ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

Detecting Violent Excerpts in Movies using Audio and Video Features

Luis Jorge Gregório Dias

A dissertation submitted in partial fulfillment of the requirements for the degree of
Master in Computer Engineering

Supervisor:

PhD Tomás Gomes Silva Serpa Brandão, Assistant Professor,

ISCTE-IUL – Instituto Universitário de Lisboa

Co-Supervisor:

PhD Fernando Manuel Marques Batista, Assistant Professor,

ISCTE-IUL – Instituto Universitário de Lisboa

October, 2016

Resumo

Esta tese aborda o problema da deteção de violência em excertos de filmes, com base em características extraídas do audio e do video. A resolução deste problema é relevante para um vasto leque de aplicações, incluindo evitar ou monitorizar a exposição de crianças à violência que existe nos vários tipos de média, o que pode evitar que estas desenvolvam comportamentos violentos. Analisámos e extraímos características áudio e vídeo diretamente do excerto de filme e usámo-las para classificar excertos de filme como violentos ou não violentos. De forma a encontrar o melhor conjunto de caracteristicas e atingir a melhor performance, as nossas experiências utilizam dois classificadores, nomeadamente: Support Vector Machines (SVM) e Redes Neuronais(NN). Foi usado um conjunto balanceado de excertos de filmes, retirado da base de dados ACCEDE, conjunto esse, que contém 880 excertos de filme, anotados manualmente como violentos ou não violentos. Durante as primeiras experiências, usando características incluídas na base de dados ACCEDE, testámos caracteristicas áudio e características vídeo, individualmente, e combinações de características áudio e vídeo. Estes resultados estabeleceram o ponto de partida para as experiências que os seguiram, usando outras características áudio, extraídas através de ferramentas disponíveis, e outras características vídeo, extraídas através dos nossos próprios métodos. As conclusões mais relevantes a que chegámos são as seguintes: 1) características áudio podem ser facilmente extraídas usando ferramentas já existentes e têm grande impacto na performance do sistema; 2) em termos de características vídeo, caracteristicas relacionadas com o movimentos e transições entre planos numa cena, parecem ter mais impacto do que características relacionadas com cor e luminância; 3) Os melhores resultados ocorrem quando se combinam características áudio e vídeo, sendo que, em geral, o classificador SVM parece ser mais adequado para o problema, apesar da performance dos dois classificadores ser semelhante para o melhor conjunto de características a que chegámos.

Abstract

This thesis addresses the problem of automatically detecting violence in movie excerpts, based on audio and video features. A solution to this problem is relevant for a number of applications, including preventing children from being exposed to violence in the existing media, which may avoid the development of violent behavior. We analyzed and extracted audio and video features directly from the movie excerpt and used them to classify the movie excerpt as violent or non-violent. In order to find the best feature set and to achieve the best performance, our experiments use two different machine learning classifiers: Support Vector Machines (SVM) and Neural Networks (NN). We used a balanced subset of the existing ACCEDE database of movie excerpts containing 880 movie excerpts manually tagged as violent or non-violent. During an early experimental stage, using the features originally included in the ACCEDE database, we tested the use of audio features alone, video features alone and combinations of audio and video features. These results provided our baseline for further experiments using alternate audio features, extracted using available toolkits, and alternate video features, extracted using our own methods. Our most relevant conclusions are as follows: 1) audio features can be easily extracted using existing tools and have a strong impact in the system performance; 2) in terms of video features, features related with motion and shot transitions on a scene seem to have a better impact when compared with features related with color or luminance; 3) the best results are achieved by combining audio and video features. In general, the SVM classifier seems to work better for this problem, despite the performance of both classifiers being similar for the best feature sets.

Palavras Chave
Keywords

Palavras chave

Deteção de Violência

Aprendizagem Automática

Características de Áudio e Vídeo

Classificação de Excertos de Filme

Rede Neuronal

Support Vector Machine

Keywords

Violence Detection

Machine Learning

Audio and Video Features

Classification of Movie Excerpts

Neural Network

Support Vector Machine

Agradecimentos
Acknowledgements

Esta tese não teria sido possível sem o contributo de muitas pessoas que, cada um à sua maneira, tornaram tudo isto realidade, e aos quais estarei eternamente grato.

Aos meus orientadores, que sempre estiveram disponíveis para me ajudar, e que me acompanharam durante todo o meu percurso, dando-me todo o apoio que necessitei, e mais algum.

À minha namorada, que durante este meu percurso, sempre esteve ao meu lado e apoiou-me, encorajou-me e sempre me fez acreditar que seria possível, sendo que teve uma paciência e força inesgotável para me aturar.

Aos meus amigos e colegas, quer dentro da faculdade, com quem partilhei experiências fantásticas, quer fora da faculdade, que sempre estiveram lá para mim e, sempre se mostraram disponíveis para me ouvir, ajudar-me nos meus problemas do dia-a-dia ou, simplesmente acompanharem-me nas noites de copos.

A toda minha família, em especial, o meu irmão que sempre foi o meu melhor amigo e com quem sempre pude contar em todo o tipo de momentos, e, acima de tudo, os meus pais, que não só me proporcionaram tudo isto, como também, sempre se sacrificaram por mim e fizeram tudo para que eu pudesse estudar, mesmo em tempos muito difíceis, e que, mais importante ainda, tiveram paciência para aturar o meu mau feitio, e com quem sempre pude contar nos melhores e nos piores momentos. A todos, muito obrigado.

Lisboa, 31 de Outubro de 2016

Luís Dias

Contents

List of Figures

List of Tables

Introduction 1

This chapter introduces the work conducted on this thesis by presenting the motivation, the goals, and the corresponding research questions. Section 1.1 introduces the starting point of this work and addresses the main motivations. Section 1.2 describes the main goals, other objectives, and the expected outcome of interest to the scientific community. Section 1.3 presents the questions addressed on this work and intended to be answers, and the corresponding research addressed in the scope of this thesis. Finally, Section 1.4 provides a description of how the document is organized and structured.

1.1 Motivation

Violence among children and their relationship with existing violence on the many kinds of media (such as movies or video games), that they are exposed to, is a subject studied for quite some time, and continues to be a very current problem. Both in the twentieth century as in the present century, many studies on the subject matter have been conducted. studies which aimed at understanding the relationship between the existing violence in various types of media and violent behavior that people, including children, who had access to these types of media, developed.

In a study that followed the growth of several children to their entry into adulthood (between 1977 and 1992), Huesmann et al. (2003) were able to conclude that children's exposure to violence on television assisted, influenced and increased the possibility that these come to develop behavior and violent attitudes during their adulthood, regardless of the context in which they lived on. This study also found that these conclusions are checked regardless of the sex of the child.

On a different research, Funk et al. (2004) concluded that the existing violence in video games and movies is reflected in an increase in feelings indicative of lack of empathy and attitudes said as demonstrative of violence, in people exposed to violent situations within these two types of media. Despite this conclusions, the study's conclusions did not reveal this very same link between violent behavior and violence to which people watch, for other types of media.

Bushman and Huesmann (2006) conducted a study that compared the effect of violence in various types of media on children and people in adulthood, coming to the conclusion that the violent content that they were exposed to had a greater short-term impact, in terms of changes on behaviors and attitudes, on adults and greater impact in the long term in children. This study also revealed that, those changes on the way of being of the targeted people, were clear and there was an evident appearance of attitudes, thoughts and aggressive feelings, both on the children and on the people that already reached the adulthood.

The aforementioned studies reinforce the need for parents and early childhood educators have some control over the content that their children watch or play, which, today, is increasingly complicated. As parents or guardians spend little time with their children, due to their professional lives and their lives outside the home environment, they tend to have difficulty on reducing their children exposure to acts of violence and other acts that are revealing of aggressive attitudes, for example, the use of inappropriate language. The development of tools and policies that assist children's parents and tutors on this task is thus imperative, as pointed out by Sheehan in the study reported in Sheehan (1997).

1.2 Objectives

It is within this context or motivation, and this need of parents and guardians to monitor and assess whether what their children watch or not violent and harmful to even arises the work described in this thesis. One of the goals of this dissertation is to provide a baseline for the development of an automatic system that helps the parents or tutor on this task of having more control on its children exposure to certain contents, by classifying movie scenes or excerpts as violent or non-violent based on its characteristics. The task address focus on, not only research and discover a set of features that are strongly correlated to violence that will be used by the system for its classification, but also build each phase of the system by describing the classifier and the techniques used for preparation and processing of the features to be used on the system. Thus, the aim of this work is to offer a contribution to help the parents of a child on the process of their decision of letting or not the children see a given movie.

The work presented on this dissertation, not only will, as stated, contribute as baseline for a tool that can be used in a future for helping parents and tutors on the task, already mentioned before, but also can be used and taken into consideration for similar tasks and other related works that can be done in a future. The use of this system or its components will be completely available so it can be considered by other projects if given the credit to the author. It can also be used as a tool for developing a complex system that, not only classify excerpts of movies (which is our contribution,) but also combined that information in order to classify the whole movie, or to give a concrete output in terms of which length of the movie is violent, and the exact temporal points where violence (detected by the system) occur, so a parent, can see that specific excerpt and decide by himself if its children can watch it.

On the task of detecting violence, an extensive research and testing of many different features is performed. those features are originated from both audio and visual domain, with different characteristics, some of them used in other tasks, like paralinguistics, or created, using Matlab development tool. A combination between features will be performed, with the intent of reaching a better system performance. A intent of this work is to understand which of the two domains performed better on the system and, which of them has an easier feature extraction step. the last point is something to be considered if there is a desire to built a real-time system. Also, there is an intent to reach a conclusion on which of the tested classifiers performed better along the experiments. This research and its conclusions, are this work's biggest contribution for the scientific community.

With this work, it is intended to enhance other works on the field of detecting violence, that were taken into consideration for developing the research step.

Having all this goals and considerations in mind, the system will have three phases:

- Separation of the audio and video signals from the movie excerpt in order to process them separately.

- Extraction of features for each signal. At this step different groups of features will be tested and extracted.

- Classification of the movie excerpt as violent or non-violent, based on the features that fed the tested classifier (SVM or Neural Network).

1.3 Research Questions

In order to fulfill the goals and objectives of this work, as well as to build the intended system, there is a focus on have the following research questions answered:

- What are the main components that a system for detecting violence in movie excerpts need to have?

- Which machine learning algorithms should be used and tested on this task?

- Are there any projects on the very same task that can be fairly compared to our work, not only in terms of the process followed, but also in terms of results?

- Are there any projects in similar tasks that show different approaches in terms of features and classifiers used?

- Which audio or video features are usually used on the task of detecting violence in excepts or other similar tasks within the field of violence detection?

- Is there any set of movies (or movie excerpts) that can be used on this task?

- Are there any already extracted features that can be used as baseline for our work?

- Are there any toolkits available for extracting audio or video features?

- Which of the two modalities of features usually have a better performance on similar tasks?

1.4 Document Structure

This dissertation is structured as follows: On Chapter 2, a presentation of a series of projects, researches and other works produced on the field of violence detection is conducted. This group is, not only composed of works with the exact same task of our work, but also with other related tasks. On Chapter 3 it is presented a view of the system architecture, its components and the process that was followed when deciding which of the available options would be better to use on each of the steps of creating the system. Still in this chapter, a description of the database used to feed the system, which made part of the work, during all process of building the system and feature extraction and testing, was performed. On Chapter 4, it is performed an overview of all the sets of features that were extracted, both of audio and video domain, and, not only extracted using programmaticly developed programs, but also using other available software and other already extracted features. On Chapter 5, there is a focus on all of the experiments conducted during this work life-time, while trying to find the best performing features and the best performing classifier for the already explained problem, dividing the experiments in two phases. Finally, Chapter 6 presents the conclusions reached out during this work, which answer the questions pledged to be answered, and the next step of this work in terms of what can be done after and based on this thesis.

2 Related Work

The main goal of the work described on this dissertation is to build a system able to classify a movie excerpt as violent or non-violent, based on an mix of selected audio and video features. In order to fulfill this goal, a research was conducted, as a way of finding different approaches and kind of works that could function as a base, or at least as something where ideas to develop this work could be found, especially on the subject of selecting features to feed the system. On the particular task of classifying violence on movie excerpts, not many bibliography was found, except the works conducted by Lin and Wang (2009), Chen et al. (2011) and Giannakopoulos et al. (2010).

Starting by the work conducted and described in Lin and Wang (2009), it is proposed a mechanism for identifying violent movies using audio and visual features. In the case of audio features, a method monitored weekly is on a set of previously extracted audio features, such as bandwidth high or zero cross rate ratio. Those features are known to have good results in the field of audio classification and, in this work, their representations are grouped together, creating an audio vocabulary. In the case of video features a classifier is used to detect violence on images. In this case, are considered relevant in the context of detection of violent scenes, the rapid movement between images, the detection of an explosion or fire in the extracts and the presence of blood. All these considered relevant information is obtained through image processing techniques somewhat complex:

- In the case of fast movement between frames, a motion vector is calculated for a frame macro-block, having the motion intensity calculated from this very same motion vector. In order to calculate the intensity associated to the frame by itself, a sum of the motion intensities of all the macro-blocks that compose a frame is calculated.

- For the case of fire and explosions, the frame is divided in areas that contain yellow

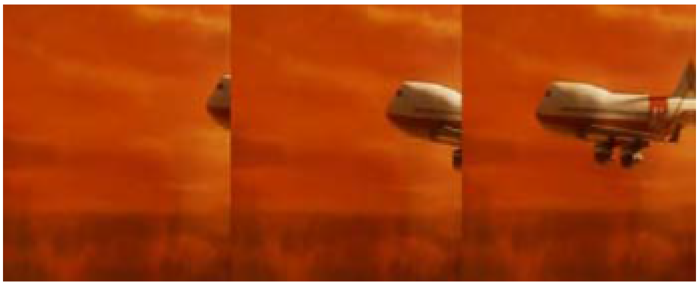


Figure 2.1: Sequence of frames not detected as blood due to lack of quick movement

tones and areas that does not contain yellow tones. Afterwards, the adjacent areas are grouped and it is studied the existence of motion within those areas to see if they can be connected to any of those phenomenon.

- In the case of blood detection, it is used a method similar to the above, but this time the division of the image is performed according to red tones. In addition, it is assumed that the appearance of blood in a violent scene is highly connected to quick movement on the scene. Figure 2.1 shows a situation where the frames that compose the excerpt are, by themselves, composed of a majority of red tones, but it is not detected blood due to lack of quick movement in the scene.

Obtained results of the two fields of information, these are combined using a machine learning method called co-training, which combines two sets of data, which typically do not have dependencies on each other. The results obtained following this approach were compared with a system using Support Vector Machines and was denoted increased performance, from one system to the other, of about 20% for the test that was made.

The work conducted and described in Chen et al. (2011) proposes an approach that not only extracts features on a frame level, but also uses the movie on its all length and extracts features related to a set of continuous frames. This work used only visual based features, as a way of being more effective. Those visual features are related to the existence of motion on a movie excerpt and the existence of blood within a frame. Their work process can be summarized in four major steps:

1. As the first step on their approach, they start by segmenting a whole movie in different shots.

2. After that, the shots are grouped in a scene

3. Thirdly, they use an SVM based classifier to detect action scenes.

4. Last, but not least, they proceed on detecting blood on the scenes.

Although their work flow is as it's described above, the referred paper has more focus on detecting action scenes and blood frames. On the first case, four different features are extracted and, intend to describe the motion within a scene:

- Average motion intensity, which is calculated by computing a motion vector for every 16x16 pixels of a frame. After that, it is computed an average of all motion vectors, which is the Average motion intensity.

- Camera motion ratio. In this case, it is verified if a frame has less than 10% of null motion vectors. If they have, that frame is considered to have camera motion. Therefore, the camera motion ratio is calculated by dividing the number of frames with motion in a scene by the number of frames that compose a scene.

- Average shot length, which is calculated by, for a scene, by averaging the shot length of each of the shots that compose a scene.

- Shot cut frequency. This feature is obtained dividing one by the number of shots in a scene.

Those features are mixed in a SVM classifier, which will return, as output, a binary decision of classifying the scene as an action scene or as a non-action scene.

For detecting the blood frames on a scene, only the frames contained on a previously classified action scene are considered. On those scenes, they select a keyframe that will represent each shot that is contained on the action scene. Then, they used the keyframe for, not only trying to detect blood on the frame, but also, trying to identify human presence on that frame, by detecting faces. While on the face detection case, they used an existing face detection algorithm, on the blood detection case, they check, on the keyframe, for aggregations of pixels with their color present on a limit of RGB indexes.

Finally, they tested their system on four known Hollywood movies, and compared their approach with the one developed by Lin and Wang (2009), previously mentioned, obtaining a better performance, especially in terms of recall.

Last, the work conducted by Giannakopoulos et al., described on Giannakopoulos et al. (2010), was performed, in order to address the referred task, proposing an audio and visual features fusion classified using Bayesian networks and k-nearest neighbor algorithms. The group of researchers created different classifiers for audio and video, to recognize distinctive characteristics in each. In the case of audio, they used a classifier, based on Bayesian Network and an One vs All classification architecture, with seven classes, four of which are considered non-violent and claimed to represent audio relating to music, normal characters speech or other audio types pertain to the normal environment in the movies. In the case of violent classes, these were meant to represent situations of shooting, fighting or shouting. For each movie excerpt, it was calculated its probability of being within each of the referred classes. In the case of image or video they were created three classes representing the interaction of existing people in the scene, based on the motion within the excerpt. The first class represented situations where there weren't people on the scene or they appeared to be immovable. The second class was used to represent situations where there was interaction of people in a natural way scene. Finally, the third class represented situations where the sudden movements, by the characters, occurred, such as falls or fights. The combination of the various classes for each film, and the consequent decision of the system was achieved using an algorithm called k-Nearest Neighbor (kNN). that returned a binary response, indicating that the film was violent or not. The system was evaluated based on a comparison of its results and the opinion of three people, for fifty movie excerpts. Figure 2.2 shows a representation of the system used by this group of researchers.

All this three works intended to address the same task as this work. Although, it isn't possible to compare the results between each work and this one, since, not only, the datasets used to build the system aren't the same, but also, the dataset of the work described on this original thesis is balanced, situation that doesn't occur on Lin and Wang (2009), Chen et al. (2011) and Giannakopoulos et al. (2010).

These three works were the only found with the exact same task as the work presented on this dissertation. However, much more studies were conducted on similar areas of violence detection. On this other works, not only the task by itself changes, but also the methodologies and features used, considering they're still within the violence in movies field.

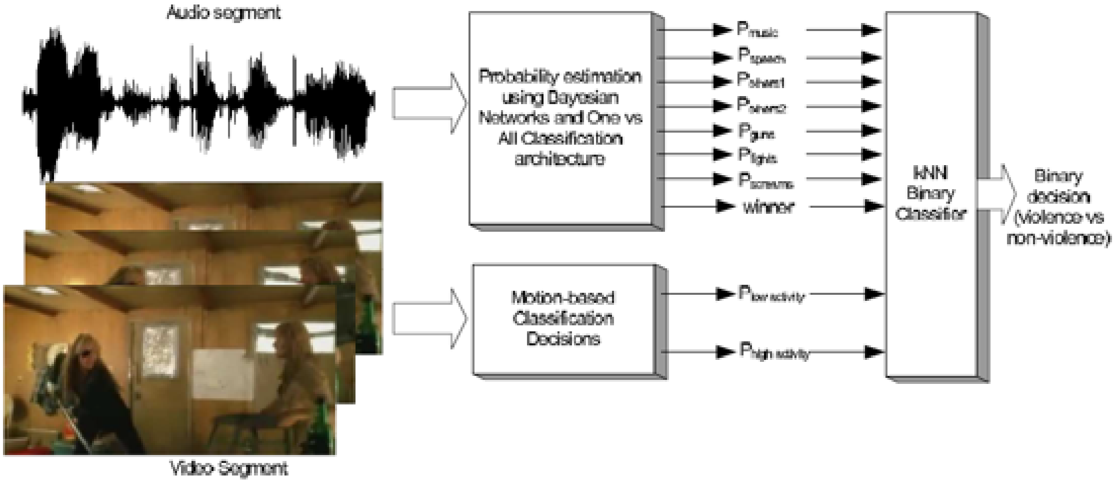

Audio segment
Probability estimation using Bayesian Networks and One vs All Classification architecture
P music
P speech
P shots1
P shots2
P guns
P fights
P screams
winner
kNN Binary Classifier
Binary decision (violence vs non-violence)
Motion-based Classification Decisions
P low activity
P high activity
Video Segment

Figure 2.2: Overview of the system reported by Giannakopoulos et al. (2010)

Acar et al. (2013) work consisted on developing two systems that used a SVM classifier for fulfilling the task of detecting the most violent scenes in Hollywood movies. Each system was fed with different groups of features: the first one contained only audio features, such as Mel-Frequency Cepstral Coefficients (MFCCs), building a bag of audio words; the second one used a mix audio and visual approach, using the features of the first system as audio features, and features from the video domain, which were motion related features, that consisted on computing the average motion between different key frames, as visual features. The results showed that the second system developed had a better performance on the task addressed, when comparing to the first one.

The MediaEval conference is a conference that since 2011 has addressed the issue of evaluating affective content in movies, including violence. One of the tasks or challenges proposed by this conference is to find the most violent excerpts among a dataset, somehow similar to the task addressed in this work. Although it is, still important to reinforce that the task that is approached on this MediaEval conference, it is different than the one approached on this work. For instance, instead on trying to evaluate which excerpts, out of a previously selected set of excerpts, are the most violent, which is the task addressed in MediaEval conference, the task addressed on this is to determine and classify if a movie excerpt is violent or not. This conference originated many different works, results and approaches, not only on the subject of the classifiers used, but also on the subject of selection of features, worth considering for the referred task, and, although the task isn't the same as the one addressed on this thesis, as the domain of the problem is the same, this work's and approaches can give a really important value to the project developed on this thesis.

In terms of the many different works conducted on this conference and task, Jiang et al. (2012) used features from both audio and video domain. For the video case, they extracted features related to the trajectory and motion in the movies, spacial and temporal points of interest, to check situations where pixels presented big variations on both spacial and temporal domains, converting this feature into a bag of words with 4000 code words and SIFT descriptors, used to locate local invariant pieces of images in a movie frame. For the audio case, the usual MFCC coefficients as a way to represent the audio signal that composes the movie. These two groups of features are subject of a process of soothing, where each frame is represented by an average frame between itself, and the frames directly before and after it, and merged in a SVM classifier with a polynomial function based kernel to fulfill the task.

Penet et al. (2012) created five different systems based on different techniques and algorithms. First, they try to avoid using machine learning algorithms and try to fulfill the task using, only, a similarity measure between the data that is contained on the training set and the data that is contained on the testing set. So for each movie, the system classify them by only compare it to violent movies and non-violent movies within the training set, using a K-Nearest Neighbor algorithm to give the final decision. The second system, used a bag of audio words as features, and a sigmoid function to give the final mean average precision, which is used to measure the system performance. The third system, added a set of 6 color related features to the audio features used, and Bayesian Networks machine learning algorithm to achieve the system decision. The fourth system, the group of researchers reimplemented a previous built system which used audio, video and text related features, and they removed the text related features, due to lack of performance. Last, they mixed all the approaches but the first one in a single system to address their problem.

Also, Dai et al. (2014) used both audio and visual domain features to address the task. They used trajectory based features, spacial and temporal points of interest and MFCC coefficient as features, in an approach that is similar to the one taken by Jiang et al. (2012) on their previously mentioned work. As classifiers, they used both SVM and deep neural networks and fused their results which were target of a score smoothing. an overview of the system can be seen on 2.3.

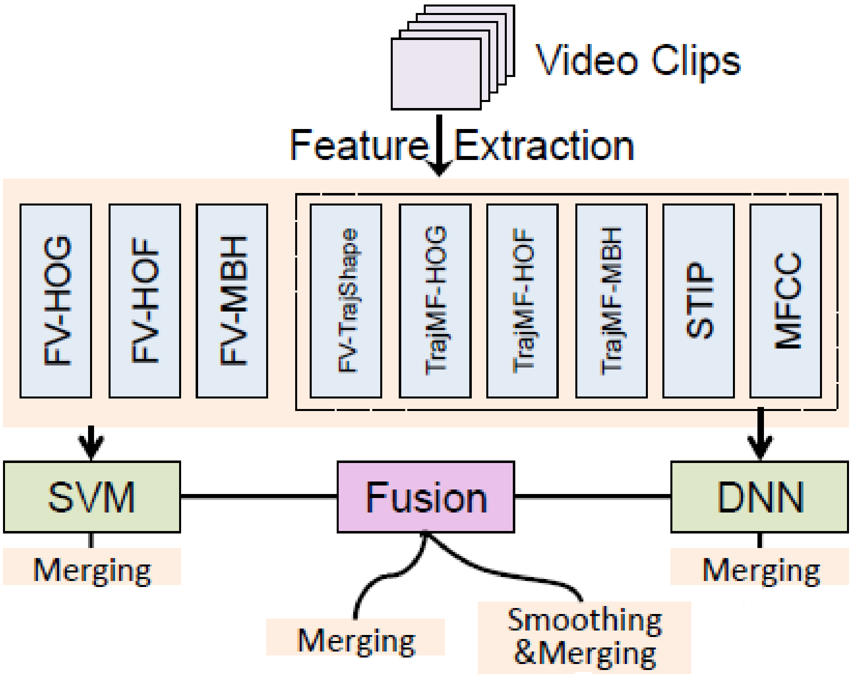


Figure 2.3: Overview of Dai et al. (2014) system

Despite many of the presented approaches use features from both audio and video domain, Lam et al. (2012) used features from, only, the video domain. This group of researchers used many global features related to color, with different configurations on the subjects of granularity, color space chosen and quantization: color moments, color histogram, edge orientation histogram, and local binary patterns, with the intent of describing the color and texture distribution of the movie frames. This features were extracted from 5 keyframes that were used to represent a shot of movie. As for the classifier used to mix this set of features, the SVM classifier was the one chosen, having a Radial Basis Function kernel. The training was performed using cross-validation with five folds.

Martin et al. (2012) used only video/visual related features on their work. They used a simple module that uses multi-scale local binary pattern histogram features mixed in a Support Vector Machine classifier with a linear kernel.

Derbas et al. (2012) developed four different runs in order to address the MediaEval conference task. On their task they used descriptors from both audio and video domain:

color histograms to describe the color distribution on each movie frames; texture descriptors; SIFT descriptors and MFCC coefficients. To these descriptors, an optimization process was applied. First, a normalization process with this descriptors as targets occurred, and then a PCA reduction with the intent of reduce the number of dimensions of the descriptors was conducted. For their work, two different classifiers were used - SVM classifier and one based on the k-nearest neighbor algorithm - and their results were mixed. Other similar works conducted on this conference with the goal of addressing the referred task are describe on the work conducted by Acar and Albayrak (2012) and Eyben et al. (2012).

In this section it is presented a set of different works produced on the area of violence detection. The works on this domain, that had the exact same task as the work conducted on this dissertation aren't much, in terms of quantity, as this is a fairly new and specific topic of the field. Nevertheless, with these related works, a series of different approaches were found, in terms of system architecture, features used and system evaluation. Although the task is the exact same on as the one addressed on this thesis, the results obtained from these works are not fairly comparable to the ones obtained during this dissertation, since they use different datasets. On a different subject, a set of different works were obtained with tasks similar to the one addressed on this work. These projects and researches resulted on many different approaches that were taken in consideration for our own work. Many different groups of features were extracted (and from different domains) and combined in many different forms using different classifiers and machine learning algorithms and techniques. This provided a starting background for the research conducted during this dissertation and highly inspired work conducted.

The Classification System 3

As stated before, the goal of this original master thesis is to develop a system that is able to correctly classify a movie excerpt as violent or non-violent, by analyzing a set of researched features, that can be from both audio or video domain, with the maximum performance possible. Having that in mind, the first step followed, before advancing to the concrete implementation, training and testing phase is to think about what are the main components of the system. This focus on addressing the system architecture and what techniques or options are going to be taken during each major component implementation, as well as what set or database of movies are going to be used as target of the experiments conducted on the scope of this work. In this last scenario, a research was conducted in order to look for an already built and available set of movie/movie excerpt data, that had the best characteristics, in order to be used on this work. Therefore, this Chapter is divided in two sections, each of them with their own self organization: On Section 3.1, it is addressed the system architecture, each of the main components of the developed system to address the problem that this original thesis is trying to solve, and the choices that were taken in each decision point of the development, in terms of system architecture. On Section 3.2, it is described the movie database that was found and adopted on this work, evaluating why that database was helpful, why it was chosen and what are its major characteristics.

3.1 Architecture

The system used to address the problem and fulfill the goal of detecting violent excerpts can be divided on three major components, with very specific functions related to them:

- Signals processing and separation component;

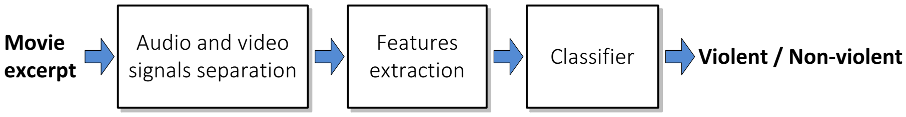


Figure 3.1: Representation of the system used and its major components

- Features extraction component;
- Classification component.

The intent of building this system is to be able to, by receiving a movie excerpt, already extracted from a whole movie (something that is not the target of this work), have the ability to classify it as a violent or non-violent excerpts, as a way of fulfilling the task addressed on this dissertation. To do that, the movie excerpt needs to go through a previous processing phase, where not only, the signal of the movie excerpt is separated on audio signal and video signal, but also, these signals are outputted in a raw format that allows them to be treated. This signal processing and separation is described on Section 3.1.1. After that, and having the audio and video signals of the movie excerpt in their raw format, it is possible to extract features from both domains from them, in order to get a set of "indicators" that will impact the final decision of the system. The process of extracting those features, the techniques used and other toolkits or software used, are described on Section 3.1.2. Having the set of wanted features or movie excerpt characteristics extracted, the system needs to evaluate them and classify the movie excerpt according to the violence within it. This is done by using a classifier to reach the system decision. The details on what classifiers were tested and how the training and system performance analysis was conducted, is addressed on Section 3.1.3. A more visual representation of the system architecture is provided on Figure 3.1.

3.1.1 Signal separation and processing

The component of signal separation and processing had a very specific and simple task: To not only separate the audio and video signal from the video excerpt, but also to make sure that the outputted format of both signal was a raw one. The reason this processing step was done, in the first place, was to make possible the extraction of independent audio and video features directly from the signal, using their raw format.

Format	Global Size (MB)	Average Size per excerpt (MB)
mp4	999.2	1.14
wav	1586.8	1.80
yuv	~74800	~85

Table 3.1: Comparison between the original files and the raw files

To complete the task explained, a tool known as FFMPEG was used. This is an open source tool to process multimedia data that allows the extraction of audio and video signals from the whole movie excerpt, with possibility of choosing the type of files the user wants to output the signals on, using the command prompt. This tool also allows the user to choose many other configurations for the output system, for example, the number of channels chosen for the audio signal. This processing was applied to the dataset used on this work, that was composed by 880 movie excerpts.

To separate the audio signal, the FFMPEG tool was used to output the .wav (raw format) from the original .mp4 files. The reason behind why that format was chosen was the fact that, not only it is a raw format, but also, it is compatible with OpenSmile toolkit, that was used to extract audio features, as it will be described further. To separated the video signal, the FFMPEG tool had the files outputted on .yuv files as it is a format that is very commonly used on video processing. It is worth mentioning that, for the video case, another phase of processing was conducted, to separated the .yuv files that contain the information of the visual signal of the whole movie excerpt in frame related informations, in order to be possible to extract features from the frame level. The table 3.1 shows a size comparison between the original .mp4 files and the raw formats from both signals: .wav files for the audio case and .yuv files for the video case.

As it can seen by analyzing the table 3.1,both raw formats had bigger size than the original format, as excepted. As the .mp4 is an encoded format and the raw files do aren't target of any type of encoding process, the raw files tend to be bigger. It is also quite visible that the .yuv files are much bigger than the .wav files, which indicated that they will take longer to process.

After this step is concluded, it is possible to extract features from the both separated signals.

3.1.2 Features extraction

After having the raw formats for the target movie excerpts, it is, now, possible to extract features, of both audio and video domain, from them. The description of each individual set of features extracted can be seen on Chapter 4.

The two big groups of features: audio features and video features, where extracted from very different ways, as they belong to different domains.

The audio related features were extracted using the OpenSmile (Eyben et al. (2010)), a toolkit that allows a transparent extraction of audio features based on different configurations of the extraction. This makes possible the extraction of many groups of features with different complexity levels and different number of features. This toolkit was, then, used to extract three different groups of features as described on 4. As it works, the OpenSmile is given a configuration and the .wav file (raw audio format) and outputs a file describing each feature value for the audio file, in arff format.

For the task of extracting video features, it wasn't found any toolkit, similar to the OpenSmile toolkit, but for video features processing. Therefore, the features had to be extracted recurring to scripts developed by us, using Matlab software. The software is commonly used for video processing and therefore was chosen to be the developing environment of this dissertation. For this particular domain, two subtypes of features were extracted, within the video features context:

- Motion related features;
- Color and luminance related features.

After the features were extracted they were outputted in .csv files, used to store the information of each individual feature, for each movie excerpt that was target of this work. More information on the features extracted is described on 4.

This component received the two types of signals as input, and outputted the .arff files containing the audio features information and the .csv files with the information about the video features.

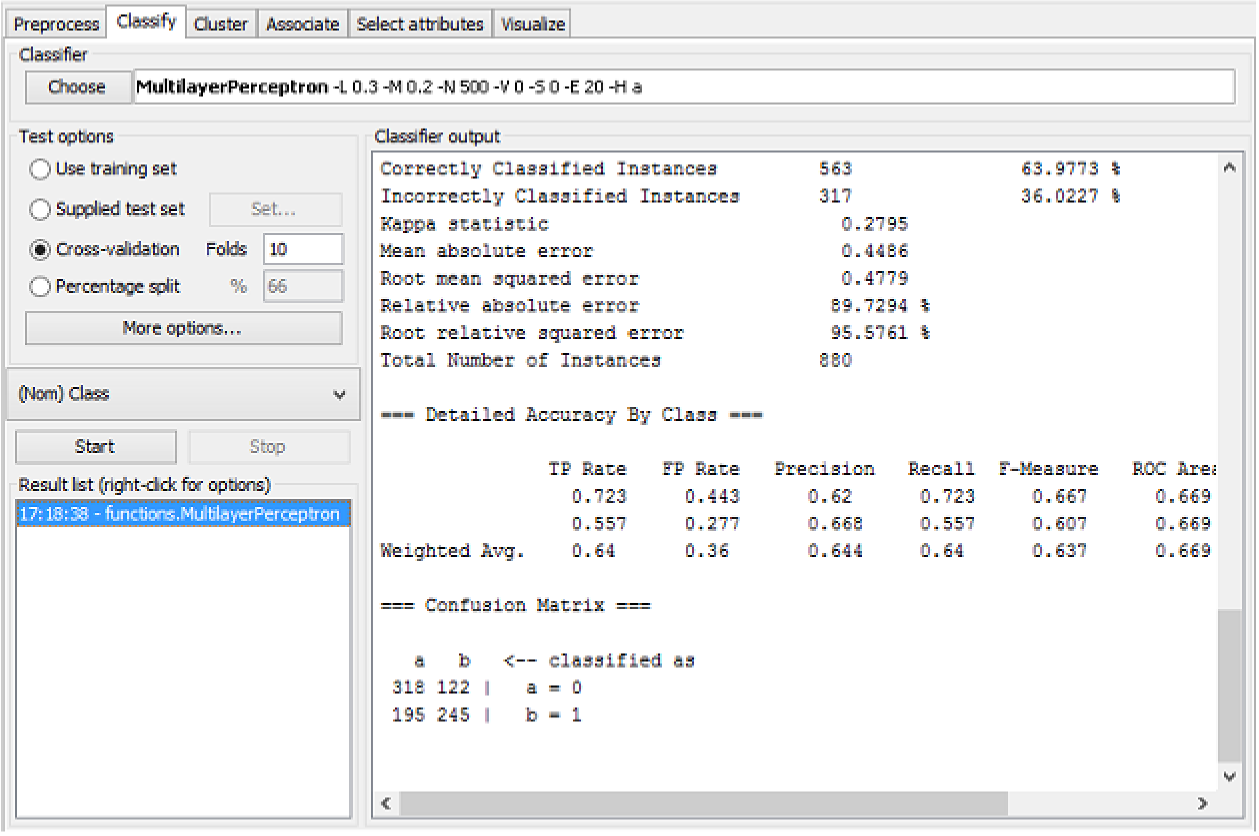
3.1.3 Classification

After having the complete sets of features extracted, from both audio and video domain, using OpenSmile and Matlab, respectively, the next step was to use their information and combine it using a classifier, in order to be able to obtain the system response. On this particular task, the classifier is used to process the input features and output a binary decision in terms of classifying the movie excerpt as violent or non-violent. Thus, two classifiers were chosen: Support Vector Machines(SVM) and Neural Network. The SVM classifier was chosen due to its good results in similar tasks as shown on many works already described on Chapter 2, while the Neural Network was chosen, as this machine learning algorithm was used, with good performance, in many different fields of knowledge, for example in DeGroff et al. (2001), Figueiredo et al. (2014) or Prasad et al. (2013), and, therefore, was tested on this task, as well, with the hope of also have a good impact.

For this phase of the system a cross-validation technique with ten folds was used, train and evaluate the system according to its performance, based on the number of cases the system evaluated correctly divided by the number of classifications the system made. To implement both systems, Weka software was used. Weka is a toolbox to implement machine learning algorithms to a given problem. It allows a user to choose the classifier he wants to use on its problem, change its configurations, like the kernel used for the SVM classifier or the neurons architecture for the Neural Network classifier, choose the way the system will be evaluated, by choosing cross-validation or by dividing the set of features he’s working with in training and testing sets, and to apply operations, such as data set standardization or normalization, on the set of features in order to try to enhance the system performance. Figure 3.2 shows a sample of the interface of Weka for implementation of machine learning algorithms.

By analyzing the Figure 3.2 , it can seen that, not only this tool contains the features described before, but also outputs a series of statistical information about the training and the system, such as the precision, recall and f-measure, and the number of correctly and incorrectly classified instances, used for the system performance checking.

Both classifiers were tested on all the sets of features that were extracted and the system performance using each classifier on each set of features, individually or mixed



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

17:18:38 - functions.MultilayerPerceptron

```
Correctly Classified Instances         563               63.9773 %
Incorrectly Classified Instances       317               36.0227 %
Kappa statistic                          0.2795
Mean absolute error                      0.4486
Root mean squared error                  0.4779
Relative absolute error                 89.7294 %
Root relative squared error             95.5761 %
Total Number of Instances              880

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Are
                 0.723     0.443     0.62       0.723     0.667      0.669
                 0.557     0.277     0.668      0.557     0.607      0.669
Weighted Avg.    0.64      0.36      0.644      0.64      0.637      0.669

=== Confusion Matrix ===

   a   b   <-- classified as
 318 122 |   a = 0
 195 245 |   b = 1
```

Figure 3.2: Sample of the Weka Interface

with other sets of features can be seen on Chapter 5.

3.2 Database

Having the system architecture selected on each of its own components level, as described on the previous Section, a set of movie excerpts data to be processed, have its features extracted, and classified as violent or non-violent, was needed. On the research for this type of data set, the main goal was to find a dataset that already had its movies separated in small excerpts, which was intended to be the input of the system described on this dissertation. The data set chosen was provided by ACCEDE movies database (Baveye et al. (2015)).

The ACCEDE movies database is a set that contains 9800 movie scenes/excerpts on .mp4 format, with about 10 seconds of duration and approximately 1 MB of file size. This movie excerpts are, individually, annotated as violent or non-violent, in a binary classifica-

Average length per excerpt (s)	9.901
Total length (s)	97033.6
Average size per excerpt (MB)	1.0492
Total size (MB)	10281.9
# of Violent excerpts	440
# of Non-Violent excerpts	9360

Table 3.2: Details about ACCEDE database content

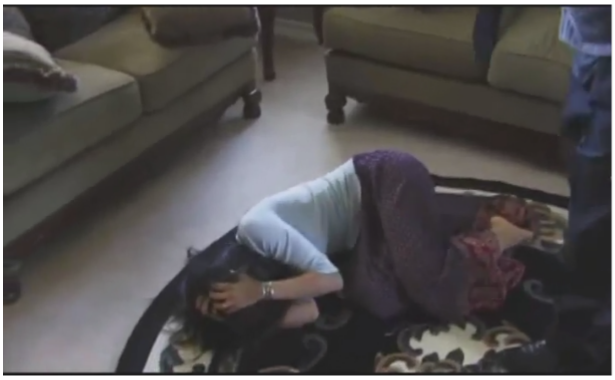


Figure 3.3: Sample of a scene considered violent

tion based on subjective opinions provided on a crowdsourcing experiment by a big number of participants.

From the 9800 movie excerpts existent on the database, most of them were considered non-violent by the crowdsourcing method: from the whole set, only 440 of the 9800 movie excerpts were actually classified as violent by the participants, having the remaining 9360 and majority of the set classified as non-violent. Some more details about the content of the ACCEDE database can be seen on Table 3.2.

A sample of an image that belong on a movie excerpt considered violent can be seen on Figure 3.3, while, an image of a scene that is not considered violent can be seen on Figure 3.4. If this two excerpts are studied, it can be seen that in the first one a man is a attacking a women, and the scene is in general composed b darker colors, lots of movement and specific sound of screaming, while the other one as a more calm atmosphere where three friends are talking to each other. This kind of observations were taken into consideration while searching for indicatives and features that represent violence.

The fact that the dataset has such a majority of data referred to the non-violent content is a big issue, since it shows that it's quite unbalanced and that threatens the training of

Figure 3.4: Sample of a scene classified as non-violent

the classifiers, especially the Neural Network based one, as there is a big possibility of the classifier only classify all the movie excerpts as non-violent, as it works for most of the cases, which is unwanted. This way of working may provide a big accuracy of the system, however, it's completely undesirable to solve the problem addressed on this dissertation. Therefore, the first step to use this database correctly was to form a balanced dataset. Since there were only 440 movie excerpts classified as violent, they were all used on this new balanced dataset, and, additionally, 440 movie excerpts classified as non-violent were randomly selected from the 9360 non-violent movie excerpts available. Doing so, after all this processment, a balanced and smaller dataset, was reached, with a proportion of 1/1 of movie excerpts classified as violent and movie excerpts classified as non-violent, that was about 1/21 on the original set.

Although this section only focused on the movie excepts contained on the database, so far, some other content was available on this dataset: A set of 41 previously extracted features from both audio and video domain were made available for each of the movie excerpt presented on the dataset. From the whole 41 features, 21 were related to video and extracted from the video signal of the movie excerpt, while the remaining 20 were related to audio and extracted form the audio signal. It's worth mentioning that some of this features represent coefficients, such as MFCC, so some of the features only have meaning together. This set of features was used as a starting point for the testing work developed on this project. More information about these features are addressed on Chapter4.

3.3 Conclusion

On this section a summary of the system that was developed is done. This system is composed by three major components in order to fulfill its intent of classifying a movie excerpt according to the violence within it. First there is a separation of the signals, followed by the extraction of features and last, the classification of the movie. Also, on this section, an overview of the ACCEDE database is conducted, regarding its content, focusing, not only on the movie excerpts that it contains, but also their characteristics and previously extracted features that were also made available on this dataset.

4 Features

During the process of building, training and testing the system, several groups of features from different domains were used in order to determine which of them have a stronger impact on the detection of violence in movie excerpts. These groups of features can be divided according to three categories: ACCEDE features – features already extracted that are contained on the ACCEDE database, from both audio and video domains; Audio Features – features extracted on the scope of this thesis in order to improve the system, using a tool known as OpenSmile Eyben et al. (2010), that enables the extraction of numerous audio features based on a configuration; and Video features – features also extracted on the context of this thesis, based on the implementation of programs developed for that objective, using the Matlab development tool. These features were computed with the intent to discriminate the behavior of visual features like the color distribution and motion. On the next sections, a deeper description of these different groups of features is performed.

4.1 ACCEDE features

As stated before, the ACCEDE database includes 41 audiovisual features previously extracted in the context of the Liris-ACCEDE project Baveye et al. (2015): 21 video features and 20 audio features. Those features were extracted in the scope of emotion recognition and affective video content analysis, a field that is highly related to the one explored on this thesis. In the context of this thesis, the ACCEDE features were used to setup a starting point, being used as a baseline that is compared throughout the different phases of testing and experimenting, in order to check how the work developed and evolved. These features were extracted and made available in a raw format, meaning they are present on the dataset without any kind of processing other than the one used to extract them. The database fields (feature values) were not normalized, therefore high differences in the range of feature values are observed from field to field. The features were separated into two groups, based on the domain they belong to (audio or video) in order to study them separately and together, as it will be shown in the next chapter that summarizes the experiments. A description of the features representative of the audio signal is performed on Section 4.1.1, while the description of the features representative of the video signal is performed on Section 4.1.2. More details on the ACCEDE features can be found in Baveye et al. (2015) and Baveye et al. (2013a).

4.1.1 ACCEDE audio features

As stated, there is a total of 20 audio related features contained on the ACCEDE database. This set of audio features aims to describe the audio signal using coefficients, volume measures and other kind of techniques.The features included on this set are as follows:

- Mel-frequency cepstral coefficients (MFCCs), which is a group of coefficients that represents the power spectrum of the sound. These coefficients are described by 12 different audio features on the ACCEDE database.

- Centroid, which measures the balance of the sound spectrum and tries to estimate the tone.

- Asymmetry and asymmetry enveloped, two features that measure the behavior of the symmetry of the sound spectrum around its centroid.

- Energy, feature that measures the power of the audio signal.

- Roll-off, which is usually described as the Nth percentile of the power spectral distribution, where the N is commonly about 85% or 95%.

- Zero-crossing Rate (ZCR), which is a rate that evaluates the signal changes along the very same signal.

4.1.2 ACCEDE Video Features

The ACCEDE database includes a total of 21 video related features. These video features can be divided according to two groups: features that intended to represent the static part of the video (still images) and features that represent the dynamic part of the video (e.g., motion). The former group contains features related to color and light distribution and were extracted on a single key frame that represents the whole video sequence. This single key frame was selected based on the color histogram: the frame with the RGB histogram that is closer to the average RGB histogram of the whole video excerpt. The latter group of video features contains features related to the motion and shot transitions (cuts) between consecutive frames, aiming to describe the video based on its flow and not only on a representative frame. The video features existent on the ACCEDE database are the following:

- Alpha, which is the orientation of the color harmonious template, as described on Baveye et al. (2013b).
- Colorfulness, which evaluates the intensity of a specific color. It is calculated as shown in Hasler and Suesstrunk (2003).
- Color Raw Energy and Strength, as well as Min Energy, three features that evaluates the energy and intensity of the color as described in Wang and Cheong (2006).
- Compositional Balance, which measures the organization of the elements in a frame, as described in Luo and Tang (2008).
- Cut length that measures the total length of the cuts on a movie excerpt.
- Depth of field, a features that measures the depth of field using a blur map based method described in Baveye et al. (2012).
- Entropy Complexity, which is used to characterize the scene/excerpt complexity based on the sum of the wavelet sub-bands, as it is calculated in Le Meur et al. (2011).
- Global Activity, the average size of the motion vectors between frames.
- Hue Count, a feature that measures the “simplicity” of a frame as stated in Ke et al. (2006).

- Lighting and Median Lightness, features that evaluates the lighting property in a frame that is used, for example to enhance the 3D visual effect. Luo and Tang (2008).

- Max Saliency Count and Saliency Disparity that counts the number of pixels with highest values in a saliency map and evaluates the differences between the pixels in a saliency map.

- NbFades, NbWhiteFrames, NbCutScenes that count the numbers of fades, white frames and cut scenes in a video.

- Spatial Edge Distribution Area that evaluates the numbers of edges in a video.

- Standard Deviation Local Max that calculates the standard deviation of the Euclidean distance between the local maxima coordinates and the centroid of the local maxima.

4.2 Additional Features Implemented

The ACCEDE related features presented on Section 4.1 consist on previously extracted features that function as the baseline for the work of this thesis. Besides that, one of the thesis goals is to test new features from both the audio and visual (video) domains, that will be used with the intent of enhancing the system's performance. Therefore, and similarly to the ACCEDE feature set, these new features can be divided into two groups according to their respective domains: audio features and video features. Before advancing to the step of extracting new audio and video features, a preliminary study was conducted in order to find possible reasons that explain why a movie excerpt is violent and which visual and audio characteristics in an excerpt are more indicative of violence. Such preliminary study was performed using movie excerpts classified as violent within the first 1000 movie excerpts that are available on the ACCEDE database. Thus, for each movie excerpt classified as violent, a list of possible reasons for the presence of violence was synthesized. Based on the list of possible reasons that lead to the classification of an excerpt as violent, an insight can be provided for the derivation of new features that may enhance the system's accuracy. Additionally, the global reasons for a movie excerpt to be classified as violent can be determined and see if those global reasons can be characterized by audio or video features. Within the first 1000 movies excerpts in the ACCEDE database, 35 were classified

Cause of Violence	#
Blood	12
Quick Movement	11
Scary Shouts	8
Pain Sounds	6
Cry Sounds	5
Angry Voices	3
Explosions	3
Harmful Objects	3
Fighting Sounds	3
Scary Music	3
Heavy Breathing	2
Eyes Expressions	2
Speech	2
Cough and Gasp	2
Image Blur	1
Gunshots	1
Breaking Glass	1
Dead Bodies	1

Table 4.1: Possible reasons for violent classification within the 35 violent excerpts out of the first 1000 excerpts.

s violent and those were the ones targeted by this preliminary study. A list with the main reasons and the number of times they occurred is synthesized on Table 4.1.

After this study was conducted, some interesting conclusions were reached out. The existence of blood in a scene and the quick movement are the most common occurrences and seem to have a stronger link to the existence of violence. However, while a measurement for the amount of movement present in a scene is easily obtained, the existence of blood is hard to quantify. The existence of blood is related to the presence of regions with reddish colors on the frame. However, looking for reddish colors can lead to many mistakes if red objects, like a red shirt, existed on the scene. Furthermore, since the presence reddish objects are more common than the presence of real blood, a feature that measures the presence of red color will most likely have a bad impact on the system.

Rapid movement in a scene is something considered to be potentially indicative of violence and worth of deriving features about. By observing the videos that were subject to this preliminary study, it can seen that most of the excerpts classified as violent that contain a form of physical violence are characterized by the presence of intense motion across the video sequence. Sometimes this observation fails, for instance, when there is a dancing scene or a sports game, which are situations where the video is constantly changing. Nevertheless, motion based features may have some potential on the system and are worth testing.

From Table 4.1, it can also be observed that, from the occurrences of violence indicators in the excerpt, 33 of them are related to video and 38 of them to audio. This fact, together with the fact that many occurrences related to video are harder to be quantified by feature values (i.e., eyes expressions, harmful objects, blood, dead bodies), may lead to the conclusion that the audio features in a movie excerpt may be a little bit more meaningful to the problem of violence classification, despite the video features also playing a big part on the matter.

With this study it was found evidence that extracting video features related to movement and color and, as well as audio features that could represent the audio behavior in a scene should be the path to follow. Therefore, new audio and video features were derived based on the results of this study and other similar analysis of the movie excerpts that occurred during the experimental work performed on the scope of this thesis, both from the ones classified as violent and from the ones classified as non-violent. New audio features that were extracted during this work are addressed on Section 4.2.1, while the new video features are presented in Section 4.2.2.

4.2.1 Audio Features

The OpenSmile Eyben et al. (2010) toolkit was used for the extraction of new features from the audio signal of the movie excerpts. This toolkit works by receiving both the .wav file that corresponds to the audio file extracted from the movie excerpt and a configuration that describes the type of features to extract from the audio file. This toolkit allows the extraction of many groups of features that can be specified through the configuration files. This software contains many configurations available varying on the number of features they intent to extract and, therefore, the level of detail to be worked with. On the experiments conducted, three different configurations were tested: IS13, GeMAPS and eGeMAPS - an extended version of GeMAPS Eyben et al. (2016). The IS13 configuration and feature set was created in the scope of the Interspeech Computational Paralinguistics Challenge 2013 (ComParE), and consists on the use of 6125 features to describe the audio signal. The GeMAPS and eGeMAPS are minimalist sets of acoustic features, proposed for several areas of automatic voice analysis, as a way of reducing the large number of existing features that are usually used to discriminate the audio signals into a much smaller set of features representative of a large percentage of the audio signal. The GeMAPS set of features contained 62 features which used, among other, frequency related features, amplitude related features and spectral related features with the advantage of having a small set, easy to process, containing, altogether, 18 low level descriptors to represent the audio signal. As for the eGeMAPS set, it contains 88 features that can be viewed as an extended version of the original GeMAPS with the addition of 7 low-level descriptors. These three groups of features were used and tested in order to see which of them had a better impact on the problem. It is worth mentioning that after the OpenSmile toolkit outputs the groups of extracted features, they still need some processing, such as the normalization of the group, in order to be served as input of the system.

4.2.2 Video Features

Similarly to the audio features, new video related features were extracted. However, for this task there is the downside of not having a proper toolkit for feature extraction such as the one used for the audio signal, which allowed a transparent extraction of features without the need of programming. Thus, for the case of video features extraction, several scripts were programmed using the Matlab tool . The extracted features were related to the color and luminance distribution on the movie excerpt and features that discriminate the motion between consecutive frames, using low complexity computations. These features values result from globally combining several measurements performed in a frame-by-frame basis.

For the color characterization the first step was to extract the first four color moments from each movie excerpt, with the intent of discriminate the color distribution and behavior in an excerpt. For this procedure, a global histogram with the representation of the color distribution for the whole movie excerpt was computed and afterwards the four color moments were extracted from the histogram:

- Average;
- Standard Deviation;
- Skewness;
- Kurtosis.

In order to characterize luminance distribution, a global histogram of the luminance was computed for the whole video sequence. Then, five features were extracted from the histogram:

- Most common luminance value (the distribution mode);
- Frequency of the most common luminance value;
- Distance (measured in histogram bins) from the mode to the second most common luminance value that is distanced by at least three histogram bins from the mode;
- Frequency of the second most common luminance value;
- Standard deviation of the histogram bins.

As for the motion characterization, the feature extraction process can be split according to two phases. On the first phase, in order to characterize the motion intensity within a movie excerpt, each video frame was split into non-overlapping blocks with dimension of 15x15 pixels. Then, the Mean Squared Error (MSE) values between pairs of blocks located in the same spatial positions of two consecutive frames were computed. After computing all block-wise MSE values along the video sequence, for all consecutive frame pairs, the resulting MSE values below a given threshold were set to 0 in order to suppress noise. Finally, three global features were extracted:

- Average of the frame-by-frame average value of the block-wise MSE values;
- Average of the frame-by-frame average value of the non-null block-wise MSE values;
- Ratio between the number of non-null block-wise MSE values and the total number of block-wise MSE values.

Still on the first phase and related to motion, a set of features was extracted in order to characterize the behavior of the motion vectors orientation, and, to characterize the direction of the movement that exists on the excerpt. Thus, using the same block-wise structure mentioned above, additional measurements were performed in a frame-by-frame basis with the intent of characterizing the distribution motion vectors orientations:

- Average of the angles amplitude for the non-null motion vectors.
- Standard Deviation of the angles amplitude of the non-null motion vectors.
- Skewness of the angles amplitude for the non-null motion vectors.
- Kurtosis of the angles amplitude for the non-null motion vectors.

Name	Description	Detection methods
Hard cut	An abrupt transition where a frame is contained on the first shot and the exact next one is part of the second.	Comparing the color histograms of both shots and having the first one much different from the second
Black frame	When the shot transition is made using a dark frame	When the luminance of the whole frame is above a threshold
Fade in	When a shot starts with a black frame and gradual transitions into the shot	Finding a black frame followed by a gradual brightness increase, in terms of luminance
Fade out	When a shot gradually derives into a black frame	Finding a black frame that is preceded by a gradual transition from a lighter frame into a dark frame
Others	When there are detected many successive hard cuts	Using the same method as hard cuts but with a bigger set of consecutive frames

Table 4.2: Shot transitions used and how were they extracted

Since the features must be representative of the whole movie excerpt and not of for individual frames, a combination of these measurements was performed and it resulted on 8 features: These 8 features result from computing the average and the standard deviation of each measures along the whole video sequence.

On the second phase, a series of features were extracted based on the very same method as the one used on the first phase, in a deeper way and with the inclusion of another paradigm: the existence of shot transitions on the movie excerpt. Using the Matlab software, a program was developed for detecting shot transitions in a movie and identify them as one of the following five categories: hard cut, black frame, fade in, fade out or others. Table 4.2 shows some more information about this shot transitions and how were they detected.

With the identification of the video shot boundaries, it is possible not only extract features related with the amount of shot transitions in a movie excerpt and their type, but also to split a movie excerpt according to smaller segments delimited by the identified shot boundaries. The identification of shot boundaries allows the extraction of features for each one of the smaller pieces of movie excerpts and to combine them, something that could not be done before. As said before, in this phase, a wider set of features was extracted. Besides the extraction of the three motion features related with the block-wise MSE described for the first phase, other sets of features were extracted using other mathematical operations to represent the motion intensity (MSE) related information, different than or combined with the average. The first operator is the maximum, from which resulted four features that represent the motion intensity information:

- The maximum of the frame-by-frame average value of the block-wise MSE values.
- The maximum of the frame-by-frame average value of the non-null block-wise MSE values.
- The average of the frame-by-frame maximum value of the block-wise MSE values.
- Ratio between the number of non-null block-wise MSE values and the total number of block-wise MSE values.

Other than that, a different kind of average was also tested, repeating the motion intensity features of the first phase, but instead of using the usual average, each MSE value was raised to the fifth power, sumed and then calculated the sum's fifth root. This calculation intended to highlight the higher values on the matrix. After these calculations, the same three statistic methods that were calculated for the MSE values, on the first phase, were extracted, but this time for this new set of features, with the intent of analyzing the motion intensity information were derived . After that, the function to detect shot transitions on the movie excerpt was used, in order to identify which frames of the excerpt belong to a shot transition and remove these frames from the computations. This process allows to remove outlier values on the motion measurements performed between consecutive frames that belong to different video shots. Therefore, the measurements performed in these frames were discarded and the three global feature values were calculated. After that, other features based on the standard deviation were tested. The first set is the following:

- Standard deviation of the average of the frame-by-frame average of the block-wise MSE values.
- Standard deviation of the average of the frame-by-frame average of the non-null block-

wise MSE values.

The second one is composed by only one feature which is the standard deviation of the average value of the MSE values of each small piece of the movie excerpt delimited by the shot boundaries.

Still on the motion features topic, and on the second phase of testing, other features related to the existence of shot transitions on the movie excerpts were derived. The first one was composed by only one feature and corresponded to the number of shot transitions present on each movie excerpt. The second one, is a more complex one, containing five different features identifying the type of shot transition and how many times it occurred on a single movie excerpt:

- The number of hard cuts in a movie excerpt.
- The number of black frames on a movie excerpt.
- The number of fade in occurrences on a movie excerpt.
- The number of fade out occurrences on a movie excerpt.
- The number of other cut types that are not any of the previously mentioned on a movie excerpt.

4.3 Conclusion

This chapter presented a brief description of all the features that will be used for testing the proposed violence classification system. The chapter started by the describing the ACCEDE features, the set of features contained on Liris-ACCEDE database. This set of features consists of 41 features, where 21 are from the video signal domain and 20 are from the audio signal domain. Then, the audio features researched and extracted in this work were described. These audio features were extracted using the OpenSmile toolkit Eyben et al. (2010), setup with three different configurations, resulting on three sets of features: IS13, consisting of 6125 features, GeMAPS, with 62 features and eGeMAPS, an extended version of GeMAPS, with 88 features. Last but not least, additional video related features were derived in the scope of this work. These features were extracted using scripts developed during the thesis using the Matlab tool. Several sets of features were extracted:

color related features that described the color behavior on the movie excerpts; luminance related features that explore how the luminance is distributed on the excerpts; motion related features, which were obtained in two phases. On the first phase, the extracted motion features are based only on measurements that use the Mean Square Error between frames and features that described the motion vectors orientation behavior. On the second phase, additional features that describe the number and type of shot transitions present on the excerpts were extract, and this information was used for the extraction of Mean Squared Error based features, similarly to the the first phase, but by removing measurements corresponding to frames located at the detected shot boundaries. Features based on the standard deviation of the Mean Squared Error blocks were also extracted. The next chapter presents experimental results for these groups of features, individually and combined with each other.

5 Experimental Results

After defining different groups of features, already contained on ACCEDE database or extracted in the scope of this thesis, a series of experiments were conducted not only for establishing the starting point of this work (using only the ACCEDE database features), but also to evaluate which of them were more relevant for the existence of violence in a movie excerpt, leading to improvements on the system's accuracy and performance. For all the experiments described along this chapter, the performance measurement used for evaluating the system's accuracy is given by the percentage of movie excerpts correctly classified by the system (both as violent and non-violent). As stated before, the two classifiers used in the experiments are Neural Network and Support Vector Machine with a Polynomial function kernel. After training the system in the different scenarios, the experimental evaluation is was performed using ten-fold cross validation technique. The experiments can be split according to two phases: the first phase, where the ACCEDE features, the audio features and the video features that correspond to the first phase of video features extraction (see Section 4.2.2) are tested separately and combined with each other; and the second phase, where the more complex video features set corresponding to the second phase of video features were combined with the results of the first experiment. This chapter is thus organized according to these experimental phases: Section 5.1 shows the experiments conducted on the first phase and Section 5.2 shows the experiments conducted on the second phase of testing that resulted on the best system's performance.

5.1 First phase of experiments

On this section, the results of the first phase of experiments are presented, divided by the type of features that feed the classification system. On Section 5.1.1 the results of

Classifier \ Features	video only	audio only	all features
Neural Network	64.20	65.23	66.02
SVM	65.80	60.57	67.39

Table 5.1: System performance using ACCEDE features

the experiments conducted using the ACCEDE features alone are presented. Section 5.1.2 presents the results of the testing of audio related features while Section 5.1.3 depicts the results of the tests whose aim was to check out the impact of using the video features on the system. Finally, Section 5.1.4 shows the results of the experiments conducted on this first phase of experiments, when combining different groups of features.

5.1.1 ACCEDE Features Testing

The first experiment consisted of testing the features included in the ACCEDE database, on two classifiers (neural Network and SVM), implemented on Weka Holmes et al. (1994). These features were first normalized and then fed to the classifier. This process of normalization consisted on arrange the column values, in order to avoid differences of magnitude. To calculate the normalized value of an element of the column, first the average value of the column is computed, as well as the standard deviation. Then the value is subtracted by the calculated average and the result of the operation is divided by the standard deviation.

The test was performed separately for the audio and video features present in the ACCEDE dataset as well as for their combination, using the whole set of ACCEDE features. This experiment sets up an initial benchmark for the system's performance and provides hints regarding which modality of features, audio or video, seems to have a stronger impact on the system. The results for the percentage of correct classifications outputted by the system are presented on 5.1.

Using all ACCEDE features, the system managed to perform decently and reached a performance of more above 66% for both classifiers. Table 5.1 also shows that, for the Neural Network classifier, using features taken from the audio signal only or from the video signal only leads to similar performance. This result is somehow different for the SVM classifier case, where using the video features included on the ACCEDE database lead to better performance than using the audio ones, with a difference of more than 5% in the accuracy. These results establish the starting point of this work, which aims to increase the performance of the system using other features. With this experiment it can also be observed that the audio features seem to be slightly more relevant for the task of detecting violence in movie excerpts, since they significantly performed better than the video features for the SVM classifier despite leading to similar results in the case of the Neural Network classifier based system.

5.1.2 Audio Features Testing

This next experience was conducted in order to evaluate the classifiers' performance using the additional audio features described in Section 4.2.1(IS13, GeMAPS and eGeMAPS). These audio features were also tested on both SVM and Neural Network classifiers. The performance achieved using each group of features, after a normalization procedure applied to all sets of features, is shown on Table 5.2. These results confirm that audio features provide good hints for the existence of violence in movie excerpts and, surprisingly, they even outperform the results achieved by using the whole group of ACCEDE features (which include audio and video features). By comparing the results achieved using the features on the ACCEDE database depicted on Table 5.1 with the results presented on Table 5.2, it can be observed that even the worst results attained by using theses additional audio features alone are better than using the ACCEDE features. Despite being the smallest audio feature set, the GeMAPS feature set reach a considerable and consistent performance for both classifiers. When comparing GeMAPS and eGeMAPS, it can be seen that the first ones perform better on the Neural Network classifier and achieve similar performance on the SVM classifier. Additionally, the GeMAPS set contains a smaller number of features than the eGeMAPS set, which speeds up the system. This statement is even more significant when comparing with the IS13 feature set, which lead to the worst performance of all three, despite the large number of features present in his set (6125 features). It is also worth to mention that the Weka software did not run properly when trying out the Neural Network classifier using the IS13 feature set and eventually stopped after processing for a few hours. This enhances the importance of having a smaller features set that also performed better.

Classifier \ Features	GeMAPS	eGeMAPS	IS13
Neural Network	71.93	70.68	-
SVM	72.95	72.95	68.75

Table 5.2: System performance using Audio features

Classifier \ Features	Color	Luminance
Neural Network	47.16	55.68
SVM	46.59	54.32

Table 5.3: System performance using Color and Luminance related features

5.1.3 Video Features Testing

After the experiments using the ACCEDE features and the additional Audio features extracted in the scope of this work, the system was evaluated using only the video features described in Section 4.2.2. In order to find out hints for the relative importance between motion intensity, motion orientation, luminance and colors based features, the classifiers were evaluated each of these groups of features separately.

In a first experiment, the color and luminance features were tested for both neural network and SVM classifiers. The corresponding results can be observed in Table 5.3.

The results presented on Table 5.3 show that the video color features do not have a positive influence on the system decision of detecting violence on the excerpts. The system's performance using color features alone is even lower than 50%. On the other hand, the use of luminance related features shows a slightly positive impact on the system, despite of not being too discriminative for violence, at least when used alone. Regardless of having a positive influence on the system (above 50%) the results are not satisfactory since there is only 55.86% of correctness for the better case. Nevertheless, it may be worth to use these luminance features combined with other types of video features, in order to check out if their use improves the system correct classifications.

After the experiments using color and luminance features, the system was tested using motion related features, also described on Section 4.2.2. These features, which measure the motion intensity and motion orientation using the motion vector angles, were fed to the system and their results are shown on Table 5.4, as they belong to the group of features extracted on the first phase of the experiments.

Classifier \ Features	Motion Intensity	Motion orientation
Neural Network	62.39	59.66
SVM	57.05	59.09

Table 5.4: System performance using motion intensity and motion orientation features

By analyzing the above table, it can be seen that both features show potential for the violence detection task, in terms of system performance. The motion intensity features reached the 62.39% of correctness when using the Neural Network classifier. Therefore, their result will be taken in consideration despite not reaching the results achieved with the audio features. As for the motion orientation features, they performed better than the motion intensity features for the SVM classifier. However, they performed considerably worst for the neural network classifier. This set of features reached a maximum of 59.66%, for the neural network classifier. Table 5.4 also shows that the best performing video features extracted in the context of this thesis (motion intensity features) reached a performance that is close to the 21 video features contained on the ACCEDE database, whose results were depicted on Table 5.1.

Up to this point, the different types of video features have been used separately and since some of them (motion intensity features, motion orientation features and luminance based features) showed some potential, additional experiments were performed in order to assess the effect of combining video features from different types. Four distinct video features sets have been defined:

- Set 1, which was a combination of luminance based features and motion intensity features;
- Set 2, which combined motion intensity features with motion orientation features;
- Set 3, which combined motion orientation features and luminance based features;
- Set 4, which was a combination of the three types of video features extracted (except color based features).

The results achieved with each set are presented on Table 5.5.

As it can be seen from the table, the best performance occurs for the set combining luminance features and the motion intensity features, using a Neural Network classifier.

Classifier \ Features	Set 1	Set 2	Set 3	Set 4
Neural Network	64.66	60.80	62.16	62.95
SVM	61.14	60.80	62.64	62.39

Table 5.5: System performance after combining different of video features types

This set of features performs better than the others, with the exception of the Set 3 and Set 4 on the SVM classifier. The maximum performance that was reach in this test was 64.66 %, which is even closer to the ACCEDE database video features' performance for the SVM classifier and better than those features for the neural network classifier. That is remarkable when considering that the luminance and motion intensity features are easier to extract and the extracted set contains less features than the 21 video features included on the ACCEDE database, as the luminance features are composed of only 5 features while the motion intensity features are composed on 3 features.

5.1.4 Multi-modal Feature Set Testing

After testing both the extracted audio and video features, it was observed that the audio features performed better than the video ones. Besides this result, audio features are easier to extract since their computation requires much less input data than the computations required to extract video features. This allows to conclude that the audio features have a bigger impact on this task than the video features. Since both features had a good impact on the system correctness, and despite the audio features had performed better, an approach that consists of combining both modalities of features was followed. Therefore, the best performing audio features were combined with the best performing video features and the ACCEDE features. In order to evaluate the effect of such combinations, three sets of features were defined and tested on the neural network and SVM classifiers:

- MultiModal1, a set containing the best performing audio features (GeMAPS) and the best individually performing video features (motion intensity features);
- MultiModal2, a set that extends MultiModal1, by adding luminance related features, since they performed well when combined with the motion intensity features;
- MultiModal3, which is a combination of the content of the set MultiModal2 and the ACCEDE features.

Classifier \ Features	MultiModal1	MultiModal2	MultiModal3
Neural Network	71.36	73.75	73.64
SVM	74.43	74.55	74.55

Table 5.6: System performance using different combinations of feature sets

The results of this experiment can be seen on Table 5.6.

The results described on the table show that the best system performance is achieved using the SVM classifier, for both MultiModal2 and MultiModal3 multi-modal sets of features. It is also worth to point out that the results attained in MultiModal1 are very close to the other two, indicating that the contribution of ACCEDE features and luminance-based features is marginal. The maximum performance of 74.55% represents a considerable increase when compared to the baseline performance achieved using the initial set of ACCEDE features. These results also confirm that the audio features are probably the most important for this task, since the inclusion of video features only resulted in a small increase of 1.5% when compared with the results obtained using the audio features alone, that can be seen on Table 5.2. Nevertheless, the use of video features lead to the highest performance. The positive impact on the system's performance showed by the reduced number of video features used in the experiments showed motivated a search for additional video features that may enhance the system performance even more, which resulted on a series of new tests described on the next section.

5.2 Second phase of Experiments

The second phase of experiments consisted on a series of tests targeting the performance achieved by using additional video features. Since this modality of features showed worst performance when compared to the audio features, it was considered to be worth to find better video features that would lead to improvements on the system. Therefore, this section was divided in two minor sections: Section 5.2.1 addresses the experiments conducted on the new video features, extracted with the intent of achieving a better performance than the video features tested on the first phase of experiments, while Section 5.2.2 describes the testing of the new video features when mixed with the best features from other domains, with the intent of enhancing the system performance obtained on Section

Classifier \ Features	AverageMSE	MaxMSE	AverageRoot5MSE
Neural Network	62.39	58.64	57.73
SVM	57.05	58.3	55.34

Table 5.7: System performance using different statistic methods without frame removal, for motion intensity characterization

Classifier \ Features	AverageMSE	MaxMSE	AverageRoot5MSE
Neural Network	60.91	59.2	60.45
SVM	55.45	58.86	57.16

Table 5.8: System performance using different statistic methods with removal of frames where existed shot transictions, for motion intensity characterization

5.1.4.

5.2.1 Second phase Video Features Testing

First, experiments were conducted having as test focus the same motion intensity features as the ones described before, on this section, but introducing other types of information combination: Maximum and Average with root 5 (explained on Section 4.2.2). These very same tests were also conducted using a preprocessing of the MSE values that consisted on removing from the calculation the frames that were identified by the shot transition detection function. The results of these tests are presented on Table 5.7, for the case that didn't include any frame removal, and on table Table 5.8 for the case which included cut frames removal.

By analyzing the two tables above, it can be concluded that the motion intensity feature used on the first phase of experiments (AverageMSE on Table 5.7) keeps on being the best performing feature among those extracted using the Mean Squared Error. Despite that, it can be seen that the remaining methods of extraction (maximum and average using root 5) showed better results when there was frame removal than when there wasn't. That and the fact that the MSE using average for the case where there was frame removal is not that far from the exact same feature on the opposite table, indicates that might be some value for the problem within the results presented on Table 5.8. Both of these results will be taken into consideration and it will be checked to see if they can improve the system, especially because the second table contains results that not only describe movement between frames, but also remove the cases where there was considered to exist movement, but there was

Classifier \ Features	STD1	STD2
Neural Network	63.07	59.77
SVM	56.02	57.05

Table 5.9: System performance using Standard deviation based features

just a change of the visual plan, a fade or other related events.

In the very same phase of video related features, they were, also, extracted features related to the standard deviation between consecutive frames. This approach intended to evaluate if the matrices' values changed in large scale which is related to the existence of motion, due to the big visual changes between frames. In this scenario, two sets of features were extracted. The first one (STD1) used the function that detects shot transitions in a movie, in order to divide the movie excerpt in small pieces and then calculate the average of each small piece's matrix values, as a way of, finally calculate the standard deviation based on the value of each small piece of movie. The second one (STD2) was basically calculating the standard deviation of all the matrices values of each movie excerpt and do the same operation but only considering the non-null values. The results of the two sets of features are shown on Table 5.9.

As it can seen by analyzing the Table 5.9, both features tend to be an indicator of existence of violence, especially the first ones. This results are good, when considering that the first set of features is composed by only one feature and the second set contains only two features. In both cases, the neural network classifier showed the best results and the best test occurred for the STD1 set (only one feature) and achieved 63.07% of correctness. As the standard deviation is a measure that describes how the values behave along the movie excerpt, these two features represent information about how much the excerpt change along the way. These are another visual features related to the movie behavior that had a reasonable performance and should be considered for the final system.

The next motion features tested are the remaining features of this second phase of motion features, which are related to the existing shot transitions on a movie excerpt. Therefore, there was focus on testing the one that only included the global number of shot transitions on a movie excerpt (NumCuts) and the one that contains the information of the different types of shot transitions that exist on a movie excerpt (TypesCuts). The results are presented on the table Table 5.10.

Classifier \ Features	NumCuts	TypesCuts
Neural Network	64.2	63.98
SVM	53.98	60.11

Table 5.10: System performance when testing shot transitions related features

As it can be seen by analyzing the Table 5.10 both sets of features seem to have interesting results, considering they are composed by small amount of features (NumCuts contains only one features, while TypesCuts contains five), which indicates that these features have a considerable impact on the task of detecting violence on movie excerpts. In both cases the best results occur when the sets of features are tested on a neural network based classifier and, although the best result happen for the NumCuts feature tested on the already mentioned classifier, the TypesCuts feature seem to have enforced results because they are just 0.22% below the other set of features and have a better result on the SVM classifier of more than 6% when comparing to the NumCuts features.

After this second phase of feature testing, a combination of features were made with the intent of enhance the system performance and to help choose which features are used in the final system.

On this second phase of motion related feature extraction and testing three features showed the most promise and had the best performance on the system: NumCuts features, that has the count of the number of shot transitions existing on a movie excerpt, TypesCuts, the set of features that has the count of shot transitions separated in five different types of features each one counting a specific type of cut, and the set of features that is calculated by applying the standard deviation of the average of each small shot of a movie excerpt extracted by dividing the excerpt when encountered a cut, STD1. As this three sets of features showed the best promise, different combinations of each other were tested. Therefore, the first combination of features tested was TypesCuts with STD1, referred as TypesSTD. Secondly, a combination of the Types Cuts features with NumCuts which is referred as TypesNum was tested. Lastly they were combined NumCuts with STD1, features that are mentioned as NumSTD. The results of the experiment are shown on table Table 5.11.

After analyzing the results obtained on the Table 5.11, it can be seen that all the experiments weren't successful at all, since in all the cases the system performance ended

Classifier \ Features	TypesSTD	TypesNum	NumSTD
Neural Network	63.98	63.75	63.40
SVM	61.83	60	59.32

Table 5.11: System Performance after combinations between NumCuts, TypeCuts and STD1 features

up dropping or at least stays as high as one of its individual sets of features, and none of the tested subjects even reach the NumCuts features performance. In the TypesSTD, the testing on the neural network classifier resulted on the performance being the same as if it was only used the TypeCuts features, as the testing on the SVM resulted on increasing the best performance when comparing to TypeCuts individually, for this classifier, in more than 1%. However, the performance is still lower than the one we get by testing the features on the neural network classifier. In TypesNum, the results of mixing NumCuts and TypeCuts resulted on a performance, in general, worse than each of them individually, for the Neural Network classifier. Even on the SVM classifier, the results weren't as good as using the Types Cuts individually. In NumSTD, the mix of NumCuts with STD1 resulted, like it happened on the previous ones, in a performance worse than the best one individually. Also, mixing the NumCuts features with STD1 gave a worse result than mixing TypesCuts with STD1, even considering that NumCuts individually showed a better performance than TypeCuts.

After testing the combination of features referred above, the focus was, next, on the features that consist on calculating the average, the maximum and the average with root five on the MSE values as described on Section 4.2.2. This features, were calculated both having the frames that contained shot transitions on the movie excerpts removed or not removed. Therefore different combinations of this features were tested on the system. Thus, the referred features were grouped together considering if they have or not the frames where a shot transition was detected, included on the calculation. A combination of both the features presented on table Table 5.7, as they were calculated without removing the frames where shot transitions were detected, and the features presented on Table 5.8 with the set of features called STD2 able to be seen on table Table 5.9 , was performed, since they were calculated with the removal of the frames where shot transitions existed and were detected using the already mentioned cut detection tool. The results of this experiment are described on Table 5.12, having the first combination mentioned as CombCuts and the

Classifier \ Features	CombCuts	CombRemCuts
Neural Network	57.27	61.25
SVM	58.52	60.23

Table 5.12: System performance using combinations of different types of motion intensity related features

Classifier \ Features	TypeCuts + CombRemCuts
Neural Network	SVM 63.98

Table 5.13: Combination of TypeCuts and CombRemCuts features

second CombRemCuts.

With the analysis of Table 5.12, it can be seen that in the first combination of features, referred as CombCuts, resulted on a system performance worse than all of the individual features that compose the set, for the neural network classifier. Despite the mix of features conducted to a better result for the SVM classifier, it is still 3% less than the best performing feature on the best performing classifier. For the second combination of features, CombRemCuts, the mix of features resulted on a small but better improvement on the system performance for both the classifiers. Although both the combination of this features presented on Table 5.12 and the features individually are somehow indicative of violence in the movie excerpts, their system performance isn't as good as some of the motion related features that were tested before.

The next step of the experiment was created with the intent of adding information about the existence of shot transitions on the movie excerpt, existing on TypeCuts features (that are represented on Table 5.10), to CombRemCuts that are shown on Table 5.12, to see if the two different features would complement each other and result on a better system performance. So the two features were mixed and the results are shown on Table 5.13.

As it can be seen by analyzing the Table 5.13, the results obtained showed that the best system performance was achieved using SVM classifier and resulted on the exact same performance as if the TypeCuts features were used all alone. That helps to conclude that the CombRemCuts didn't mattered that much during the process. That is reinforced by the fact that on the Neural Network classifier, the performance is worst when comparing with the one obtained by using, only, the TypeCuts features.

After testing all this hypothesis and features, not only individually, but also by com-

Classifier \ Features	Final1	Final2	Final3	Final4
Neural Network	74.66	73.75	73.98	74.55
SVM	75	76.59	75.34	76.70

Table 5.14: System performance when swapping the former motion features for new ones

bining them, the impact of the better ones on the final system described on the previous section was tested. The features, on the second phase of experiments that result on a better system performance were the NumCuts, TypeCuts and STD1. However, when combined the performance dropped.

5.2.2 Final Multi-modal Features Testing

After testing this different kinds of video or visual related features that were intended to find new and good representative video features, it can be seen that some of them show a reasonable potential and can actually contribute to the final system's performance. Therefore, the best performing system from the first phase of experiments was used, which is a combination of GeMAPS features, first phase motion related features and ACCEDE features, and make small additions of features extracted on second phase of experiments to it, with the intent of improving the system. The Luminance related features were left out, since despite their addition had conducted to an increase of the final system performance for the SVM classifier, it had no impact on the Neural Network classifier, which is the best performing one.

Having that in mind, the best performance features that were tested on the second phase of experiments were selected and added to the system. Thus, the first step was to swap the former used motion features with the three different motion features calculated by removing the frames where a cut was happening, in order to see if they would be better on the system and provide additional information, that the other ones didn't, despite they performed worst, individually. So, three different tests, each one swapping the motion features for AverageMSE, MaxMSE and AverageRoot5MSE, presented on Table 5.8 (referring to the experiments as Final1, Final2 and Final3, respectively), were conducted. Also, it was also tested a mix of all of Table 5.8 features, to see if they, together, had a better impact on the system (Final4). The results are shown on Table 5.14.

As it can be seen by analyzing the table above, just by swapping the former used motion

Classifier \ Features	Final4+Num	Final4+Types	Final4+STD
Neural Network	76.47	76.02	75.34
SVM	76.02	75.68	76.25

Table 5.15: System performance obtained by mixing the best performance features of both stages of experiments

related features, extracted during the first phase of experiments, with the new motion intensity related features calculated with a previous removal of the frames where a cut was detected, an important improvement on the system was achieved. The first thing that can be seen is that all of the new experiments resulted on a better system performance than the performance obtained using the former motion features and consequently the better combination of features until this stage. That means that the newly extracted motion features had a better impact on the system. The best result got was achieved by mixing all three motion related features and resulted on an improvement of more than 2% when comparing to the former best solution, achieving 76,70%. It is also worth mentioning that the best individually used feature, from the set of the motion features, extracted on the second phase of experiments and tested above, was the one calculated using the maximum.

After the test described above, different and new features were tested with the intend of improving the system results, in this case, the best performing ones from the second phase of experiments. Therefore, to the best performing system (including the addition of the new motion related features), were added NumCuts, TypeCuts and a mix of both the standard deviation related features described on Table 5.9, separately. The results can be seen on Table 5.15, having in consideration that the best performing features are mixed, in Final4+Num with NumCuts, in Final4+Types with TypeCuts and in Final4+STD with the Standard deviation related features.

By analyzing the Table 5.15, it can be concluded that the system performance didn't overcome the best result until now. However, the best result for the neural network classifier was obtained in Final4+Num: 76.47. This, also come to reinforce the fact that the NumCuts features where the best and more linked with violence detection of all three. Despite that, the best system in terms of performance continued to be the mix of ACCEDE features, GeMAPS features and motion features from the second phase of experiments, and that's the final system.

5.3 Conclusion

This section focused on the experiments conducted to test the features described on Chapter 4, on both SVM and Neural Network based classifiers. The experiments were divided in two phases. On the first phase of experiments, the focus was on testing the ACCEDE features, from both audio and video domain, which represented a baseline for this work. Next, the audio features that were extracted on the scope of this work were tested, which, by themselves showed a good performance. After that the focus was on testing the first video features, which showed worst performance than the audio features. Finally a mixed all of them was performed, resulting on a top system performance of 74.55%. The second phase of experiments intended to test new video features, due to the fact that the first ones were considerably worst that the audio features. That fact also explains why there were extracted more sets of video features than sets of audio features. First they were tested the new video features and found some of them, especially the ones related to the shot transitions within the movie excerpts, very interesting and with some potential. After that a mix of the best performing features from this phase with the best performing features from the first phase of experiments was conducted, resulted on a top system performance of 76.70% which represents an increase of almost 2% when comparing to the multi-modal features from the first phase of experiments.

6 Conclusions and future work

The work presented on dissertation focus on developing a system that is able to detect violence on movie excerpts using multi-modal features from both audio and video domain. This task and idea to built this system, comes from the current problem that many parents and tutors are having nowadays, which is, not being able to control their children's access to violent content on the many types of media. Therefore this system was built with the intent of giving a baseline for the built of a tool that is able to detect if a movie is violent or not, which would help the decision that a parent or tutor has to make of letting their children watch certain content.

The goal was to develop a system that, given an excerpt of movie, was able to classify it as violent or non-violent. To do that the system was divided on three major components. The first one had the function of separating the audio and video signals from the movie excerpt, using the FFMPEG tool, outputting them in their raw format so they could be processed, in order to extract features from them. Which is exactly, what is the second major component is about: Extract features from the audio and video signals using the available and open-source toolkit OpenSmile (Eyben et al. (2010)) for the audio features case, and using Matlab as a development platform for developing scripts to extract video features, since it wasn't found any toolkit, for video features, similar to the OpenSmile, for audio features. Finally, the last component of the system was the classification component, where two selected classifiers (SVM and Neural Network) use the features of a movie to output the binary decision of classifying the movie as violent or non-violent.

In order to train and test the system, there was a need of a dataset of movie excerpts, which was provided by the ACCEDE database (Baveye et al. (2015)), which contained 9800 movies already classified as violent or non-violent according to a crowdsourcing exercise. The dataset had to be balanced due to a big disparity between the number of movie excerpts classified as violent and the number of movie excerpts classified as non-violent. After some tests, it was concluded that a dataset that balanced usually drives the classifier to always decide that the movie excerpt is non-violent, due to the fact that the proportion between non-violent movies and violent movies were 21/1. Therefore, the final dataset was composed by 880 movie excerpts, 440 classified as violent and 440 classified as non-violent.

The major contribution of this work, for the scientific community, was the study that was conducted in order to search for possible features that might have influence on the system decision, the combination and testing of them, on the system and the conclusions that resulted from that. A series of many experiments for testing every single group of features and other combinations of features on both the SVM based and Neural Network based classifier was conducted.

The first features tested was a set of 41 features contained on the ACCEDE database that included both audio a video features. This ACCEDE set of features was used to establish the starting point for the work presented on this dissertation, in terms of system performance. This set of features achieved a maximum performance of 67.39% of correction using the SVM classifier.

After testing the ACCEDE features on the system, the focus was on finding and testing audio features. To do that, using OpenSmile toolkit, they were extracted three different groups of audio features based on different configurations: IS13 features, a group of 6125 features, GeMAPS, a minimalist set that contained 62 features, and eGeMAPS, an extended version of the previous ones with 88 features. While testing this features, the results were somewhat interesting: Despite being a set composed by less features then the other ones, GeMAPS features actually achieved the best performance of all three for the Neural Network classifier, and for the SVM classifier (Having the same performance, in this case as the eGeMAPS features), while the IS13 features, which was composed by an enormous number of features, achieved the worst performance, with the aggravating of not being able to be tested on Neural Network classifier, due to its large number. The best performance was of 72.95% for the GeMAPS and eGeMAPS on SVM classifier and, already presents an improvement when comparing to the ACCEDE features. With this experiments it was concluded that this audio features not only have a big influence on matter, but also are relatively easy and quick to process.

Having the audio features in a good level, already, more focus was given to the video features, especially because there wasn't any toolkit for their extraction. In a first phase, two major groups of features were tested: color and luminance related features, and motion related features. The first one were measures extracted from the color and luminance histogram, such as the color moments, for the color case, and measures that described the most common luminance values within the histogram for the luminance case. For the motion related features, measures were extracted based on the Mean Squared Error (MSE) between block of consecutive frames, as a way to discriminate the motion intensity, and about the motion vectors angles distribution and behavior, to characterize motion orientation. While the color and luminance histogram didn't performed as expected and achieved very poor performance, motion features, especially the motion intensity related ones achieved a considerable performance of 62.39% for the Neural Network case.

At this point many combinations of features from the both domains were tested in order to achieve a better performance. Also, the ACCEDE features were also included on this testings, since their performance was, still, very interesting, which means they were correlated with violence within the movie excerpts. The best combination of features was achieved combining the GeMAPS features, the motion intensity features and the luminance related features, which, despite having a poor individually performance manage to have a very small improvement on the system when comparing to the very same system that didn't include them, and resulted on a performance of 74.55%.

Finished the first phase of experiments, some conclusions were taken. Not only the video features add a worst performance when comparing to the audio ones, but they are much harder to extract. Although, since at this point the audio features were already in a very interesting level, a second phase of experiments was conducted in order to try to find better video related features, that would have a better impact on the system, as they represented the weakest part of this work, so far. Therefore, the second phase of experiments addressed new video features extracted based on the ones from the first phase, but with a paradigm change, and, also, taking shot transitions into account. This paradigm change resulted on removing frames from the equation, where a shot transition occurs, and, in some cases, instead of extracting features directly from the frame level to the excerpt level, it was performed a division of the excerpt in small shots divided by a transition, extract features for them and them combined them to achieve the excerpt features. Thus, there were extracted the motion intensity features again but, this time, with the removal of the frames where a shot transition occurred, and not only using the average statistic: maximum and average with root 5. Also, features related to the standard deviation of the MSE values were extracted, both using division of the excerpt in shots and not. Last but not least, two sets of features were also created, to address the number and the type, respectively, of shot transitions. After testing all sets of features, all of their maximum performance was between 60 and 65% which is considerably good, especially since this sets of features were mostly composed by a small number of features.

Having the features from the second phase of experiments extracted, many different combinations of features from both phases of experiments were tested, in order to try to improve the top performance achieved during the first phase of experiments, of 74.55%. The combination that resulted on the best performance was one that used the final combination of features of the first experiments, but replaced its hold motion intensity features with the motion intensity features that went through a process of removing the frames where a shot transition occurred, from the calculation and resulted on a final performance of 76.70%, for the SVM classifier. This result showed that despite the motion features' performance, individually, do not reach the performance obtained with the audio features, the fact that a set of motion features were swapped for others resulted on an improvement of more than 2% which is considerably good, specially when the performance was already high.

Finalized the experiments some conclusions can be taken: As stated before, audio features seem to have a bigger impact on the system and are far easier to process. Also, it is possible to extract set of audio features, based on a configuration, using available toolkits, something that is not that easy to find for the video features. For the video features case, unlike it was initially thought, the motion features and features that describe the activities related to shot transitions on a scene, seem to have a better impact on the task when comparing to color or luminance related ones. Also, in general, the SVM classifier seem to work better for this problem, when comparing to the Neural Network, despite their close final performance. This advantage for the SVM classifier, happens, mostly, due to the fact that the audio related features (the features that had the best individual performance) had the best results using the SVM classifier, since for some of the video related features, the Neural Network based classifier had better results. Finally, it was clear that the best results are achieved using selected features from both domains, instead of using only audio or video related features individually.

As future work, the most important task there is, is to find other options for the video features the have a better impact on the system, when comparing to the ones used and described on this thesis, and also, video features that are quicker to extract. The speed of extracting and processing feature is quite important, since another topic that may be addressed on the future, it to built the fully functional and real-time violence in movies detector, with the possibility of web integration, as a way of being available for the biggest number of users possible. Also, a step that might be taken in the future is to use and test other ways of classification and machine learning algorithms to see if that techniques improve the system performance. Other aspect that may be interesting to address is test other kinds of features, for example, text related features, that can be extracted from the subtitles, to see if that kind of features have a connection to violence. The same is valid to speech recognition which can also be considered. Even the color related features, that didn't perform as expected may be target of a more incisive research to find possible solutions to increase their value and performance on the system, based on different types of information representation, or addressing other topics related to color.

Since the field of detecting violence in movies/movie excerpts using machine learning algorithms is a relatively new topic, many steps can be taken following this original work. Many different approaches can be taken in so many levels to continue to contribute for the area.

Bibliography

Acar, E. and Albayrak, S. (2012). Dai lab at mediaeval 2012 affect task: The detection of violent scenes using affective features. In *MediaEval*. Citeseer.

Acar, E., Hopfgartner, F., and Albayrak, S. (2013). Violence detection in hollywood movies by the fusion of visual and mid-level audio cues. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 717–720. ACM.

Baveye, Y., Bettinelli, J.-N., Dellandréa, E., Chen, L., and Chamaret, C. (2013a). A large video database for computational models of induced emotion. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 13–18. IEEE.

Baveye, Y., Dellandrea, E., Chamaret, C., and Chen, L. (2015). Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55.

Baveye, Y., Urban, F., and Chamaret, C. (2012). Image and video saliency models improvement by blur identification. In *International Conference on Computer Vision and Graphics*, pages 280–287. Springer.

Baveye, Y., Urban, F., Chamaret, C., Demoulin, V., and Hellier, P. (2013b). Saliency-guided consistent color harmonization. In *Computational Color Imaging*, pages 105–118. Springer.

Bushman, B. J. and Huesmann, L. R. (2006). Short-term and long-term effects of violent media on aggression in children and adults. *Archives of Pediatrics & Adolescent Medicine*, 160(4):348–352.

Chen, L.-H., Hsu, H.-W., Wang, L.-Y., and Su, C.-W. (2011). Violence detection in movies. In *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pages 119–124. IEEE.

Dai, Q., Wu, Z., Jiang, Y.-G., Xue, X., and Tang, J. (2014). Fudan-njust at mediaeval 2014: Violent scenes detection using deep neural networks. In *MediaEval*.

DeGroff, C. G., Bhatikar, S., Hertzberg, J., Shandas, R., Valdes-Cruz, L., and Mahajan, R. L. (2001). Artificial neural network–based method of screening heart murmurs in children. *Circulation*, 103(22):2711–2716.

Derbas, N., Thollard, F., Safadi, B., and Quénot, G. (2012). Lig at mediaeval 2012 affect task: use of a generic method. In *MediaEval 2012-Workshop on MediaEval Benchmarking Initiative for Multimedia Evaluation*, volume 927, page 2p. CEUR Workshop Proceedings.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Eyben, F., Weninger, F., Lehment, N. H., Rigoll, G., and Schuller, B. (2012). Violent scenes detection with large, brute-forced acoustic and visual feature sets. In *MediaEval*.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Figueiredo, M., Vicente, L., Vicente, H., and Neves, J. (2014). School dropout screening through artificial neural networks based systems. *Advances in Educational Technologies*, pages 22–27.

Funk, J. B., Baldacci, H. B., Pasold, T., and Baumgardner, J. (2004). Violence exposure in real-life, video games, television, movies, and the internet: is there desensitization? *Journal of adolescence*, 27(1):23–39.

Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., and Theodoridis, S. (2010). *Audio-Visual Fusion for Detecting Violent Scenes in Videos*, pages 91–100. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hasler, D. and Suesstrunk, S. E. (2003). Measuring colorfulness in natural images. In *Electronic Imaging 2003*, pages 87–95. International Society for Optics and Photonics.

Holmes, G., Donkin, A., and Witten, I. H. (1994). Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361. IEEE.

Huesmann, L. R., Moise-Titus, J., Podolski, C.-L., and Eron, L. D. (2003). Longitudinal relations between children's exposure to tv violence and their aggressive and violent behavior in young adulthood: 1977-1992. *Developmental psychology*, 39(2):201.

Jiang, Y.-G., Dai, Q., Tan, C. C., Xue, X., and Ngo, C.-W. (2012). The shanghai-hongkong team at mediaeval2012: Violent scene detection using trajectory-based features. In *MediaEval*.

Ke, Y., Tang, X., and Jing, F. (2006). The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE.

Lam, V., Le, D.-D., Le, S. P., Satoh, S., and Duong, D. A. (2012). Nii, japan at mediaeval 2012 violent scenes detection affect task. In *MediaEval*. Citeseer.

Le Meur, O., Baccino, T., and Roumy, A. (2011). Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 373–382. ACM.

Lin, J. and Wang, W. (2009). Weakly-supervised violence detection in movies with audio and video based co-training. In *Pacific-Rim Conference on Multimedia*, pages 930–935. Springer.

Luo, Y. and Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, pages 386–399. Springer.

Martin, V., Glotin, H., Paris, S., Halkias, X., and Prevot, J.-M. (2012). Violence detection in video by large scale multi-scale local binary pattern dynamics. In *MediaEval*. Citeseer.

Penet, C., Demarty, C.-H., Soleymani, M., Gravier, G., and Gros, P. (2012). Technicolor/inria/imperial college london at the mediaeval 2012 violent scene detection task. In *Working Notes Proceedings of the MediaEval 2012 Workshop*.

Prasad, K., Nigam, D. C., Lakhotiya, A., and Umre, D. (2013). Character recognition using matlab's neural network toolbox. *International Journal of u-and e-Service, Science and Technology*, 6(1):13–20.

Sheehan, P. W. (1997). The effects of watching violence in the media: Policy, consensus, and censorship. In *conference: Violence, Crime and the Entertainment Media, held in Sydney*, pages 4–5. Citeseer.

Wang, H. L. and Cheong, L.-F. (2006). Affective understanding in film. *IEEE Transactions on circuits and systems for video technology*, 16(6):689–704.