiments, we adopted mainly the MediaEval-2013-VSD dataset [15], which contains 25 Hollywood movies of diverse genres. This is an early iteration of the dataset, that later incorporated YouTube clips in its test-set for the 2015 edition [16]. Later, we also tested our method with these additional clips.

The definition of violence used is that a scene is violent if “one would not let an eight-year-old child see it”. The dataset is released with separated training and testing partitions. The training set includes 18 movies, while the test set comprises seven movies.

The annotations were carried out by three expert annotator groups. At first, two groups conducted all the annotations independently. Then, a third master group merged the two sets of annotation and made decisions for the inconsistent cases. The subjective definition of violence required panel discussions for borderline cases. Further information on the evolution of the dataset as well as more insight of the annotation method can be found in the latest work by Constantin et al. [10].

Remarkably, for all the movies in the MediaEval-2013 dataset, only 20% of the shots have been categorized as violent. Although the dataset provides annotations for individual concepts (e.g., blood, fights), these annotations are only available for the training set.

Given that only the movies in the official training set contain annotations for the sub-concepts, we partitioned this set into its own training, validation, and test sets. This allows for a better comparison between different methods that aim to classify specific concepts. Table 4.1 details the movies used in each partition.

The VSD task motivation was to foster 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

		Movie	Length (m)	Violence (%)
Dev Set	Training	Armageddon	145	7.78
		Billy Elliot	106	2.46
		Dead Poets Society	124	0.58
		Eragon	100	13.26
		Harry Potter 5	133	5.44
		Midnight Express	116	7.12
		Pirates of the Caribbean	137	18.15
		Reservoir Dogs	95	30.41
		Saving Private Ryan	162	33.95
		The Bourne Identity	114	7.18
		The Wicker Man	98	6.44
		The Wizard of Oz	98	1.02
	Validation	Fight Club	133	18.83
		Leon	106	16.36
	Test	I am Legend	96	15.64
		Independence Day	147	13.13
		The Sixth Sense	103	2.00
Test Set		Fantastic Four	102	20.53
		Fargo	94	15.04
		Forrest Gump	136	8.29
		Legally Blond	92	0.00
		Pulp Fiction	148	25.05
		The Godfather	170	5.73
		The Pianist	143	15.44

Table 4.1: Details of the Mediaeval 2013 VSD task dataset and how was used for the experiments, with Length of videos in minutes and percent of violent scenes within. 'Dev Set' and 'Test Set' are the official segmentation of the database, while Training, Validation, and Test in the Dev Set are our segmentation to validate each concept individually. The movie *Kill Bill 1* had annotation problems and was later excluded from the dataset.

return the largest number of violent shots at the first positions of the top-k retrieved shots. For that, the competition suggests using the Mean Average Precision (MAP) at the 100 top-ranked violent shots (MAP@100) as the official evaluation metric.

To calculate this metric, first, each movie is segmented into different shots, which are roughly equivalent to the movie scenes. This segmentation is provided by the dataset at the frame level, indicating the first and last frame of each shot.

MAP is calculated by first taking the average precision scores for all movies in the test set and then getting the arithmetic mean of these scores, as shown in Equation 4.1:

$$MAP@k = \frac{1}{q}\sum_{i=1}^{q} AP@k(i), \tag{4.1}$$

where k is the quantity of shots within the rank of retrieved shots, in this case, $k = 100$, and q is the quantity of system queries for obtaining ranked violent shot lists. In the VSD case, $q = 7$, which is the number of movie titles within the test set. $AP@k(i)$, in turn, is the average precision of the i-th query(movie), when returning a k-shot ranked list, as it follows:

$$AP@k(i) = \frac{1}{q} \sum_{j=1}^{k} (precision(i, j), \tag{4.2}$$

where $precision(i, j)$ is the system precision when retrieving the top-j violent shots, within the i-th query (movie).

When calculating the MAP@100, we only use the precision scores of the 100 highest ranked shots. This way, a solution is considered better than another if it presents a higher MAP@100 value as it indicates that such a solution returns fewer false-positive shots in the first positions of a 100-violent-shot ranked answer.

Since our objective is to simply detect violence, without ranking or taking into consideration only the most violent scenes, we mainly report our solutions' balanced accuracy at the frame-level.

Since there is a high unbalance of classes within the dataset, as shown in Table 4.1, most of the scenes of each movie are annotated as non-violent, all our experiments adopt a dataset balancing approach.

For each movie, we count the number of frames annotated as violent and randomly select the same number of non-violent frames, respecting the shot-segmentation of the dataset and ignoring frames that are all black, due to fade or cut effects and the beginning and end credits.

The accuracies we report throughout the experiments are all derived from this balanced dataset.

To validate our method in a non-VSD dataset, we also test our method with the NTU-CCTV fights dataset [49], which consists of 1000 Youtube videos with more than 17 hours of footage containing different types of fights. This dataset contains 280 CCTV videos ranging from 5 seconds to 12 minutes each (average of 2 minutes), totaling 8.54 hours, while the remainder of the dataset consists of 720 shorter videos (45 seconds on average) extracted mainly from mobile cameras.

4.2 Validation and Results

In the following, we present the step-by-step search for a reliable method that can incorporate the many different kinds of violence to give us a classification for such a broad concept. In order to classify violence, we first classify each of the defined sub-concepts individually. Here we can already see that some sub-concepts convey movement, such as fights or explosions. In Section 4.2.1, we explore different ideas on how to represent motion information to a network by manipulating the input or using architectures that incorporate this kind of feature. In Section 4.2.2, we use the best visual features for each sub-concept in addition to audio features to complement the information available for our fusion network. Then, in Section 4.2.3, we start to explore how this method holds up when using a separate dataset to train a specific sub-concept while still using features from the other sub-concepts trained with the Mediaeval VSD dataset.

4.2.1 Visual and Motion Detection

To find the best way to represent each sub-concept visually, we compare the results from the different strategies of *Input Transformations* and *Time-based Architecture Modeling* alongside a straight *2D Static Raw* input classification. Our baseline is to run the Inception-v4 with raw frames for each sub-concept and for the direct violence classification.

Here and in all of the following experiments, we present the balanced accuracy results for each sub-concept, and a comparison between three methods of classifying violence:

- **Violence**, which is a direct classification of the frames, using the subjective definition of the VSD task. This method uses an independent network to classify violence as if it was yet another sub-concept, using the violence annotation of the dataset as guideline and serving as baseline for our fusion method.
- **Concatenation** is the classification done with the features of each sub-concept put together. That is, for each frame of a movie, we extract the features from the networks of each sub-concept and concatenate them into a single feature vector for that frame. This is the input for the classifier.
- **Fusion Network** is the classification that uses the concatenation feature vector as input for our proposed fusion network.

All of these classify violence based on the dataset annotation of violence that states that a scene is violent if "one would not let an eight-year old child see".

Input Transformations

The first experiments with input manipulation were based on the TRoF Combinations. For these, we compared the results of the classification done by the inception-v4 own softmax layer with a SVM classifier that received the features from the last fully connected layer of the network. We tested three different kernels for the SVM: i) Linear; ii) RBF; and iii) Power Mean (PmSVM).

Of the kernels tested, the one that performed better was the PMSVM [69]. This method uses the Power Mean family of additive kernels, going from the X^2 to the Histogram Intersection Kernel (HIK). It first performs a Principal Component Analysis (PCA) to reduce the dimensionality of the features and later performs an optimized grid search for the best parameters and kernel.

Tables 4.2, 4.3, and 4.4 show the results for each SVM used. It is notable that the fusion of sub-concepts performed better than the subjective definition of violence in all of the experiments. It is also notable that the Extremities combination was the best type of combination overall, independent of the SVM kernel.

Compared to the simple softmax layer of the Inception-v4, though, the accuracy of the SVM does not give many advantages. For example, the best fusion accuracy achieved by any tested SVM method was the Extremities combination with the PMSVM, 73.1%. The same combination input when classified by the inception-v4 achieved a 73.3% accuracy.

	Linear SVM Classification			
		TRoF Combinations		
	Raw Frames	Central	Extremities	Average
Blood	**59.0**	57.6	58.5	54.7
Cold Arms	**72.5**	61.9	65.0	67.0
Explosions	65.4	72.1	**73.2**	71.5
Fights	54.4	69.4	**71.0**	67.8
Fire	**73.9**	62.7	70.3	68.9
Firearms	**60.1**	57.3	56.6	59.3
Gunshots	56.6	52.4	**61.9**	53.2
Violence	**68.3**	67.3	63.1	64.8
Concatenation	69.4	70.3	**72.7**	70.1

Table 4.2: Balanced classification accuracy (in percentage) for all TRoF Combinations with a linear kernel SVM. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Fusion' does not include the 'violence' concept.

	RBF Kernel SVM Classification			
		TRoF Combinations		
	Raw Frames	Central	Extremities	Average
Blood	54.3	58.7	**61.1**	60.1
Cold Arms	**71.1**	63.7	65.8	66.1
Explosions	73.1	69.5	**74.3**	73.1
Fights	72.3	70.7	**74.2**	73.6
Fire	**76.7**	72.4	72.4	73.7
Firearms	**59.4**	58.7	58.6	58.8
Gunshots	61.1	63.9	**65.3**	57.0
Violence	**68.5**	67.4	63.2	63.1
Concatenation	69.2	69.4	70.1	**70.3**

Table 4.3: Balanced classification accuracy (in percentage) for all TRoF Combinations with a RBF kernel SVM. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Fusion' does not include the 'violence' concept.

	PMSVM Classification			
		TRoF Combinations		
	Raw Frames	Central	Extremities	Average
Blood	61.8	63.5	**69.6**	62.8
Cold Arms	68.7	60.5	**68.9**	63.5
Explosions	**71.0**	68.7	67.9	65.9
Fights	72.3	72.1	73.5	**74.3**
Fire	57.7	55.2	**61.6**	55.9
Firearms	**69.8**	64.4	65.9	63.8
Gunshots	**73.9**	78.4	72.6	72.0
Violence	69.2	66.2	**69.4**	66.7
Concatenation	71.3	70.6	**72.1**	71.9

Table 4.4: Balanced classification accuracy (in percentage) for all TRoF Combinations with a Power Mean kernel SVM. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Fusion' does not include the 'violence' concept.

The last classification methods analyzed were the classification done by the inception-v4 and our fusion network, detailed in the Section 3.5.

Table 4.5 shows the results of the classification done directly by the inception-v4 for each sub-concept as well as two types of fusion: i) a simple concatenation of the features extracted from the last fully-connected layer of each sub-concept; and ii) using these same features as inputs to our fusion network.

The results show that our fusion network is better than the simple concatenation of features for the inception-v4, in all types of input tested it achieved a better accuracy, peaking at 74.4% for the Raw Frames. We can also see that the individual results for the classification of each sub-concept have an overall better accuracy when done directly by the neural network. Notably, the cold arms concept, which for the first time achieved more than 80% accuracy in all of the experiments. The TRoF combinations also performed worse than the raw frames with no movement information associated. The best one being the extremities combination, which had a 73.2% accuracy for the Fights sub-concept.

In search of a different representation of movement, we used Optical Flow and Optical Acceleration. Table 4.6 also shows our results with these experiments, alongside the results for the Network Classification as shown in Table 4.5.

	Network Classification			
		TRoF Combinations		
	Raw Frames	Central	Extremities	Average
Blood	**74.2**	73.8	70.6	69.9
Cold Arms	**81.6**	64.4	71.5	70.8
Explosions	**79.4**	71.3	75.9	74.2
Fights	73.1	70.4	**73.2**	71.5
Fire	70.1	**70.7**	70.6	69.9
Firearms	**60.8**	58.4	58.5	59.1
Gunshots	**69.3**	66.8	66.4	64.0
Violence	66.7	**68.4**	62.8	65.3
Concatenation	72.4	70.5	73.3	**73.6**
Fusion Network	**74.4**	74.1	73.8	74.2

Table 4.5: Balanced classification accuracy (in percentage) for all TRoF Combinations using a neural network classification layer. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Concatenation' is the classification done by the inception-v4 using the features from the last fully-connected layer of every sub-concept network. 'Fusion Network' is the classification done by our fusion network with the same features. These fusions do not include the 'violence' concept.

From all these experiments, we can see that for every type of input the results for combining sub-concepts (the results from the 'Fusion' row) are consistently better than just classifying violence solely through its subjective definition (the 'Violence' row).

We can also see that using just TRoF combinations leads to better results for the fusion than the full-optical flow inputs. This result seems to indicate that the TRoF combinations, when used together in a fusion network, might be better suited to use as inputs than the optical flow.

However, no combination from TRoF was better than the other types of inputs (Raw frames or optical flow) with the individual concepts. This result can be explained by the nature of the concepts and how the inputs are designed. For example, for blood and cold arms, where there is little movement, it was expected that a movement-based input would not provide better results. For fights, fire, and gunshots, the movement provides

		TRoF Combinations			Optical Flow Inputs	
	Raw Frames	Central	Extremities	Average	Flow	Acceleration
Blood	**74.2**	73.8	70.6	69.9	68.3	58.2
Cold Arms	**81.6**	64.4	71.5	70.8	61.9	76.5
Explosions	**79.4**	71.3	75.9	74.2	77.8	70.6
Fights	73.1	70.4	73.2	71.5	**76.8**	74.3
Fire	70.1	70.7	70.6	69.9	68.1	**71.2**
Firearms	60.8	58.4	58.5	59.1	62.3	**66.8**
Gunshots	69.3	66.8	66.4	64.0	63.6	**73.1**
Violence	66.7	**68.4**	62.8	65.3	65.0	58.7
Concatenation	72.4	70.5	73.3	**73.6**	67.9	72.1
Fusion Network	**74.4**	74.1	73.8	74.2	68.2	72.8

Table 4.6: Balanced classification accuracy (in percentage) for all types of input manipulation visual dynamic features with corresponding raw static frames accuracy for comparison. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. Individual sub-concept classification was done by inception-v4. 'Concatenation' is the classification done by the inception-v4 using the features from the last fully-connected layer of every sub-concept network. 'Fusion Network' is the classification done by our fusion network with the same features. These fusions do not include the 'violence' concept.

additional information, hence the raw frames' results were worse. The optical flow ability to convey this kind of information could be the deciding factor.

The exceptions to this reasoning are the concepts of explosions and firearms, but we can look at the dataset itself for some explanation. Many explosions in the movies within the dataset receive a close-up treatment, filling the whole screen and visual effects that prolong their duration, making it easier to detect it with raw frames than with movement detectors, even when the explosion does not involve fire itself. On the other hand, firearms are often portrayed alongside gunshots rather than as a static object, explaining why the optical flow inputs captured this better than the raw frames.

Figure 4.1 show some examples of positive samples for firearms and explosions that were classified as positive by both the raw frames and optical flow inputs and which were classified as positive only by one of them. Both classifiers correctly detected the more straightforward scenarios, but the harder ones were split, making it challenging to decide which one is better.

Time-based Architecture

With these results, we then ran experiments with time-based architectures. We used the raw frames with two distinct networks: C3D and CNN-LSTM, as detailed in Section 3.3.1. Table 4.7 shows our results for each of the sub-concepts and the same comparison between classifying violence directly and through our fusion network. From Table 4.7, we can see that once again, the results for the fusion of concepts are classified better than just using the subjective definition directly. The best results for the direct definition of violence were 68.3%, while our fusion had 69.2% accuracy. Overall, the C3D architecture performed better, especially with the optical acceleration input, but the results with a deep CNN — Inception-v4 — outperformed the best results with C3D.

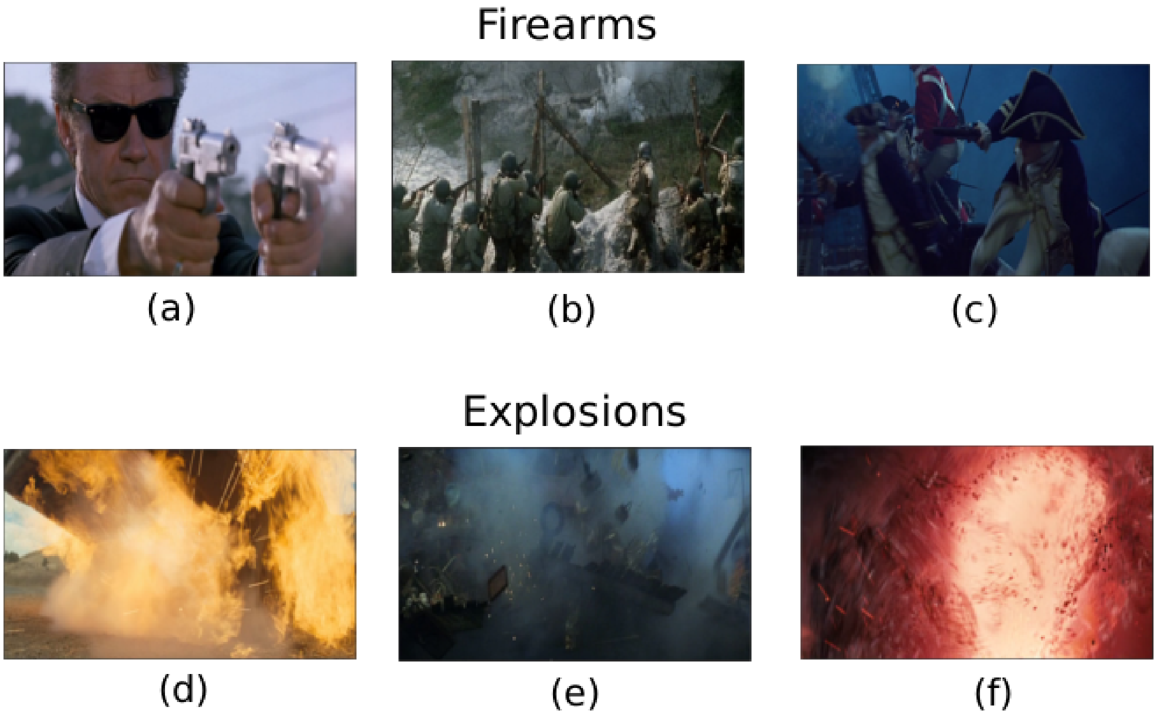

Figure 4.1: Some positive examples for firearms and explosions. The first column, scenes (a) and (d) were correctly classified by both the raw frames and the optical flow inputs. The middle column, scenes (b) and (e) were correctly identified only by the optical flow input. The last column, scenes (c) and (f) were only detected correctly with the raw frames input.

However, one thing to note is that even if the results with a dCNN were better, not a single type of input led to better results across all sub-concepts. Namely, the sub-concepts that convey motion were expected to perform better with networks that use temporal information, but both *Explosions* and *Fights*, the most prominent sub-concepts associated with motion, not only perform better with the dCNN, but the best results with C3D and CNN-LSTM were from the Raw Frames input. This could mean that the Optical Flow for these concepts may interfere with how these networks calculate temporal information.

4.2.2 Combined Visual and Audio Features per Concept

Since our fusion network relies on the extracted features from the last layer of the sub-concept classifier network, it allows for a combination of features from different networks. We can select the networks' features with the best performance accuracy for their specific sub-concept and use them as input for our fusion network. With this, we expect to use our best individual classifiers as feature feeds to a fusion network that can analyze a scene with information from multiple violence sub-concepts and learn how they combine to give us a broader classifier for violence.

With this reasoning, the best performing networks for each sub-concept come from the Inception-v4 architecture, shown in Table 4.6. We also take into consideration that

	Raw Frames		**Optical Flow**		**Optical Acceleration**		**Central TRoF**	
	C3D	CNN-LSTM	C3D	CNN-LSTM	C3D	CNN-LSTM	C3D	CNN-LSTM
Blood	58.0	57.2	59.2	60.2	60.2	58.2	56.8	57.4
Cold Arms	58.3	54.2	66.5	66.2	75.3	69.0	63.5	64.0
Explosions	77.1	61.4	66.4	64.8	73.0	68.1	69.4	69.0
Fights	70.5	53.7	68.0	66.9	65.4	61.7	68.2	66.3
Fire	60.2	55.6	60.3	61.3	64.9	61.9	62.7	63.6
Firearms	61.0	60.3	63.2	65.0	66.5	62.3	62.0	63.4
Gunshots	65.3	56.8	62.6	64.1	68.6	66.8	63.8	64.5
Violence	62.3	55.9	58.1	58.6	68.3	63.6	62.3	60.1
Fusion Network	67.3	63.3	67.2	64.8	69.2	64.2	66.8	65.2

Table 4.7: Balanced classification accuracy (in percentage) for each time-based architecture for all types of input transformation, dynamic features, and raw static frames. All seven concepts were trained and tested with the same subsets of movies. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Fusion' does not include the 'violence' overall concept.

since we are combining visual and audio features, it is possible that a visual feature that performed worse individually can be significantly improved by the addition of audio features, resulting in a better performance. Through our experiments, though, we found that not to be the case. Table 4.8 shows one example of using the best features from the C3D, which performed worse than the Inception-v4, combined with our audio features. The features we used are from the experiments highlighted in Table 4.7. Even though our final fusion results were better than using only the visual or only the audio features, it was not better compared with our classification of visual features using Inception-v4, reaching a balanced accuracy of 74.1% compared to our previously discussed 74.4%.

	Best C3D Features	**Audio Features**	**Visual + Audio Features**
Blood	60.2	61.0	**62.1**
Cold Arms	75.3	66.9	**78.4**
Explosions	**77.1**	65.3	76.8
Fights	70.5	61.9	**72.1**
Fire	64.9	**67.8**	62.3
Firearms	**66.5**	62.4	65.6
Gunshots	68.6	70.7	**70.7**
Violence	68.3	**72.8**	71.4
Fusion Network	69.2	63.0	**74.1**

Table 4.8: Balanced classification accuracy (in percentage) for the best C3D and audio features for each sub-concept and their result for the fusion of visual and audio features. All seven concepts were trained and tested with the same subsets of movies. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Fusion' does not include the 'violence' concept.

For this experiment, we settled with the features from the best-performing networks for each sub-concept. Since the C3D and CNN-LSTM networks underperformed, we used the best results with Inception-v4 architecture from Table 4.6 as discussed in the previous section, that is, for each sub-concept, we used the following inputs:

- **Blood**: Raw Frames
- **Cold Arms**: Raw Frames
- **Explosions**: Raw Frames
- **Fights**: Optical Flow
- **Fire**: Optical Acceleration
- **Firearms**: Optical Acceleration
- **Gunshots**: Optical Acceleration

	Best Visual Features	**Audio Features**	**Visual + Audio Features**
Blood	**74.2**	61.0	66.5
Cold Arms	81.6	66.9	**83.2**
Explosions	**79.4**	65.3	77.3
Fights	76.8	61.9	**77.2**
Fire	**71.2**	67.8	64.5
Firearms	66.8	62.4	**73.3**
Gunshots	73.1	70.7	**74.5**
Violence	68.4	**72.8**	72.1
Fusion Network	75.3	63.0	**78.5**

Table 4.9: Balanced classification accuracy (in percentage) for the best visual (from Inception-v4) and audio features for each sub-concept and their result for the fusion of visual and audio features. All seven concepts were trained and tested with the same subsets of movies. The 'violence' concept refers to the MediaEval VSD definition of a violent scene. 'Fusion' does not include the 'violence' concept.

Table 4.9 shows our results using the individual best feature for each sub-concept. We obtained the highest fusion results using the best features for each sub-concept. With the audio features discussed in Section 3.4, we performed another set of experiments: classifying each sub-concept using only audio and then combining the best visual features with such audio features. Table 4.9 shows results for both experiments.

Interestingly, with only audio features, we had our best results classifying violence from the direct definition (the more subjective one), even if no individual sub-concept was better classified using only audio. This could result from the nature of the dataset itself, which contains Hollywood movies that, by design, have distinctive sounds and music cues for action scenes that could help the network identify these scenes as violent without learning the specific aspects of violence well.

When adopting both visual and audio features, though, we had our best results yet for most sub-concepts and the final fusion, with 78.5% accuracy. This indicates the complementary nature of the visual and audio features since joining them improves the classifier accuracy compared to each separately.

Our decision to use balanced accuracy to compare results was to have a better sense of the generalization of the methods. However, the official metric used by the MediaEval competition is the mean average precision(mAP). Table 4.10 compares our best results for the fusion of visual and audio features with other works adopting the same dataset in the prior art. We also report results using this method in the latest 2015 MediaEval VSD dataset.

We can see the fusion method compared to the competition's best results and some other works that used the same dataset. In the 2013 version, our method was not the best;

	MediaEval 2013	**MediaEval 2015**
TRoF [38]	0.508	-
LIG [17]	**0.690**	-
FUDAN [11, 12]	0.587	0.296
MIC-TJU [72]	-	0.285
Li et al. [34]	-	**0.303**
Proposed Solution (w. Fusion)	0.656	0.301

Table 4.10: Mean Average Precision (mAP) of our method, in the Mediaeval VSD task in 2013 and 2015, compared with the top performers in both years and later works

it performed better than most competitors, behind only the competition's first place. For the 2015 version, our method was on par with the best results, even from the later work of Li et al. [34]. This is a significant result, mainly because we did not train with the other movies from the training set of the 2015 dataset.

4.2.3 Specialized Dataset

We also tested the pre-trained networks in a more specialized dataset: The NTU-CCTV fights dataset [49]. In Table 4.11, we can see our results for the fights specific network and the fusion in this set. To test this dataset, we extracted every individual frame and used the frame-level annotation provided. Like all of our experiments, the accuracy is balanced to account for the different positive and negative sets. To calculate the mAP, we segmented the shots in the edge frames when a scene becomes violent or becomes non-violent.

	mAP
Two-Stream	0.795
C3D	0.645
TRoF	0.692
Fights Detector	0.623
Proposed Solution (w. Fusion)	0.652

Table 4.11: Mean Average Precision in the NTU-CCTV dataset, with the different methods portrayed in their work and our detectors, specific for fights and the fusion of all concepts.

	Original Training		**Specialized Training**	
	Accuracy	mAP	Accuracy	mAP
Fights	77.2	-	**78.8**	-
Proposed Solution (w. Fusion)	78.5	0.656	**79.6**	**0.661**

Table 4.12: Classification Accuracy (in percentage) and Mean Average Precision of the fights detector and fusion network in the Mediaeval 2013 VSD dataset when the fights detector is trained with the NTU-CCTV-Fights dataset.

One of the main aspects of our method is its modularity, such that it can be easily updated. Here we illustrate one example with the NTU-CCTV dataset, which focuses mainly on fights. We trained a fight-specific network in the NTU-CCTV dataset and plugged in the extracted features from the other concepts, previously tested with the Mediaeval 2013 VSD dataset. Table 4.12 shows our results with this setup when tested in the VSD dataset. This change improved the detector's accuracy, going from 77.2% to 78.8%. This training also improved the accuracy (78.5% to 79.6%) and mAP of the fusion (0.656 to 0.661), making this our best result thus far. This improvement is thanks to a more comprehensive set of examples of fights in the NTU-CCTV than that present in the MediaEval dataset. Given this result, we searched for other available datasets with different concepts, but unfortunately, such datasets do not exist to date.

Chapter 5

Conclusion and Future Work

Detecting violence is a challenging problem, and there is not much work invested in detecting violence in general. Prior work, notably, focuses on a single aspect of violence. We designed a method that breaks the broader subjective concept into smaller and more concrete ones to combine them later and better understand how violence can be portrayed.

We trained different dCNNs (static and motion-based), each responsible for detecting a single aspect of violence. We focused our efforts on the Mediaeval 2013 VSD task and dataset, which has annotations for several violence concepts. We later combined such individual solutions with another network to classify violence using features from all the networks.

We could follow our results as they became better when introducing audio features and trained with specialized datasets. We tested our trained networks in a different, more challenging dataset composed of drastically different videos and had better results than the contestants.

This methodology's modularity allows for the inclusion and removal of any concept of violence relevant to the problem, which we consider a potential solution to its inherent subjectivity. The obtained results with the specialized fights dataset point us in a promising direction, in which we can train each different aspect of violence with its specialized dataset to fuse them into a yet more generalized violence detection network.

Each concept of violence could be trained not only with its own dataset, but taking into consideration the best features and architectures for them.

This idea of training specialized networks with other datasets, however, creates yet another problem in the lack of annotated datasets with different aspects of violence. Most of the available datasets focus on fights. To get a wider range of violence sub-concepts, one approach could be gathering non-labeled data. The challenge then becomes training these networks with non-labeled data.

The big advancement of this idea is taking yet another step to an entirely data-driven approach to the violence detection problem. In this direction, we believe the most exciting future research direction is developing self-supervised learning techniques capable of using just a few annotated examples of each concept and expanding them as more data arrives at the pool. Self-supervised learning using techniques such as triplet loss demonstrated by Schroff et al. with Facenet [53] and Zhai et al. with S4L [76] is just beginning to show its potential, but we anticipate it would be appropriate for such investigations.

Bibliography

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

[2] Scott Blunsden and Robert Fisher. The behave video dataset: ground truthed video for multi-person. *Annals of the British Machine Vision Association*, 4:1–11, 04 2009.

[3] 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 *ACM international Conference on Multimedia*, pages 855–858. ACM, 2013.

[4] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. Automatic reliability estimation for speech audio surveillance recordings. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6. WIFS, 2019.

[5] 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.

[6] 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.

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

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

[9] Christine Clarin, M. Dionisio, Michael Echavez, and Prospero Naval. Dove: Detection of movie violence using motion intensity analysis on skin and blood. Technical report, University of the Philippines, 2005.

[10] Mihai Gabriel Constantin, Liviu Daniel Stefan, Bogdan Ionescu, Claire-Helene Demarty, Mats Sjoberg, Markus Schedl, and Guillaume Gravier. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*, pages 1–1, 2020.

[11] 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 *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[12] 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.

[13] Ankur Datta, Mubarak Shah, and Niels Da Vitoria Lobo. Person-on-person violence detection in video data. In *Object recognition supported by user interaction for service robots*, volume 1, pages 433–438 vol.1, 2002.

[14] 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 - CIARP Pucón, Chile, November 15-18, 2011. Proceedings*, pages 248–255, 2011.

[15] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Bogdan Ionescu, Vu Lam Quang, and Yu-Gang Jiang. Benchmarking Violent Scenes Detection in movies. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2013.

[16] 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.

[17] Nadia Derbas, Bahjat Safadi, and Georges Quénot. LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[18] 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*, pages 517–531. Springer, 2016.

[19] Anitha Edison and Jiji Charangatt Victor. Optical acceleration for motion description in videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1642–1650, 2017.

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

[21] Alex Hanson, Koutilya Pnvr, Sanjukta Krishnagopal, and Larry Davis. Bidirectional convolutional lstm for the detection of violence in videos. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *IEEE International Conference on Multimedia and Expo*, volume 1, pages 113–116. IEEE, 2002.

[27] 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.

[28] 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.

[29] 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.

[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] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.

[32] 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.

[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] Xirong Li, Yujia Huo, Jieping Xu, and Qin Jin. Detecting violence in video using subclasses. *CoRR*, abs/1604.08088, 2016.

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

[36] Zihan Meng, Jiabin Yuan, and Zhen Li. Trajectory-pooled deep convolutional networks for violence detection in videos. In Ming Liu, Haoyao Chen, and Markus Vincze, editors, *Computer Vision Systems*, pages 437–447. Springer International Publishing, 2017.

[37] 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.

[38] 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.

[39] 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.

[40] David E Morrison and Andrea Millwood. The meaning and definition of violence. *International Journal of Media Cultural Politics*, 3(3):289–305, Nov 2007.

[41] Aqib Mumtaz, Allah Bux Sargano, and Zulfiqar Habib. Violence detection in surveillance videos with deep network using transfer learning. In *European Conference on Electrical Engineering and Computer Science (EECS)*, pages 558–563, 2018.

[42] 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.

[43] 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 - Volume Part II*, pages 332–339. Springer-Verlag, 2011.

[44] Bruno Peixoto, Bahram Lavi, Paolo Bestagini, Zanoni Dias, and Anderson Rocha. Multimodal violence detection in videos. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2957–2961, 2020.

[45] 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)*, pages 8276–8280. IEEE, 2019.

[46] Bruno M. Peixoto, Bahram Lavi, Zanoni Dias, and Anderson Rocha. Harnessing high-level concepts, visual, and auditory features for violence detection in videos. *Journal of Visual Communication and Image Representation*, 78:103174, 2021.

[47] 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*, page 50. ACM, 2018.

[48] 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.

[49] Mauricio Perez, Alex Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666, 2019.

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

[51] Giuliano Pontara. The concept of violence. *Journal of Peace Research*, 15(1):19–32, 1978.

[52] 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.

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[54] 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.

[55] Tobias Senst, Volker Eiselein, Alexander Kuhn, and Thomas Sikora. Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation. *IEEE Transactions on Information Forensics and Security*, 12(12):2945–2956, 2017.

[56] 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.

[57] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computing Research Repository (CoRR)*, abs/1409.1556, 2015.

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

[59] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, June 2018.

[60] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[61] 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.

[62] Eleni Tsironi, Pablo Barros, Cornelius Weber, and Stefan Wermter. An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. *Neurocomputing*, 268:76–86, 2017.

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

[64] World Health Organization. Violence and Injury Prevention. Violence: a public health priority. *WHO Global Consultation on Violence and Health, Geneva, 2-3*, Dec 1996.

[65] Marin Vlastelica, 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.

[66] 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.

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

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

[69] Jianxin Wu. Power mean svm for large scale visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2344–2351, June 2012.

[70] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[71] 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*, pages 157–165. Springer, 2018.

[72] 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.

[73] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2012.

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

[75] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7094–7103, 2019.

[76] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.

[77] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019.