Affective Analysis of Videos:
Detecting Emotional Content in Real-Life Scenarios

vorgelegt von
Master of Science
Esra Acar Celik
geb. in Afyonkarahisar

Von der Fakultät IV – Elektrotechnik und Informatik –
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Thomas Wiegand
Berichter:	Prof. Dr. Dr. h.c. Sahin Albayrak
Berichter:	Prof. Dr. Adnan Yazıcı
Berichter:	Dr. Frank Hopfgartner

Tag der wissenschaftlichen Aussprache: 10. Januar 2017

Berlin 2017

to my beloved father

ABSTRACT

As the amount of available multimedia content becomes more and more abundant, the use of automatic multimedia analysis solutions in order to find relevant semantic search results or to identify illegal content present on the World Wide Web has reached a critical importance. In addition, the advances in digital media management techniques have facilitated delivering digital videos to consumers. Therefore, accessing online video content has become extremely easy. As emotions play an important role for multimedia content selection and consumption in peoples' daily life, analyzing the emotional content (i.e., affective content) of videos in order to structure mostly unstructured or ill-structured data is of high value.

This thesis focuses on the high-level research question of discriminative feature representations and modeling methods for affective content (including violence) analysis of videos, while keeping the domain knowledge about the problem or application at hand to a minimum. Towards addressing this research question, analysis frameworks which let the video data itself construct mid-level steps to narrow the "affective gap" between low-level audio visual elements and high-level semantics are presented.

In the first part of the thesis, we first address the issue of feature engineering in the field of video affective content analysis. We present a deep learning architecture to construct audio and static visual higher level representations from raw data instead of handcrafting such higher level representations. Second, a comprehensive analysis of supervised machine learning algorithms for emotional content modeling is performed. Finally, the importance of temporal information for the generation of discriminative analysis models is investigated.

In the second part of the thesis, we concentrate on a special case of video affective content analysis: Violence detection. A comprehensive analysis on the discriminative power of different modalities including audio, static and dynamic visual is performed. A "divide-et-impera" approach is presented to model a complex

concept (namely, violence) present in videos, where kernel-based and deep learning methods are used as base building blocks. In addition, a “coarse-to-fine” analysis setup is introduced to address the time efficiency of the video analysis process.

The effectiveness of the frameworks presented in both parts are discussed with extensive experiments on standard video datasets using official evaluation metrics.

ZUSAMMENFASSUNG

Die Menge an verfügbaren Multimedia-Inhalten wächst ständig. Automatische Multimedia Analyselösungen sind erforderlich, um relevante semantische Suchergebnisse zu finden oder unzulässige Inhalte im World Wide Web zu identifizieren. Darüber hinaus haben Fortschritte im Bereich digitaler Medien-Management-Techniken den Zugriff auf digitale Videos deutlich vereinfacht. Emotionen spielen eine entscheidene Rolle bei Auswahl und Konsum von Multimedia-Inhalten. Dadurch gewinnt die Analyse von emotionalen Inhalten (so genannten affektiven Inhalten) von Videos an Bedeutung für die Strukturierung meist unstrukturierter oder schlecht strukturierter Daten.

Diese Dissertation beschäftigt sich mit der Darstellung diskriminativer Merkmale und Modellierungsmethoden von affektiven Inhalten (einschließlich Gewalt) in Videos. Der Fokus liegt mehr auf der übergeordneten Funktionalität als auf dem konkreten Anwendungsfall. Es werden Frameworks zur Analyse von Videodaten betrachtet, die in der Lage sind, die „affektive Lücke“ („affective gap“ auf Englisch) zu schließen, indem sie übergeordnete semantische Repräsentationen aus einfachen audio-visuellen Daten ableiten.

Der erste Teil der Dissertation befasst sich mit Feature-Engineering im Kontext affektiver Inhaltsanalyse von Videos. Wir stellen eine „Deep Learning“ -Architektur vor, mit deren Hilfe sich automatisch übergeordnete („higher level“ auf Englisch) Darstellungen von Audio und statischem Video aufbauen lassen, ohne dass manuelle Eingaben nötig werden. Danach wird eine umfassende Analyse überwachter maschineller Lernalgorithmen („supervised machine learning“ auf Englisch) zum Modellieren emotionaler Inhalte durchgeführt. Schließlich wird die Bedeutung von zeitlichen Faktoren bei der Erzeugung diskriminativer Analysemodelle untersucht.

Der zweite Teil der Dissertation konzentriert sich auf einen speziellen Fall von affektiver Inhaltsanalyse von Videos: Erkennung gewalttätiger Inhalte. Eine umfassende Analyse der Diskriminierungsfähigkeit verschiedener Modalitäten

einschließlich Audiomodalität, statischer und dynamischer visueller Modalität wird durchgeführt. Eine „teile und herrsche“ („divide-et-impera“) Lösung zur Modellierung eines komplexen Konzepts (nämlich Gewalt) wird präsentiert, die auf Kernelmethoden und „Deep Learning“ beruht. Zusätzlich wird ein „grob zu fein“ („coarse-to-fine“ auf Englisch) Analyseaufbau zur Zeiteffizienzsteigerung im Videoanalyseprozess eingeführt.

Die Wirksamkeit der in beiden Teilen betrachteten Frameworks wird im Zuge der Durchführung umfangreicher Experimente auf anerkannten Multimedia-Datensätzen anhand offizieller Bewertungsmetriken diskutiert.

ACKNOWLEDGMENTS

On the completion of this work I wish to express my appreciation to all the people who have made this dissertation possible.

First of all, I would like to express my gratitude to my thesis supervisor Prof. Dr. Şahin Albayrak for giving me the chance, support and freedom to work on my PhD topic at the DAI Laboratory at TU Berlin. I also take this opportunity to thank Prof. Dr. Adnan Yazıcı and Dr. Frank Hopfgartner, for sitting on my thesis committee and their valuable feedback that helped me improving this dissertation.

Further, I would like to thank my colleagues from the Information Retrieval and Machine Learning research group for their support and for providing a friendly working environment during my PhD studies.

Special thanks are due to Dr. Brijnesh Johannes Jain for fruitful discussions and for the feedback on the publications upon which this PhD dissertation is based, and also for proofreading and providing valuable comments to mature this dissertation.

Another special thanks goes to Benjamin Kille for helping me finalizing the "Zusammenfassung" of this thesis, and also for the interesting discussions on diverse topics during our lunch breaks at the "Mathe Kantine".

I also would like to thank the people – no need to name each of you explicitly, you know who you are :) – from the Badminton group. The evenings spent playing Badminton were very relaxing moments, which kept me away from the daily stress.

Last but not least, I would like to express my deepest gratitude to my dearest family. I am very much indebted to my parents who supported me through all my university studies. I am sure that my beloved father is watching me proudly from somewhere! A very big thanks is for my husband Hasan who was a big support during the PhD with all my ups and downs.

LIST OF PUBLICATIONS

The content of this thesis builds on the following publications by the author:

- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material", Multimedia Tools and Applications (MTAP) journal[1], June, 2016.
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies", Neurocomputing journal[2], October, 2016.
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Fusion of learned multimodal representations and dense trajectories for emotional analysis in videos", IEEE Int. Workshop on Content-Based Multimedia Indexing (CBMI), 2015.
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Detecting violent content in Hollywood movies and user-generated videos", Smart Information Systems, Springer Int. Publishing, 2015.
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Understanding affective content of music videos through learned representations", Int. Conference on MultiMedia Modelling (MMM), 2014.
- Dominique Maniry, **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "A visualization tool for violent scenes detection", ACM Int. Conference on Multimedia Retrieval (ICMR), 2014.

[1] `http://link.springer.com/journal/11042`
[2] `http://www.journals.elsevier.com/neurocomputing/`

- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "TUB-IRML at MediaEval 2014 violent scenes detection task: Violence modeling through feature space partitioning", MediaEval Multimedia Benchmarking, 2014.
- **Esra Acar**, "Learning representations for affective video understanding", ACM Int. Conference on Multimedia (ACMMM), 2013.
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Violence detection in Hollywood movies by the fusion of visual and mid-level audio cues", ACM Int. Conference on Multimedia (ACMMM), 2013.
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Detecting violent content in Hollywood movies by mid-level audio representations", IEEE Int. Workshop on Content-Based Multimedia Indexing (CBMI), 2013.

Other publications of the author describing research carried out outside the scope of this thesis:

- **Esra Acar**, Marco Lützenberger and Marius Schulz, "Intermodal mobility assistance for megacities", Smart Information Systems, Springer International Publishing, 2015.
- Jan Keiser, Nils Masuch, Marco Lützenberger, Dennis Grunewald, Maximilian Kern, Frank Trollmann, **Esra Acar**, Cigdem Avci Salma, Xuan-Thuy Dang, Christian Kuster and Sahin Albayrak, "IMA–an adaptable and dynamic service platform for intermodal mobility assistance", IEEE Int. Conference on Intelligent Transportation Systems (ITSC), 2014.
- David Scott, Zhenxing Zhang, Rami Albatal, Kevin McGuinness, **Esra Acar**, Frank Hopfgartner, Cathal Gurrin, Noel E O'Connor and Alan F Smeaton, "Audio-visual classification video browser", Int. Conference on MultiMedia Modelling (MMM), 2014.
- Dominique Maniry, **Esra Acar** and Sahin Albayrak, "TUB-IRML at MediaEval 2014 visual privacy task: Privacy filtering through blurring and color remapping", MediaEval Multimedia Benchmarking, 2014.
- Dominique Maniry, **Esra Acar** and Sahin Albayrak, "DAI at the MediaEval 2013 visual privacy task: Representing people with foreground edges", MediaEval Multimedia Benchmarking, 2014.
- Serdar Arslan, Adnan Yazıcı, Ahmet Saçan, Ismail H. Toroslu and **Esra Acar**, "Comparison of feature-based and image registration-based retrieval of image data using multidimensional data access methods", Data & Knowledge Engineering, 2013.
- **Esra Acar**, Tobias Senst, Alexander Kuhn, Ivo Keller, Holger Theisel, Sahin Albayrak and Thomas Sikora, "Human action recognition using Lagrangian descriptors", IEEE Int. Workshop on Multimedia Signal Processing (MMSP), 2012.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1

INTRODUCTION

In ever-increasing multimedia sources, emotions play an important role in multimedia content selection and consumption. This thesis focuses on the analysis of emotional content including the violence of videos to narrow the affective gap between low-level audio-visual data and the high-level emotional content of videos. In this introductory chapter, we first provide in Section 1.1 the basic concepts used throughout this thesis. The motivation of the present work is explained in Section 1.2. The focus of the thesis including addressed issues and related research questions is discussed in detail in Section 1.3 and, subsequently, Section 1.4 provides real-life scenarios that are the subject of this thesis. In the final two subsections, the contributions as well as the impact of this thesis are summarized (Section 1.5) and the organization of the manuscript is introduced (Section 1.6).

1.1 Basic Concepts

Multimedia can essentially be defined as content that involves different modalities such as audio, visual information and text. Video is one instance of such multimedia contents and, due to its multi-modal nature, constitutes a complex instance. A video consists of a sequence of frames where each frame can be thought as an individual image. In addition, a video usually contains an audio signal synchronized with the frames and, in some cases, textual data related to its content (e.g.,

metadata, script). One of the biggest research questions in the field of multimedia is how to make a computer (partly) understand the content of a multimedia data item (i.e., an image, audio or video file); the related research field is called *multimedia content analysis* [51]. In other words, this research field studies how to close the so-called "semantic gap" between raw multimedia data and high-level semantics describing its content.

The semantic meaning or content of a video can be described from different perspectives. Two basic levels of video content perception can be differentiated: These are *cognitive-level* and *affective-level* [57]. The aim of the studies that analyze video content at the cognitive-level is to extract information (so called "facts" according to the terminology used in [57]) in terms of objects or people appearing in a video, events or scenes (e.g., someone playing a guitar). The other basic content analysis perspective is to explore the content of videos at the affective-level. The *affective content* of videos can be basically defined as the intensity and type of *affects* (i.e., emotions) that are contained in videos and expected to arise in users while watching the videos [57].

Semantic content analysis approaches – whether at the cognitive or affective level – that are based on machine learning techniques make use of feature representations at different abstraction levels to model multimedia content. The feature representations can be classified according to different schemes. One type is the classification based on the level of semantic information which a given feature carries. In the terminology which we adopt, at one extreme, a feature is said to be "low-level" if it carries (almost) no semantic information (e.g., value of a single pixel or audio sample); at the other extreme, it is said to be "high-level" if it carries maximally semantic information (e.g., a guitarist performing a song in a clip). Between both, "mid-level" feature representations are derived from raw data, but are one step closer to human perception. Another possible type of classification, which is particularly relevant in video analysis, where data items are not a single image but sequences, is the distinction between "static" and "dynamic" (or temporal) features.

1.2 Motivation

The last two decades have witnessed an extraordinary development in the various technologies relating to audio, visual or textual data acquisition, storage, transmission and computing power. Illustration of those developments are ubiquitous; we just have to take a scan around us to realize it, be it at home, in the street, in concert halls, or on squares during various celebrations. In all of these locations or venues, devices equipped with cameras, microphones, gigabytes of storage, wireless connections, and powerful processors are present: PCs, Home Entertainment devices, smart phones or tablets enable the capture of multimedia in general, and sound and videos in particular, which then can be stored, processed, published or transmitted.

Multimedia data can represent innumerable events or scenes. A non-exhaustive list of what multimedia can represent includes:

- Surveillance data, thanks to the presence of CCTV cameras placed in streets or public places;
- Personal recordings or photo collections, obtained through consumer video cameras or portable devices equipped with cameras and microphones;
- Entertainment data items.

Of the above given list, the last item – entertainment data items – plays a particular role in our daily life. Entertainment data items encompass not only professionally edited videos such as concert recordings, shows, movies, documentaries but also user-generated videos which are shared or posted on various platforms such as YouTube[1], Facebook[2], Instagram[3], Vimeo[4] or Dailymotion[5]. Concerning user-generated videos, we witnessed an unprecedented boom as evidenced by statistics. For instance, according to YouTube statistics[6], the website has over 1 billion of users and YouTube CEO Susan Wojcicki announced on the 25th of July'2015 that 400 hours of video are uploaded every minute [22]. We can only expect these figures to grow [104]. Videos uploaded to YouTube can include redundant content. However, the number and, more importantly, the growth over time (Figure 1.1) is quite significant considering that YouTube is merely one of the various online platforms where videos are uploaded.

Concerning professionally edited videos, movies account for a large part of this type of multimedia data. In their habits of viewing movies, consumers have shifted away from the traditional broadcasting means such as conventional television or materialized supports (e.g., tapes, DVDs)[7] to adopt new broadcasting solutions such as video-on-demand (VOD) services which attest an ever growing dematerialization trend. Dematerialization which was popularized by iTunes in the years '90 and '00 – initially for music, but later also for audio-visual content – has recently enjoyed a new growth thanks to services such as Amazon Prime[8], Hulu[9] or Netflix[10]. As an illustration of this growth, the latter features already 74 millions of users and announced its expansion to more than 130 countries[11].

[1] `https://www.youtube.com/`
[2] `https://www.facebook.com/`
[3] `https://www.instagram.com/`
[4] `https://vimeo.com/`
[5] `http://www.dailymotion.com/`
[6] `https://www.youtube.com/yt/press/statistics.html`
[7] `http://www.wsj.com/news/articles/SB10001424052702304887104579306440621142958`
[8] `https://www.amazon.de/prime/`
[9] `http://www.hulu.com/`
[10] `http://www.netflix.com`
[11] `http://www.forbes.com/sites/laurengensler/2016/01/19/netflix-fourth-quarter-earnings`

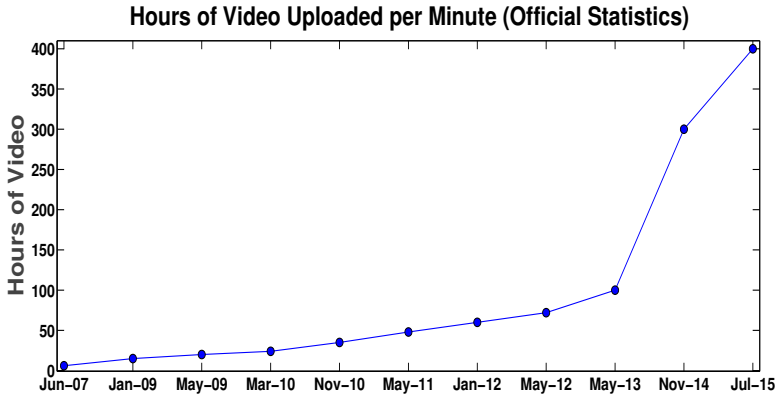


Figure 1.1: Video upload rate (in hours of video per minute) on YouTube from June 2007 to July 2015.

User-generated and professionally edited videos share a number of common features:

(a) Their number is enormous, and continuously growing (e.g., Figure 1.1);

(b) They are mostly unstructured or ill-structured, i.e., they are organized according to poor schemes – e.g., for user-generated videos, organization and retrieval are based on file names or user-generated tags, or comments on sites hosting the videos (like YouTube). User-generated tags are useful information sources. However, they might be problematic in the sense that the tagging intent of users cannot be always used to index the multimedia content (i.e., the tags are not related to the actual content of multimedia items, but correspond more to reputations or judgments [98]). For movies, on the other hand, structuring is based on genres (which is a rough classification scheme), actors playing in the movie, directors, or summaries which are generally performed manually by professionals;

(c) Both types of videos (i.e., user-generated and professionally edited) contain emotions, or are meant to provoke emotions by spectators. For instance, in user-generated videos which get a high number of hits, scenes with strong emotions, which can be considered as "funny" or "violent" are often present. A similar consideration can be derived for movies, because they are typically designed to convey emotions (e.g., romance, comedy, thriller).

In view of the properties **(a)** and **(b)**, there appears a need for a better – finer – organization scheme. In view of **(c)**, organizations based on analyzing emotions

more likely evoked in users (i.e., the *affective content* of videos) can provide a scalable and adaptable solution as an alternative to cognitive-level video analysis where an extensive research is going on, since emotions play an important role in multimedia content selection and consumption.

Emotional analysis is basically the subject of *affective computing* introduced in 1995 by Rosalind Picard [101]. Affective computing refers to emotional intelligence of technical systems in general; yet so far, research in this domain has mostly been focusing on aspects of human-machine interaction [49]. However, there are fewer studies that address the analysis of affective information contained in the audio-visual content itself. In this thesis, we therefore undertake an **analysis of emotions** to perform different tasks relating to the organization of user-generated videos and movies.

1.3 Focus of the Thesis

As said in Section 1.2, our ultimate goal is to develop solutions for affective content analysis of videos, more specifically professionally edited or user-generated videos. Throughout the thesis, we address this goal at different levels. In the first part of the thesis (Chapters 3 and 4), we approach the affective content analysis as understood in its broadest sense, i.e., the **analysis of emotions in general**.

Next to this general analysis, we dedicate the second part of the thesis (Chapters 5 and 6) to the **concept of violence** as a special case of affective content analysis, since violence plays an important part in movies or in user-generated videos. For instance, several movie genres involve violence as one of the dominant concepts present in a movie – horror, action, adventure to name a few. Violence is usually used to evoke strong emotions by the user. In addition, due to the potential harmful effects of violence among sensible audiences, spectators may have the desire to avoid violent contents. The importance of violence justifies, therefore, that we specifically dedicate a part of this thesis to describe algorithms for analyzing violent contents.

Achieving automatic emotion analysis or violence detection in professionally edited or user-generated videos requires the analysis of multiple modalities, typically audio and visual. This observation implies **(1)** multimedia data to be properly represented by means of appropriate audio-visual features, so that **(2)** automatic analysis can be performed thanks to inferring techniques such as pattern recognition and machine learning.

The literature survey considering the two points (1) and (2) on the affective content analysis of videos reveal the following issues which are worth investigating in this thesis. The first issue is related to feature representation. The most of approaches in the field of affective content analysis either use general-purpose audio-visual features or construct domain-specific and problem-dependent representations. The general-purpose features are easy to construct and generic, whereas the problem-dependent ones are time-consuming and generally require domain

knowledge. The second issue is related to model generation. When we analyze the approaches in the field, most of them concentrate on the feature engineering part of the analysis and use commonplace approaches such as SVM modeling for its modeling part. Hence, this thesis focuses on these two issues. The ultimate research question of this thesis through the aforementioned two parts concerns building discriminative feature representations and modeling methods for affective content (including violence) analysis of videos, while keeping the domain knowledge to a minimum and using the data itself during the construction of analysis models. Consequently, each part of this work contains research elements which tackle the problem of feature representation or the problem of classification.

1.4 Targeted Real-life Scenarios

Systems analyzing audio-visual signals in terms of their affective content can be used in many applications including summarization, indexing and retrieval in large archives. It enables, for instance, to choose "sad" moments in a drama when generating the highlights of the multimedia content. Another application scenario is the automatic affective tagging of audio-visual contents which can be used to structure multimedia items in a cold start scenario where no information (e.g., preferences) about users is available. The framework we present in the first part of the thesis (Chapters 3 and 4) can typically be applied on these real-life scenarios.

Another important application scenario is related to youth protection services. Nowadays, children are submerged by connected equipment, whether this is at school, at home or even in the car. Notable examples of such equipment include TV, cable or satellite set-top boxes, tablets or the smartphones of the parents, when those let their children play with it. Parents can always try to track these "undesired" contents (e.g., violent content). However, this tracking might not be constantly possible. In these situations, automatically detecting and filtering violent content would show a considerable utility. In the second part of the thesis (Chapters 5 and 6), we address this real-life scenario with the presented framework.

1.5 Contributions of the Thesis

The main contributions of the thesis as well as the related publications are given in this section.

- *Audio and static visual affective feature learning.* The majority of existing affective content analysis methods (discussed in Chapter 2) use handcrafted low and/or mid-level audio-visual representations. Different from these approaches we apply a deep learning approach to learn audio and static visual features to represent videos for affective content analysis. More specifically, we use a convolutional neural network (CNN) architecture to learn higher

level acoustic patterns from the audio signal of videos using an acoustic spectral descriptor-time domain representation as 2D raw data. Next to this, we apply the CNN architecture on keyframes of videos to extract static visual patterns from raw visual data. The commonplace approach of using CNNs on images is to apply them on the RGB color space. In this study, we extend this approach and work in the HSV color space in addition to the RGB color space and evaluate the learned static visual representations within the context of affective video content analysis. The findings of the feature learning provide valuable insights for the field of affective content analysis of videos, as we experimentally show that superior classification performance can be achieved while keeping the domain knowledge to a minimum and let data define patterns for content analysis. This enables to concentrate more on the modeling part than the feature engineering part of the analysis. The details of representation learning are discussed in Chapter 3. Part of the work presented in Chapter 3, and work pertaining to this topic have been presented in:

  - **Esra Acar**, "Learning Representations for Affective Video Understanding", ACM International Conference on Multimedia (ACMMM), 2013 [1].
  - **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Understanding affective content of music videos through learned representations", International Conference on MultiMedia Modelling (MMM), 2014 [5].

- *Ensemble learning with enriched multi-modal cues for affective modeling.* The contributions for this part of the study (Chapter 4) are twofold. First, we distance ourselves from the prior art – discussed in detail in Chapter 2 – and generate analysis models using ensemble learning (decision tree based bootstrap aggregating to be more specific). This learning method is shown to further enhance the classification accuracy of the system compared to SVM modeling which is the commonplace approach. Second, we extend our audio-visual feature set with motion and domain-specific representations whose applicability, to the best of our knowledge, have not been investigated yet. The use of motion-based features within the context of affective content analysis is limited – as discussed later in Chapter 2 – to simple motion features (such as frame differencing). We apply dense trajectory features to derive higher level representations via sparse coding to boost the classification performance of the proposed system. In addition, we explore optimal late fusion mechanisms, since we are in the presence of multi-modal features to represent videos. These contributions are important, as it is experimentally shown that ensemble learning based analysis outperforms the SVM based one – which is the dominant modeling approach – on both professionally edited and user-generated videos. Besides, the importance of using advanced motion features such as dense trajectories for affective content anal-

ysis is presented through extensive experiments. Part of the work presented in Chapter 4, and work related to this matter have been presented in:

  - **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos", IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), 2015 [6].
  - **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material", Multimedia Tools and Applications (MTAP) journal[12], June, 2016 [8].

- *A comprehensive study of the correlation between audio-visual features and the concept of "violence".* Within the context of violence detection in videos, we perform an extensive analysis of audio-visual features at different levels. More specifically, we evaluate the discriminative power of vector quantization and sparse coding based mid-level audio representations and the extent to which the fusion of these representations with static and dynamic visual representations further improves the violence analysis in videos. The findings related to this contribution provide valuable insights for violent content analysis of videos. It is experimentally shown that videos need to be described from different perspectives such as audio, static and dynamic visual features to achieve state-of-the-art performance. This is also an indication of the diverse nature of "violence" concept. Part of the work presented in Chapter 5, as well as other related work on this topic have been presented in:

  - **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Violence detection in Hollywood movies by the fusion of visual and mid-level audio cues", ACM International Conference on Multimedia (ACMMM), 2013 [4].
  - **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Detecting violent content in Hollywood movies by mid-level audio representations", IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), 2013 [3].
  - Dominique Maniry, **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "A visualization tool for violent scenes detection", ACM International Conference on Multimedia Retrieval (ICMR), 2014 [86].

- *Feature space partitioning and coarse-to-fine analysis for violence detection.* The contributions for this part of the study (Chapter 6) are twofold. First, our work differs from the prior art (discussed in detail in Chapter 2) by modeling the concept of "violence" as a plurality of subconcepts (i.e., *feature space*

[12] http://link.springer.com/journal/11042

partitioning); this stems from our observation that "violence" is usually expressed under diverse forms, rather than as a single type. Second, we introduce a *coarse-to-fine analysis* analysis strategy for violence detection in videos in order to cope with the heavy computations resulting from the advanced dynamic and static visual representations introduced to ameliorate the performance. The feature space partitioning approach eliminates the need for time-consuming manually designed violence-related concepts (e.g., fight, explosion) and lets data define subconcepts in terms of audio-visual characteristics. As we use no subconcept hardwired to "violence", the approach can be extended to other complex concepts. The coarse-to-fine analysis enables computational efficiency without sacrificing too much from performance. Besides, it provides a scalable solution adjustable depending on the processing power or accuracy requirements. Part of the work presented in Chapter 6 and related work in this topic have been presented in:

- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "TUB-IRML at MediaEval 2014 Violent Scenes Detection Task: Violence Modeling through Feature Space Partitioning", MediaEval Benchmarking, 2014 [2].
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Detecting Violent Content in Hollywood Movies and User-Generated Videos", Smart Information Systems, Springer International Publishing, 2015 [9].
- **Esra Acar**, Frank Hopfgartner and Sahin Albayrak, "Breaking Down Violence Detection: Combining Divide-et-Impera and Coarse-to-Fine Strategies", Neurocomputing journal[13], October, 2016 [7].

1.6 Organization of the Thesis

This thesis is organized as follows. In Chapter 2, the background on emotion representation models, affective content analysis and violence detection is presented. In Chapter 3, audio and static visual feature learning using deep learning architectures and the evaluation of these learned representations on a video dataset consisting of music video clips from different genres is introduced. Chapter 4 extends the framework presented in Chapter 3 and concentrates primarily on the modeling part of affective content analysis in videos. Also, the evaluation part is extended with a more challenging dataset which contains user-generated video clips from YouTube and Flickr[14]. The next chapter (Chapter 5) addresses the detection of "violence" in videos and presents a comprehensive evaluation on audio-visual representations for violent content analysis in videos. The subsequent chapter (Chapter 6) deepens the modeling aspects of violence detection and introduces feature space partitioning and coarse-to-fine analysis for more discriminative model generation and time efficiency, respectively. Finally, Chapter 7 concludes the thesis

[13] http://www.journals.elsevier.com/neurocomputing/
[14] https://www.flickr.com/

and gives future research directions within the context of affective content (including “violence” concept) analysis.

2

BACKGROUND

This chapter introduces the basic concepts of affective analysis used throughout this thesis. In addition, a literature review on affective content analysis and on the special case of violence detection in videos is provided. We start by defining the term *emotion* in Section 2.1 and by providing emotion representation models that are commonly used in affective analysis in Section 2.2. Subsequently, Section 2.3 discusses a review of existing works related to affective content analysis. In Section 2.4, the concept of violence is defined, then uni-modal and multi-modal violent content analysis approaches, and the MediaEval[1] benchmarking task initiated for violence analysis in movies are discussed. Finally, Section 2.5 concludes this chapter with a short summary.

2.1 Definition of an Emotion

In [108], Scherer defines an *emotion* as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism". It is important here to remark that an emotion is different from a feeling or mood which are two other affective phenomena. *Feeling* is the

[1] `http://www.multimediaeval.org/`

term used for the subjective experience component of emotion and is not a synonym for emotion [108]. Similarly, mood is also not synonymous with emotion. *Mood* can be defined as a diffuse affective state characterized by a long-lasting predominance of certain types of feelings that affect the experience and behavior of a person (e.g., cheerful, depressed) [108]. The main difference between mood and emotion is that mood is a long-lasting and low-intensity affective state compared to emotion and is not related to a specific stimulus, whereas an emotion is a relatively short-term state as a reaction to an internal or external stimulus and is usually of higher intensities [108].

Besides the definition of an emotion and the characteristics which discriminate it from feeling and mood, an important discussion concerns the different types of emotion: On the one hand, **felt or actual** emotions; on the other hand, **expected** emotions. Within the context of affective video content analysis, an actual emotion can be defined as the emotion that is evoked in a user while watching a video. Hence, it is personal (i.e., subjective) and context-dependent [57]. In opposition, an expected emotion is the one that is either intended to be communicated toward an audience by a film director or that is likely to be elicited from the majority of an audience who is watching a particular video [57]. For instance, "disturbing" scenes in a horror movie are edited in a way by film makers to elicit particular emotions in the audience. "Fear" is usually the expected emotion in this example. Although this might be the case for most of the people watching these scenes, there are people who find this kind of scenes "funny" (i.e., actual emotion). In this thesis, we follow the latter definition of emotion (i.e., **expected emotion**) within the context of affective content analysis of videos and use the audio-visual content that is being watched by people to infer the expected emotion. Trying to directly understand the audio-visual or multimedia content has the advantage of being non-invasive, which translates into a broader applicability. Besides, the expected affective response to a specific stimulus can be mapped to a user-specific affective response to that stimulus using the profile of a particular user [57].

2.2 Emotion Representation Models

One of the first steps in affective multimedia content analysis is to decide how to represent emotions. Despite the existence of various other models, *categorical* and *dimensional* approaches are the most commonly used ones for automatic affective analysis [56].

In the categorical representation of emotions, there are predefined discrete emotion categories such as anger, fear, sadness, joy, to name a few. Different lists of emotion categories have been proposed in the literature. These categories can be regarded as fixed or graded (e.g., weak, medium, strong), or as pure or mixed, or sometimes even antagonistic (e.g., a mixture of anger, joy and irony) [56]. As an example, we present in Figure 2.1 Plutchik's emotion wheel [102] which defines a wheel-like diagram of emotions consisting of eight basic (i.e., primary) emotions

and their derivatives, and illustrates various relations among these emotions. The eight basic emotions of the Plutchik's wheel are *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger* and *anticipation*. The wheel is coordinated in pairs of opposite emotions (e.g., *joy* versus *sadness*, *anger* versus *fear*). The intensity of an emotion increases toward the center of the wheel and is indicated with an increased color intensity as shown in Figure 2.1. A final remark on Plutchik's wheel is about secondary emotions which are illustrated between the primary emotion leaves in Figure 2.1 with no color. These emotions are a mixture of two neighboring primary emotions. For instance, *love* is a combination of *joy* and *trust*, whereas *anticipation* and *joy* combine into *optimism*. The secondary emotions of Plutchik's wheel are *love*, *submission*, *awe*, *disapproval*, *remorse*, *contempt*, *aggressiveness* and *optimism*.

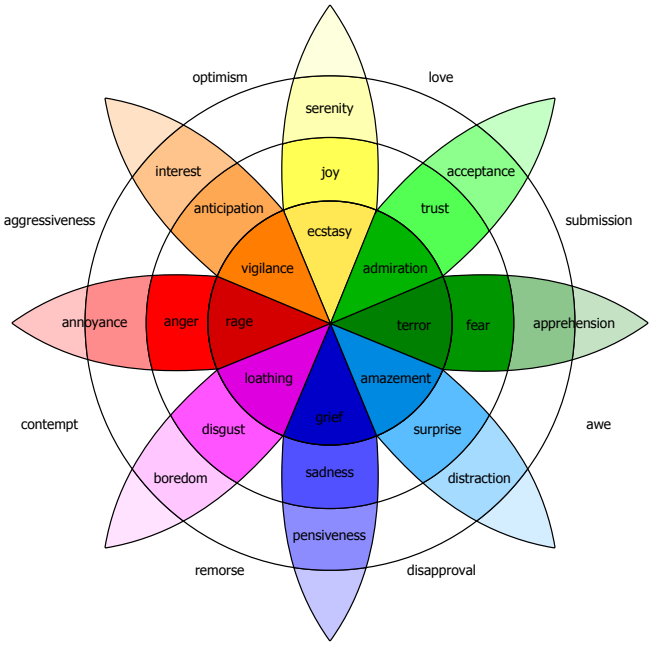


Figure 2.1: Plutchik's wheel of emotions [102] used for the representation of emotions of user-generated videos.

In the dimensional representation of emotions, affective analysis is performed along affect dimensions instead of using predefined emotion categories (i.e., emotions are represented as points in an n-dimensional space, where n is the number of affect dimensions being used in the modeling of an emotion). Traditionally, emotions are modeled by means of the arousal and valence dimensions [56]. There

are studies that include an additional dimension called the control/dominance dimension (e.g., [145]). Arousal can be defined as the *intensity* of an emotion, while valence is the *type* of an emotion. Valence is typically characterized as a continuous range of affective states extending from "pleasant/positive" to "unpleasant/negative", while arousal is characterized by affective states ranging on a continuous scale from "energized/excited" and "calm/peaceful" [57]. The third dimension (i.e., control) is used to differentiate between emotional states having similar arousal and valence (e.g., differentiating between "grief" and "rage") and usually ranges from "no control" to "full control" [57]. An example of dimensional representation of emotions is Russell's circumplex model [105] of emotions, presented in Figure 2.2.

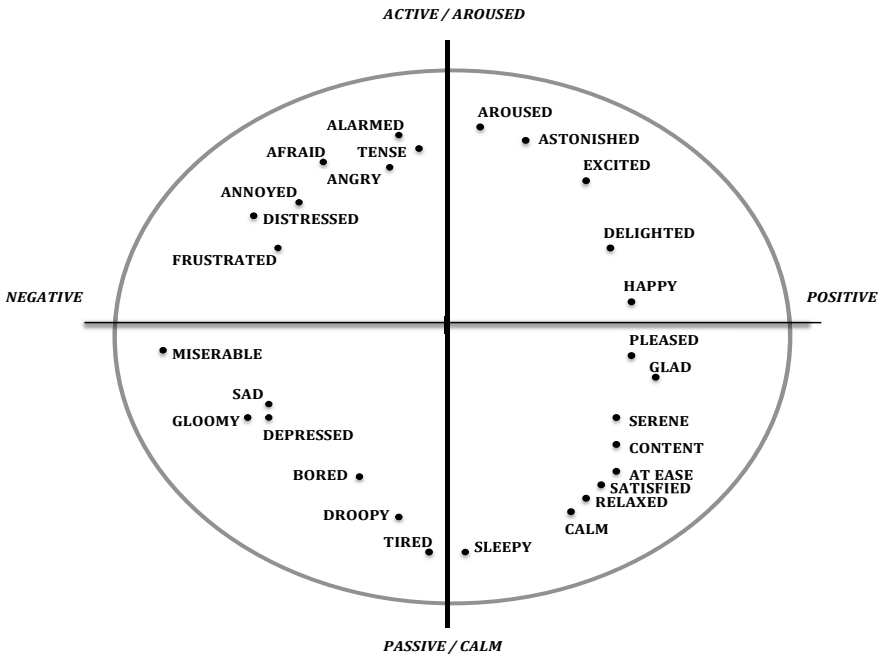


Figure 2.2: The emotion model based on Russell's circumplex model of emotions.

As discussed in [56], the choice of a categorical or dimensional analysis is not critical, as in practice, categories can always be mapped onto dimensions and vice versa. For example, Figure 2.2 illustrates Russell's circumplex model of emotions which are distributed in the Valence-Arousal space (VA-space). This implies that emotion categories can also be placed on the VA-space. Another example would be defining the four quadrants of the VA-space as four emotion categories and perform categorical emotion analysis of videos as in [145].

2.3 Affective Content Analysis: Literature Review

The recent survey by Wang and Ji [129] provides a thorough overview of the subject and distinguishes two kinds of research. On the one hand, *direct* approaches (mainstream) deduce the emotional content of videos directly from audio and/or visual features extracted from the video data; on the other hand, *implicit* ones deduce them from the user's reaction while being exposed to the video [129]. The literature on emotional content analysis – and in particular on direct methods – can be analyzed from different points of view. As also evoked in the survey by Wang and Ji [129], a direct affective content analysis framework requires two essential elements; these are video feature extraction (i.e., *representation*) and classification or regression (i.e., *modeling*). As, in this work, we focus both on the representation and on the modeling aspects of emotional content analysis, we present a review of existing solutions examined from these two perspectives in Sections 2.3.1 and 2.3.2, respectively. In addition, stressing the importance of motion related features enables us to highlight one of our contributions; hence, in Section 2.3.3, we also position existing studies with respect to the use of motion information.

2.3.1 From a feature representation point of view

Among video affective content analysis methods, using low-level audio-visual features as video representations is one type of commonplace approach. In [47], Eggink and Bland present a method for mood-based classification of TV Programs on a large-scale dataset. Various low-level features are used as video representations; these are audio (e.g., MFCC, sound energy, spectral rolloff, zero crossing rate) and visual features (luminance, motion, cuts and presence of faces), either taken individually or in combination. SVMs are used as classifiers. In [130], a combined analysis of low-level audio and visual representations based on early feature fusion is presented for facial emotion recognition in videos. The audio features include MFCC and zero crossing rate (ZCR). The visual features are extracted based on a deformable model fitted on faces and correspond to various face-related information such as facial pose, opening of the mouth or status of eyebrows. The combined feature vectors are classified with SVMs. In order to retrieve videos containing resembling emotions, Niu et al. [96] develop a similarity measure of videos based on affect analysis. They use four low-level features (audio: sound energy, audio pitch average; visual: motion, shot-change rate) to construct Valence-Arousal (VA) graphs. Those VA-graphs are normalized to account for unequal video durations and are used to derive the similarity measures. The set of videos is further partitioned using spectral clustering based on the similarity measure. The result is used in a recommender system to retrieve similar contents. Baveye et al. [15, 17] address the issue of a common emotional database and introduce the LIRIS-ACCEDE dataset which is an annotated creative commons emotional database. In that work, a baseline framework was also presented, which employs low-level audio (energy, ZCR, etc.) and still image (colorfulness, spatial edge distribution, etc.) features,

in an SVM framework. Yazdani et al. [145] present a method which uses low-level audio-visual features as representation and the k-nearest neighbor classifier as the learning method for the affective analysis of music video clips.

The major drawback of the approaches of this category – which mainly use low-level video representations for emotional content analysis – is the risk of losing the global structure in the video data as well as having a minimal semantic meaning compared to higher level representations which we discuss in the following paragraphs.

Another type of commonplace approach is to use mid-level or hierarchical representations of videos to narrow the affective gap between the raw video data and its emotional content by means of these higher level video representations. Irie et al. [63] present an affective video segment retrieval method based on the correlation between emotions and so-called emotional audio events (EAEs) which are *laughter*, *loud voice*, *calm music* and *aggressive music*. The main idea is to use EAEs as an intermediate representation. The detection of EAEs is based on the audio information only. Xu et al. [138] present a 3-level affective content analysis framework, in which the purpose is to detect the affective content of videos (i.e., horror scenes for horror movies, laughable sections for sitcoms and emotional tagging of movies). They introduce mid-level representations which indicate dialog, audio emotional events (i.e., horror sound and laughter – similar in concept to the EAEs of Irie et al. [63]) and textual concepts (i.e., informative emotion keywords). In [64], Irie et al. propose to represent movie shots with so-called Bag-of-Affective Audio-visual Words and apply a latent topic driving model in order to map these representations to affective categories, and, hence, to achieve movie classification. The audio-visual words are formed by bringing together audio and visual words, which are constructed based on audio (pitch, short-term energy, MFCC) and visual (color, brightness, motion intensity and shot duration) features, respectively. In [24], Canini et al. introduce a framework where movie scenes are represented in a 3-dimensional connotative space whose dimensions are natural, temporal, and energetic. The aim is to reduce the gap between objective low-level audio-visual features and highly subjective emotions through connotation. As audio-visual representation, they employ low-level audio descriptors (representing intensity, timbre and rhythm), low and mid-level color (e.g., color energy, average saturation of pixels) and motion (average of motion vector modules and their standard deviation) descriptors. Ellis et al. [48] introduced a framework for emotional analysis of movie trailers using mid-level representations corresponding to specific concepts (e.g., "gunshot", "fight"). The rationale behind their method is that human emotions are closely related to such concepts rather than low-level features. They define 36 concepts (annotated at the video shot level) and build a detector for each concept. Each concept detector is realized with an SVM using low-level audio (e.g., MFCC, pitch) and visual (e.g., SIFT, number of faces) features. The concepts are used to infer emotions in the videos. In an attempt to close the so-called emotional gap between low-level features and emotions, Borth et al. [20] construct a visual

sentiment ontology containing 3,000 concepts (i.e., mid-level visual representations), each corresponding to an Adjective-Noun pair (ANP). Each pair is composed of a neutral noun associated to a strong sentiment (e.g., "beautiful flower"). Those concepts are detected using a pool of detectors (SentiBank). Visual features such as color histograms and local binary patterns as well as additional features (e.g., faces or objects) are employed in SVM detectors. The approach was evaluated on images and videos retrieved from online repositories in order to infer emotions. In an improvement to the work by Borth et al. [20], Chen et al. [31] aim at solving two problems related to ANP concepts, namely the localization of objects and the ambiguity of annotations due to adjectives. They solve the localization issue by limiting their study to only six common objects appearing in social media content and by detecting those objects with a parts-based detector. The ambiguity issue between semantically similar concepts is tackled by using a new type of SVM, which is capable of learning overlapping class labels. The method is reported to outperform conventional SentiBank. In [70], Jiang et al. propose a comprehensive computational framework, where they extract an extensive set of features from the dataset, ranging from well-known low-level audio-visual descriptors (audio: MFCC, energy entropy, etc.; visual: SIFT, HOG, etc.) to high-level semantic attributes such as ObjectBank and SentiBank representations.

Handcrafted higher level features (e.g., ANP in [20] or EAE in [63]) provide representations which are closer to human perception. However, they are mainly time-consuming in design, problem-dependent and require domain-knowledge. Hierarchical approaches (e.g., [138]) may in addition have the disadvantage of suffering from cascaded classification errors introduced by an additional classification layer.

All of the above-mentioned works represent videos with low or mid-level handcrafted features. However, in attempts to extend the applicability of methods, there is a growing interest for directly and automatically learning features from raw audio-visual data rather than representing them based on manually designed features, deep learning being a widespread illustration of such direct and automatic learning. For example, Schmidt et al. [109] address the feature representation issue for automatic detection of emotions in music by employing regression based deep belief networks to learn features from magnitude spectra instead of manually designing feature representations. By taking into account the dynamic nature of music, they also investigate the effect of combining multiple timescales of aggregated magnitude spectra as a basis for feature learning. These learned features are then evaluated in the context of multiple linear regression. Among deep learning solutions, Convolutional Neural Networks (CNNs) have become particularly popular in video content analysis in the last years. Li et al. [79] propose to perform feature learning for music genre classification and use CNNs for the extraction of musical patterns directly from audio features. Ji et al. [68] address the automated recognition of human actions in surveillance videos and develop a novel 3D-CNN model to capture motion information encoded in multiple adjacent frames. As an improvement of the SentiBank paper by Borth et al. [20], Chen et al. [30] propose to ex-

tend it using deep learning (*DeepSentiBank*). CNNs are used instead of binary SVM classifiers. The increased computational load induced by deep learning is dealt by a GPU-based learning framework. CNNs are shown to provide substantial improvement compared to binary SVMs. Another work making use of DeepSentiBank is the one by Dumoulin et al. [45], where DeepSentiBank is combined with low-level audio-visual and CNN-based features in a hierarchical classification scheme. In [134], Xu et al. predict sentiments in still images using CNNs. CNNs are trained as object classifiers to classify image content according to labels such as "zebra" or "lemon"; subsequently, transfer learning is employed for generating mid-level representations. Logistic regression is then used to predict sentiments using the generated mid-level representations. More recently, Baveye et al. [16] compared CNNs applied on video keyframes against the combination of Support Vector Regression (SVR) and low-level features, and reached the conclusion that CNNs constitute a promising solution. Xu et al. [133] perform emotion recognition in videos using a BoW approach, where the dictionary is constructed by clustering features obtained from so-called auxiliary images by means of CNNs. The same CNNs are then applied on video frames to encode videos according to the BoW scheme. Unlike most studies where features are independently extracted from audio and visual modalities, Pang and Ngo [97] propose to extract joint features from the raw data directly using Deep Boltzman Machines (DBM), i.e., they extract mid-level features which simultaneously capture audio, visual and text information. Visual, audio and text features are input to a multi-modal DBM which returns the joint representations. Those joint representations are used to train SVM classifiers with RBF kernels in order to predict emotions.

As discussed above, deep learning approaches are proven to be promising in affective content analysis, while having the advantage of learning features from raw data. Different from the aforementioned deep-learning-based works, we apply 2D CNN modeling not only on visual data but also on audio data to learn features automatically and also investigate the performance of different color spaces in the visual feature learning part of our affective content analysis system.

2.3.2 From a modeling point of view

A closer look at the methods proposed in the literature (Section 2.3.1) reveals that most of them (a non-exhaustive list of examples including [20, 47, 70, 130, 138]) concentrate on the representation aspect of content analysis by proposing new handcrafted or learned (via deep learning methods) audio-visual descriptors. The modeling part, on the other hand, is rather neglected; we see, indeed, that authors predominantly use SVMs as the learning method in an "Out-of-the-Box" fashion.

Although the SVM-based learning technique is the dominant modeling scheme for video affective content analysis, certain works in the literature employ other techniques. In [20], Borth et al. use logistic regression in addition to linear SVM and show that logistic regression is superior to SVM. Inspired by its success in [20],

Xu et al. also adopt logistic regression as the learning scheme in their work [134]. Dumoulin et al. [45] introduced a hierarchical classification scheme, where input features are classified by traversing a 3-level tree. At the highest level, the features are determined as stemming from an emotional or non-emotional video segment; then, those classified as emotional are labeled as corresponding to positive or negative emotions; finally, at the lowest level, the class of emotion is determined. This hierarchical scheme was tested using different types of classifiers (e.g., SVM, random forests).

As already presented in detail in Section 2.3.1, alternative learning methods utilized in the field of affective content analysis include topic extraction via the latent Dirichlet allocation [63], spectral clustering [96], *k*-nearest neighbor classifier [145] and SVR [16].

2.3.3 The issue of motion information

One observation about the works mentioned in Section 2.3.1 is that the use of the temporal aspect of videos is either limited or totally absent. In other words, videos are generally analyzed as a sequence of independent frames rather than as a whole. A few works (e.g., [24, 47, 64]) use motion-based features, and these are limited to simple features (e.g., features based on frame differencing). For instance, Canini et al. [24] use average motion vector magnitudes derived from the motionDS descriptor [67]. Eggink and Bland [47] use motion based on the difference between every tenth frame. Irie et al. [64] employ motion intensity as the average magnitude of motion vectors; they also take into account the duration of shots as another feature.

The only notable exception is the work of Ji et al. [68], where multiple adjacent frames are used. However, they take into account only 7 adjacent frames. Increasing this number to higher dimensions would probably render the learning of the 3D-CNNs intractable. Therefore, in our opinion, a more effective mid-level motion representation is needed.

2.4 Violent Content Detection

An analysis of existing works relating to the detection of the concept of violence in videos brings to light two dichotomies from two different perspectives: (1) One from the representation perspective, namely the distinction between uni-modal approaches and multi-modal approaches, and (2) one from the modeling perspective, i.e., SVM-based modeling against non-SVM-based modeling; it also shows the lack of a proper, commonly adopted definition of violence. Hence, in a preliminary part (Section 2.4.1), we tackle the issue of violence definition. Then in Sections 2.4.2 and 2.4.3, we review violence detection approaches from the representation and modeling perspectives, respectively. Finally, Section 2.4.4 presents a synopsis of the significant studies demonstrated at the MediaEval multimedia benchmarking

task addressing the issue of violent content detection in videos by using a standard violence definition and dataset.

2.4.1 Defining violence

In the following two subsections, we first discuss the issues relating to the definition of violence in the literature (Section 2.4.1.1), and then how violence is defined within the related task of MediaEval multimedia benchmarking (Section 2.4.1.2).

2.4.1.1 The difficulty in defining violence

In spite of the existence of institutions, the task of which is to assign a recommended age to movies in France, the ratings determined by those institutions are not as strict and differentiated as the ones from the German FSK (*Freiwillige Selbstkontrolle der Filmwirtschaft*). One explanation of the sources of discrepancies is the fact that employees from the film making industry are allowed to participate in the recommendation process in France.

A lot of movies labeled as FSK 16 (i.e., recommended for an audience of age higher than 16 years old) in Germany are released without restrictions in France[2]. There also is no equivalence for the FSK 6 label in France, where movies are recommended only for the age of 0, 12, 16 and 18[3]. Another illustration of differences between countries is the age rating labels used in the USA, where one example of label is "NC-17". The latter does not mean that audience should be at least 17 years old, but that audience should not be 17 or under 17, while, for instance, FSK 16 means audience should be at least 16. This can also be a source of confusion among consumers.

In addition, the degree of violence one is able or willing to bear might vary strongly even within a group of persons of identical age. That is probably why parents should get from video producers information which is not limited to rating only but also a preview of the most violent scenes, in order to help them decide if the movie is adequate to be watched by their child.

Although video content analysis has been studied extensively in the literature, violence analysis of movies or of user-generated videos is restricted to a few studies only. Consequently, a first difficulty stems from the fact that the research community had limited possibilities to properly define violence. We discuss some of the most representative ones which use audio and/or visual cues in Sections 2.4.2.1 and 2.4.2.2. In these studies, other difficulties arise regarding the definition of violence. For instance, in the publications [29] and [80], violent scenes are defined as those scenes that contain action and blood, whereas in [36] it is defined as sudden changes in motion in a video footage. In some of the works presented in Sections 2.4.2.1 and 2.4.2.2, the authors do not even explicitly state their definition of

[2] `http://fsf.de/jugendmedienschutz/international/filmfreigaben/`
[3] `http://www.fsk.de/`

violence. In addition, nearly all papers in which the concept is defined consider a different definition of violence; therefore, whenever possible, we also specify the definition adopted in each work discussed.

2.4.1.2 Towards a common definition of violence: The definition provided by the MediaEval multimedia benchmarking

The MediaEval multimedia benchmarking Violent Scenes Detection (VSD) task is a yearly venue which allows participating teams to test their algorithms and to position themselves in terms of performance with respect to other teams. As MediaEval provides not only training and test datasets but also a definition of "violence", participants have the possibility to evaluate their algorithms using standard data and definitions. In this thesis, we use the violence definition provided by the MediaEval multimedia benchmarking VSD task. This enables us to evaluate our framework against others using the same definition and standard datasets provided by the task organizers. The violence definition in the task evolved over the years. The very first adopted definition of violence was *objective violence*, defined as "physical violence or accident resulting in human injury or pain" [37]. This definition was used in the VSD task in the first two editions of MediaEval (2011 and 2012). According to this objective definition, actions such as self injuries, or other moderate actions such as an actor pushing or hitting slightly another actor are considered violent.

In 2013, the need of defining violence closer to human perception (a lesson-learned from the tasks of 2011 and 2012), incited the task organizers to introduce the *subjective* definition of violence, in addition to the objective definition. According to this definition, violent scenes are the ones that "one would not let an 8 years old child see in a movie because they contain physical violence" [37].

The last change occurred in 2014. Since then, only the subjective definition of violence is considered for the task (2014 and 2015 editions).

2.4.2 Literature review: Feature representation perspective

In this section, we present a review of existing solutions from the representation perspective. First, we focus on works that consider a single modality for violence analysis in Section 2.4.2.1. Second, we review multi-modal violence analysis approaches in Section 2.4.2.2.

2.4.2.1 Uni-modal violence detection approaches

Uni-modal approaches are those that are based exclusively on one modality, either audio or visual, extracted from the video stream in order to detect violence. Giannakopoulos et al. [52] define violent scenes as those containing shots, explosions, fights and screams, whereas non-violent content corresponds to audio segments containing music and speech. Frame-level audio features both from the

time and the frequency domain such as energy entropy, short time energy, ZCR, spectral flux and rolloff are employed. The main issue with this work is that audio signals are assumed to have already been segmented into semantically meaningful non-overlapping pieces (i.e., shots, explosions, fights, screams, music and speech).

In their paper [36], de Souza et al., similarly to other works related to violence detection, adopt their own definition of violence, and designate violent scenes as those containing fights (i.e., aggressive human actions), regardless of the context and the number of people involved. Their approach is based on the use of Bag-of-Words (BoW), where local Spatial-Temporal Interest Point Features (STIP) are used as the feature representation of video shots. They compare the performance of STIP-based BoW with SIFT-based BoW on their own dataset, which contains 400 videos (200 violent and 200 non-violent videos). The STIP-based BoW solution has proven to be superior to the SIFT-based one. The issue with this work is the assumption that violent and non-violent video samples are perfectly balanced, i.e., equally represented in the dataset.

Hassner et al. [59] present a method for real-time detection of breaking violence in crowded scenes. They define violence as sudden changes in motion in a video footage. The method considers statistics of magnitude changes of flow-vectors over time. These statistics, collected for short frame sequences, are represented using the Violent Flows (ViF) descriptor. ViF descriptors are then classified as either violent or non-violent. The authors also introduce a new dataset of crowded scenes on which their method is evaluated. According to the presented results, the ViF descriptor outperforms the Local Trinary Patterns (LTP) [146], histogram of oriented gradient (HoG) [78], histogram of oriented optical flow (HoF) [78] descriptors as well as the histogram of oriented gradient and optical flow (HNF) descriptor [78]. The ViF descriptor is also evaluated on well-known datasets of videos of non-crowded scenes such as the Hockey [95] and the ASLAN [72] datasets in order to assess its performance in action-classification tasks of "non-textured" videos (i.e., non-crowded). With small vocabularies, the ViF descriptor outperforms the LTP and STIP descriptors, while with larger vocabularies, STIP outperforms ViF. However, this performance gain comes with a higher computational cost.

In [135], Xu et al. propose to use Motion SIFT (MoSIFT) descriptors to extract a low-level representation of a video. Feature selection is applied on the MoSIFT descriptors using kernel density estimation. The selected features are subsequently summarized into a mid-level feature representation based on a BoW model using sparse coding. The method is evaluated on two different types of datasets: Crowded scenes [59] and non-crowded scenes [95]. Although Xu et al. do not explicitly define violence, they seem to focus on violence-related concepts such as fights or sudden changes in motion. The results show that the proposed method is promising and outperforms HoG-based and HoF-based BoW representations on both datasets.

Uni-modal approaches have the disadvantage of having an "incomplete view" of videos, as they take into account only one modality (e.g., [52] is audio only, whereas [59] is dynamic visual only). The video data itself is already complex and

the difficulty of the task of violence detection is even more amplified as "violence" is also a complex concept. Therefore, there is a need to incorporate information about videos from different perspectives.

2.4.2.2 Multi-modal violence detection approaches

Multi-modal methods, which constitute the most common type of approach used in violent content detection in videos, consist in fusing audio and visual cues at either feature or decision level. Wang et al. [28] use audio and static visual features in order to detect horror in videos. Color and texture features are used for the static visual representation of video shots, while MFCC are used for the audio representation. More specifically, the mean, variance and first-order differential of each dimension of MFCC are employed for the audio representation. As observed from their results [28], using color and textural information in addition to MFCC features slightly improves the performance. However, the authors do not explicitly state their definition of horror. Therefore, assessing the performance of their method and identifying the situations on which it properly works is difficult.

Giannakopoulos et al. [53], in an attempt to extend their approach based solely on audio cues [52], propose to use a multi-modal two-stage approach. In the first step, they perform audio and visual analysis of segments of one second duration. In the audio analysis part, audio features such as energy entropy, ZCR and MFCC are extracted and the mean and standard deviation of these features are used to classify scenes into one of seven classes (violent ones including shots, fights and screams). In the visual analysis part, the average motion, motion variance and average motion of individuals appearing in a scene are used to classify segments as having either high or low activity. The classification probabilities obtained in this first step are then used to train a binary classifier (violence versus non-violence).

In [54], a three-stage method that uses audio and visual cues is proposed. The authors employ audio-visual features such as motion activity, ZCR, MFCC, pitch and rhythm features to characterize video segments with fast-tempo. Although not explicitly stated, the authors define violent scenes as those which contain action and violence-related concepts such as gunshots, explosions and screams. The method was only evaluated on action movies. However, violent content can be present in movies of all genres (e.g., drama). The performance of this method in genres other than action is, therefore, unclear.

Lin and Wang [80] employ audio, static and dynamic visual features to analyze violence in videos. As audio features, spectrum power, brightness, bandwidth, pitch, MFCC, spectrum flux, ZCR and harmonicity prominence features are extracted. An audio vocabulary is subsequently constructed by k-means clustering. Audio clips of one second length are represented with mid-level audio features with a technique derived from text analysis. However, this approach presents the drawback of only constructing a dictionary of twenty audio words, which prevents having a precise representation of the audio signal of video shots. In the visual classification part, the degree of violence of a video shot is determined by using mo-

tion intensity, the (non-)existence of flame, explosion and blood appearing in the video shot. Violence-related concepts addressed in this work are fights, murders, gunshots and explosions. This method was also evaluated only on action movies. Therefore, the performance of this solution in genres other than action is uncertain.

Chen et al. [29] define a violent scene as one that contains action and blood. The average motion, camera motion, and average shot length are used for scene representation to classify scenes as action and non-action. Subsequently, using the "Viola-Jones" face detector, faces are detected in each keyframe of action scenes and the presence of blood pixels near detected human faces is checked using color information. The approach is compared with the method of Lin and Wang [80] because of the similar violence definitions, and is shown to perform better in terms of precision and recall.

Ding et al. [43] observe that most existing methods identify horror scenes only from independent frames, ignoring the context cues among frames in a video scene. In order to consider contextual cues in horror scene recognition, they propose a joint sparse coding based MIL technique which simultaneously takes independent and contextual cues into account and use frame-based audio and static visual features for video scene representation (commonplace audio-visual representations such as MFCC, color and texture features). Their definition of violence is very similar to the definition in [28]. They perform experiments on a horror video dataset collected from the Internet and the results demonstrate that the performance of the proposed method is superior to other existing well-known MIL algorithms.

The approaches of this category consider a more complete view of videos, as they use multi-modal representations. However, the most of reviewed solutions are problem-dependent and concentrate more on the feature engineering part of violence analysis. Besides, the definition of violence is either not even provided or very restrictive. For instance, Chen et al. [29] define violent scenes as the scenes that contain action and blood, which is a highly restricted description.

2.4.3 Literature review: Modeling perspective

In the previous section (Section 2.4.2), we presented existing works from a representation perspective and gave an overview of representations used for describing videos. In this section, the papers addressing violence detection are presented from the modeling perspective.

One popular type of approach adopted in the literature is classification based on SVM models using different types of kernels. An illustration to SVM-based solutions is the work by Giannakopoulos et al. [52], where a polynomial SVM based on audio features is used as the classifier. Some authors construct SVMs using motion descriptors; for instance, de Souza et al. [36] apply a linear SVM-based approach using local motion descriptors, whereas Hassner et al. [59] present a method based also on a linear SVM using global motion descriptors. The work introduced in [135] concentrates on the feature engineering part and simply uses an SVM with RBF ker-

nel to construct violence analysis models. Multiple kernel learning is also employed in SVM-based violence modeling (e.g., [55]). Using SVM modeling in a staged setup also appears to be common for violence modeling. The three-stage method presented in [54] is one example. In the first stage, they apply a semi-supervised cross-feature learning algorithm [139] on the extracted audio-visual features for the selection of candidate violent video shots. In the second stage, high-level audio events (e.g., screaming, gun shots, explosions) are detected via SVM training using RBF kernel for each audio event. In the third stage, the outputs of the classifiers generated in the previous two stages are linearly weighted for final decision. Similar to [54], the above-mentioned work by Chen et al. [29] constitutes a two-phase solution. In the first phase, an SVM with RBF kernel based on motion features is used for the classification of video scenes into action and non-action. In the second phase, regions near facial areas are analyzed in each keyframe of action scenes to detect the presence of blood pixels.

Next to SVM-based solutions, approaches which make use of other types of learning-based classifiers exist. For instance, Yan et al. [142] adopt a Multi-task Dictionary Learning approach to detect complex events in videos. Based on the observation that complex events are made of several concepts, certain concepts useful for particular events are selected by means of combination of text and visual information. Subsequently, an event oriented dictionary is learned. The experiments are conducted on the TRECVid Multimedia Event Detection dataset. The same authors have experimented Multi-task Learning in other situations. For instance in [141], Yan et al. employ a variant of Linear Discrimant Analysis (LDA – used to find a linear combination of features capable of characterizing or discriminating several classes) called Multi-task LDA to perform multi-view action recognition based on temporal self-similarity matrices. More recently, Yan et al. [140] have developed a Multi-task Learning approach for head-pose estimation in a multi-camera environment under target motion.

There also exist approaches which apply Bayesian networks to model violence (e.g., [53] and [99]). Penet et al. [99] investigate Bayesian networks and different kinds of Bayesian network structure learning algorithms for violence detection at the video shot level. Giannakopoulos et al. [53] apply a two-stage approach. In the first stage, Bayesian networks are used to build binary classifiers to model violence-related concepts such as fights, screams and activity level, and non-violence-related concepts such as music and speech. The classification probability outputs of the first stage constitute the feature space for the k-NN classifier used in the second stage to model violence.

Considering a video scene as a set of shots and analyzing violence levels using MIL algorithms is another technique. For instance, the Multi-view MIL (M^2IL) modeling algorithm proposed in [43] considers a video scene as a bag of video shot instances and construct both representations from the independent view and the contextual view to model violence. Aiming at improving SVM-based classification, Wang et al. [28] apply MIL (MI-SVM [11]) using audio-visual features. Video scenes

are divided into video shots, where each scene is formulated as a bag and each shot as an instance inside the bag for MIL.

Co-training is another learning approach applied to model violence. This is for instance used by Lin and Wang in [80], where separate classifiers for audio and visual analysis are trained. Probabilistic latent semantic analysis is applied in the audio classification part. In the visual classification part, the degree of violence of a video shot is determined by using violence-related concepts (i.e., high activity, flame or explosion, and blood). Finally, the audio and visual classifiers are combined in a co-training framework.

As discussed in this section, different modeling algorithms are applied in the literature to construct violence analysis models. However, violence detection approaches mostly build a unique model for the "violence" concept (e.g., [99]) or incorporate domain-knowledge such as violence analysis through violence-related concepts (e.g., [53]).

2.4.4 MediaEval Violent Scenes Detection Task

The MediaEval Violent Scene Detection (VSD) task derives from a use case attributed to the company *Technicolor*[4]. The French producer of video content and entertainment technologies adopted the aim of helping users to select movies that are suitable to watch with their children. Detailed description of the task (from 2012 to 2015) including the dataset, the ground-truth and evaluation criteria are given in [38, 40, 37, 112]. The works discussed in the following paragraphs employ the same definitions of violence adopted in the MediaEval VSD task (i.e., *objective* and/or *subjective* violence as discussed in Section 2.4.1.2). We present here the most promising approaches proposed for the task since its introduction in 2011 in terms of the official evaluation metric of the task.

Penet et al. [99] propose to exploit temporal and multi-modal information for "objective" violence detection at the video shot level. In order to model violence, different kinds of Bayesian network structure learning algorithms are investigated. The proposed method is tested on the dataset of the MediaEval 2011 VSD Task. Experiments demonstrate that both multimodality and temporality add valuable information into the system and improve the performance in terms of the MediaEval cost function [38]. The best performing method achieves 50% false alarms and 3% missed detection, ranking among the best submissions to the MediaEval 2011 VSD task.

In [61], Ionescu et al. address the detection of "objective" violence in Hollywood movies using neural networks. The method relies on fusing mid-level violence-related concept predictions inferred from low-level audio-visual features. The mid-level concepts used in this work are gory scenes, the presence of blood, firearms and cold weapons (for the visual modality); the presence of screams and gunshots

[4] `http://www.technicolor.com/`

(for the audio modality); and car chases, the presence of explosions, fights and fire (for the audio-visual modalities). The authors employ a bank of multi-layer perceptrons featuring a dropout training scheme in order to construct the aforementioned 10 violence-related concept classifiers. The predictions of these concept classifiers are then merged to construct the final violence classifier. The method is tested on the dataset of the MediaEval 2012 VSD task and ranked first among 34 other submissions, in terms of precision and F-measure.

In [55], Goto and Aoki propose an "objective" violence detection method which is based on the combination of visual and audio features extracted at the segment level, using machine learning techniques. Violence detection models are learned via multiple kernel learning. The authors also propose mid-level violence clustering in order to implicitly learn mid-level concepts without using manual annotations as in [61]. The proposed method is trained and evaluated on the MediaEval 2013 VSD task using the official metric Mean Average Precision at 100 (MAP@100). The results show that the method outperforms the approaches which use no external data (e.g., Internet resources) in the MediaEval 2013 VSD task.

Derbas and Quénot [41] explore the joint dependence of audio and visual features for both "objective" and "subjective" violent scene detection. They first combine the audio and the visual features and then determine statistically joint multimodal patterns. The proposed method mainly relies on an audio-visual BoW representation. The experiments are performed in the context of the MediaEval 2013 VSD task. The obtained results show the potential of the proposed approach in comparison to methods which use audio and visual features separately, and to other fusion methods such as early and late fusion. They also compare k-NN and SVM learning approaches within the task of 2013 and report superior results for k-NN classification.

In 2014 and 2015, using deep neural networks both for the representation and modeling aspects of violence detection became popular among participants of the MediaEval VSD task. As in previous years, the focus of most of the proposed approaches lay in the feature engineering part with SVM-based modeling (especially linear SVM) being applied in the modeling part. For the representation part, using CNNs emerged as the most common way of feature generation. Deep learning was also applied in modeling (e.g., [34]). The work in [34] presents a deep learning approach to capture complex relationships between audio and visual features during violence modeling. The authors propose a regularized deep neural network (DNN) which first applies a feature abstraction for each input feature and then uses another layer to identify feature relationships (i.e., feature fusion), and finally builds a classification model in its last layer. Input features to the DNN are MFCC audio features and advanced motion features such as dense trajectory based motion features and STIP. The best performing approach of 2015 [35] employs deep learning architectures for feature learning and simply uses SVM as classifier. In addition to conventional motion and audio features such as MFCC, STIP and dense trajectory based motion features, the proposed system consists of several learned representa-

tions based on CNN and Long Short Term Memory (LSTM) [132]. First, the authors train a CNN with a violence-related subset of ImageNet[5] classes. They adopt a two-stream CNN framework [111] to generate features both from static frames and optical flow vectors, and subsequently apply LSTM models on top of the two-stream CNN features to capture long-term temporal dynamics [35]. All these learned and handcrafted features are finally used to build SVM-based violence analysis models.

The drawbacks of the approaches presented at the MediaEval VSD task (except [55]) are similar to the ones discussed in the previous section (Section 2.4.3). We discuss the drawbacks of the work [55] in detail in Chapter 6.

2.5 Summary

This background chapter laid down essential definitions and presented a review of the related research in the literature. First, we explained the main affective analysis concepts and terms used in this thesis. More specifically, the definition of affective phenomena including emotions and emotion representation models were discussed. Second, we gave a literature review in the field of affective content analysis in videos. The definition of violence and the related research within the context of violence detection were also provided. More specifically, we discussed violence analysis approaches from the representation and modeling perspectives. Finally, we presented details on the promising approaches which took part to the MediaEval Violent Scenes Detection (VSD) task which is running since 2011.

The remainder of this thesis is divided in two: A first part (I), relating to emotion analysis in general; a second one (II), addressing violence detection. In the following chapter (first in Part I), an affective video content analysis framework that learns audio and static visual representations from raw audio-visual data is presented.

[5] http://www.image-net.org/

Part I

Affective Content Analysis

3

LEARNING AUDIO AND STATIC VISUAL CUES: APPLICATION ON EDITED VIDEOS

Video affective content analysis approaches that are based on machine learning techniques make use of feature representations at different abstraction levels to model the emotional content of videos. When designing an affective content analysis algorithm for videos, one of the most important steps is, therefore, the selection of discriminative features for the effective representation of video segments. One usually employed representation is based on audio and static visual cues. The majority of existing methods concentrate on the feature engineering part and either use low-level audio-visual representations or generate handcrafted higher level ones. These features suffer from several deficiencies, such as *limited discriminative power* (especially for the low-level ones) or *poor applicability* to new environments and scenarios (especially for the higher level ones). We tackle these problems first by introducing a deep learning architecture that eliminates the need for feature engineering by learning a hierarchy of audio or static visual features from the raw audio or visual data; those learned representations are shown to be more discriminative than handcrafted low-level and mid-level ones of the same modality (audio or visual). We also investigate the extent to which the fusion of the learned audio and static visual features further improves the uni-modal analysis on professionally edited music video clips. The work presented in this chapter has been published in [1, 5].

3.1 Introduction

Classifying, retrieving and subsequently delivering personalized video content corresponding to the needs of the consumers is a challenge which still has to be resolved. Video affective analysis – which consists in identifying videos segments that evoke particular emotions in the user [129] – can bring an answer to such a challenge from an original perspective. In particular, in the context of categorical affective analysis, where human emotions are defined in terms of discrete categories (opposed to dimensional affective analysis where they are non-discrete [56]), one direction followed by many researchers consists in using machine learning methods (e.g., [64, 138, 145]). Machine learning approaches make use of a specific data representation (i.e., features extracted from the data) to identify particular events or concepts. Their performance is heavily dependent on the choice of the data representation on which they are applied [19]. As in any pattern recognition task, one key issue is, therefore, to find an effective representation of video content.

The common approach for feature representation used to model the affective content of videos is either to employ low-level audio-visual features or to build handcrafted higher level representations based on the low-level ones (e.g., [32, 62, 85, 136]). Low-level features have the disadvantage of losing global relations or structure in the data, whereas creating handcrafted higher level representations is time consuming, problem-dependent and requires domain knowledge. In this context, in the field of audio, video and more generally multi-dimensional signal processing, automatically and directly learning suitable features (i.e., mid-level features) from raw data to perform tasks such as event detection, summarization, retrieval has attracted particular attention, especially because such learning kept the amount of required supervision to a minimum and provided scalable solutions. To achieve this, deep learning methods such as convolutional neural networks (CNNs) and deep belief networks are shown to provide promising results (e.g., [68, 79, 109]). The advantages of deep learning architectures are: (1) *Feature re-use*: constructing multiple levels of representation or learning a hierarchy of features, and growing ways to re-use different parts of a deep architecture by changing the depth of the architecture; (2) *abstraction and invariance*: more abstract concepts can often be constructed in terms of less abstract ones and have potentially greater predictive power (i.e., reduced sensitivity to changes in the input data) [19]. The recent success of deep learning methods incited us to directly learn feature representations from automatically extracted raw audio and color features by deep learning (more specifically, the CNN) to obtain mid-level audio and visual representations.

We develop in the present chapter a direct and categorical affective analysis framework and consequently address the following research question: **(RQ) What is the discriminative power of learned audio and static visual representations against handcrafted ones of the same modality (audio or static visual)?** We create CNN architectures for audio and static visual data of videos to learn hierarchical audio and visual representations. Subsequently, we review the discriminative

power of audio and static visual representations learned from raw audio-visual data through deep learning against handcrafted audio and static visual ones at different abstraction levels (low-level or mid-level). To sum up briefly, our aim in this chapter is to extract higher level audio and static visual representations from raw video data without performing feature engineering to create effective audio and visual representations for affective content analysis of videos.

Our approach differs from the existing works (presented in Section 2.3) in the following aspects: **(1)** Both audio and static visual feature representations are learned from automatically extracted raw data by 2D CNN modeling and are fused at the decision-level by SVM modeling for the affective classification of music video clips; **(2)** it is experimentally shown that the learned mid-level audio-visual representations are more discriminative and provide more precise results than low-level and handcrafted mid-level audio-visual ones. The 2D CNN modeling allows to recognize particular inherent audio or static visual patterns in the video data which facilitate the generation of more discriminative classification models.

The chapter is organized as follows. Section 3.2 gives an overview of the affective content analysis system. In Section 3.3, we introduce our audio and static visual representation learning method used in this work. Section 3.4 discusses how affective content analysis models are built with the learned representations. We provide and discuss evaluation results on a subset of the DEAP dataset [73] in Section 3.5. Finally, concluding remarks and future directions to expand our method are presented in Section 3.6.

3.2 Overview of the Affective Content Analysis System

In this section, we present our categorical affective analysis approach. An overview of our method is illustrated in Figure 3.1. The presented system contains the following steps: (1) Highlight extracts of music video clips are first segmented into fixed-length pieces; (2) audio and visual feature extraction (i.e., MFCC and color values in the RGB and HSV color spaces); (3) learning mid-level audio and static visual representations *(training phase only)*; (4) generating mid-level audio-visual representations; (5) generating an affective analysis model *(training phase only)*; (6) classifying a fixed-length video clip segment into one of related emotion categories *(test phase only)*; and (7) classifying a complete music video using the results obtained on the fixed-length segments constituting the video *(test phase only)*.

3.3 Representation Learning for Video Affective Content Analysis

In this section, we discuss the details of our audio and static visual representation learning architectures. Section 3.3.1 presents the architecture for audio, whereas the one for static visual is discussed in Section 3.3.2.

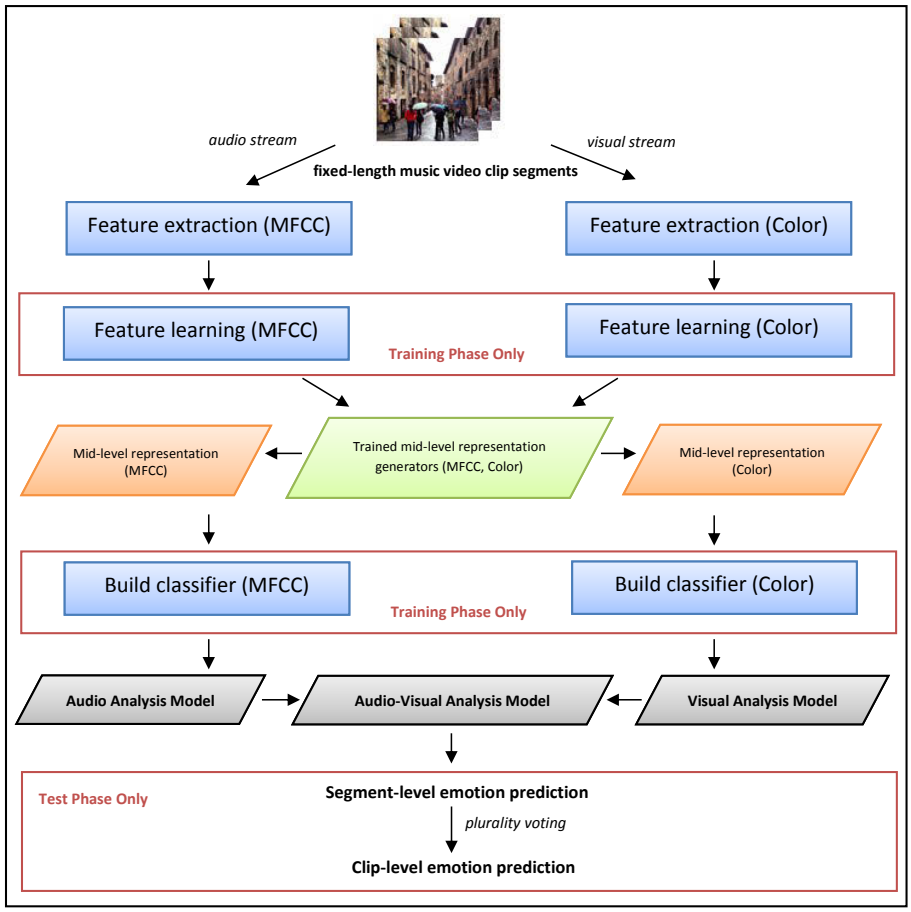


Figure 3.1: A high-level overview of our method for music video affective analysis. Final classification decisions are realized by a *plurality voting* process. (MFCC: Mel-Frequency Cepstral Coefficients)

3.3.1 Learning audio representations

MFCC values are extracted for each fixed-length video segment. The set of resulting MFCC feature vectors, when concatenated, constitutes an MFCC-time domain representation that can be regarded as an "image", on which we apply 2D CNN modeling. As CNNs were successfully applied to detect patterns from raw image data (e.g., [69, 74]), we considered them for the representation we obtained. The aim is to capture both timbre (basically the tone color or tone quality of sound) and temporal information.

The first layer (i.e., the input layer) of the CNN is an $M \times N$ map (M = number of feature vectors, N = feature dimension) which contains the MFCC feature vectors from the frames of one segment. In Figure 3.2, the CNN architecture used to generate audio representations is presented. The CNN has three convolution and two subsampling layers, and one output layer which is fully connected to the last convolution layer (this network size in terms of convolution and subsampling layers has experimentally given satisfactory results). In the VA-based classification, the output layer consists of four units: One for each quadrant of the VA-space. In the wheel-based classification, the output layer consists of nine units (annotation according to Russell's circumplex model): One unit for each emotion category. Each unit in the output layer is fully connected to each of the units in the last convolution layer. After training, the output of the last convolution layer is used as the *mid-level audio representation* of the corresponding video segment. Hence, the MFCC feature vectors from the frames of one segment are converted into a multi-dimensional feature vector (which constitutes a more abstract audio representation) capturing the acoustic information in the audio signal of the video segment.

3.3.2 Learning static visual representations

Existing works (e.g., [83, 124, 150]) have shown that colors and their proportions are important parameters to evoke emotions. This observation has motivated our choice of color values for the generation of visual representations for music videos. The traditional approach of using CNNs on images for feature learning is performed in the RGB color space (e.g., [74, 119]). Hinted by the prior art evidence that perception of emotions by humans is enhanced in a hue-saturation type color space [124], we close this gap here by working in the HSV color space in addition to RGB and evaluate the discriminative power of learned representations based on both color spaces (RGB versus HSV).

Keyframes (i.e., representative frames) are extracted from each music video clip segment, as the frame in the middle of the video segment. For the generation of mid-level visual representations, we extract color information in the RGB and HSV color spaces from the keyframe of each segment. The resulting color values in each channel are given as input to a separate CNN. In Figure 3.3, the CNN architecture used to generate visual representations is presented. The input layer of the CNN is an $X \times Y$ map (X = number of pixels in the x-dimension, Y = number of pixels in the y-dimension) which contains the color values from one color channel of the keyframe. The CNN has three convolutional and two subsampling layers, one full connection and one output layer (this network size in terms of convolution and subsampling layers has also experimentally given satisfactory results). The training of the CNN is done similarly to the training of the CNN in the audio case. As a result, the color values in each color channel are converted into a multi-dimensional feature vector. The feature vectors generated for each of the three color channels (RGB or HSV) are concatenated into a final feature vector which forms a more ab-

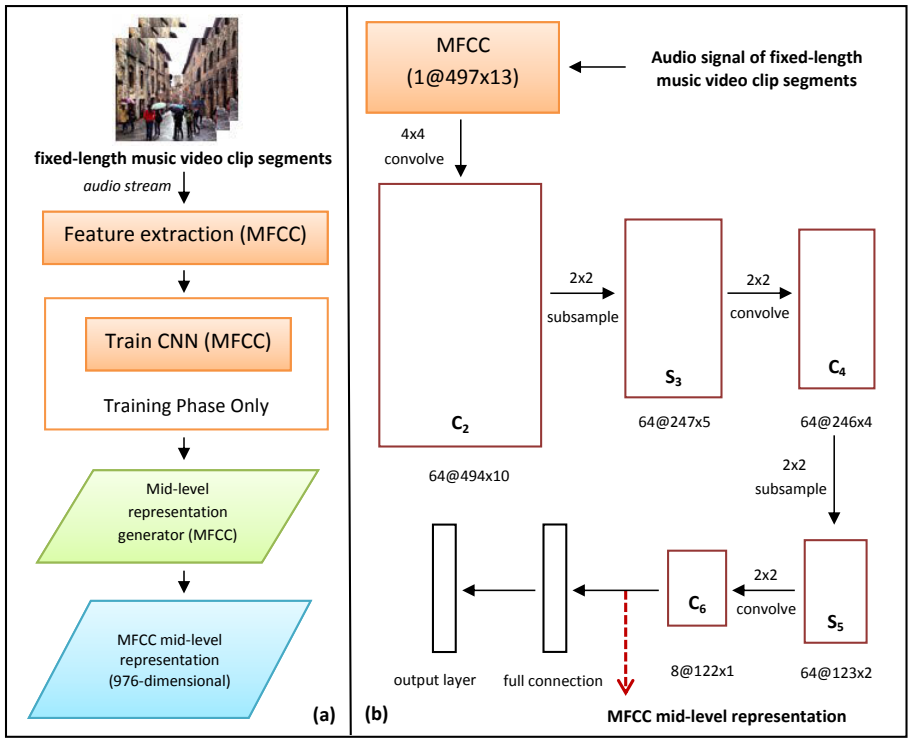


Figure 3.2: (a) A high-level overview of our mid-level audio representation generation process, (b) the detailed CNN architecture for audio representation learning. The architecture contains three convolution and two subsampling layers, one output layer fully connected to the last convolution layer (C_6). (CNN: Convolutional Neural Network, MFCC: Mel-Frequency Cepstral Coefficients)

stract visual representation capturing the color information in the keyframe of the music video segment.

3.4 Generating the Affective Analysis Model

As discussed in the previous sections (Sections 3.2 and 3.3), we only use neural network structures to extract higher level audio and static visual representations and employ multi-class SVMs for model generation. There are several reasons that incited us to choose SVM as classifier in the system. First, in the literature, SVM is proven to be powerful for semantic multimedia analysis including affective content analysis [70, 130, 138]. Second, during the optimization of SVM, there are only two hyper parameters (γ and C in this work) to determine. Third, since the learned

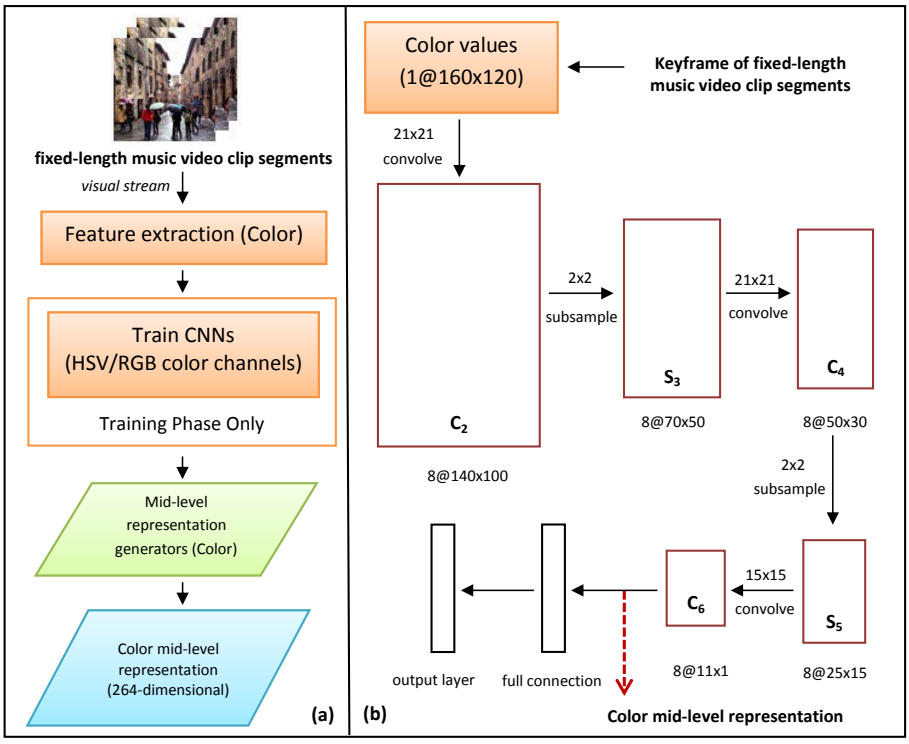


Figure 3.3: (a) A high-level overview of our mid-level color representation generation process, (b) the detailed CNN architecture for color representation learning. The architecture contains three convolution and two subsampling layers, one output layer fully connected to the last convolution layer (C_6). (CNN: Convolutional Neural Network)

audio and visual representations are compared to handcrafted audio and visual features in this work, a classifier like SVM is needed to deal with both type of representations (learned and handcrafted). Last but not least, we work with a small dataset and think that a kernel method such as SVM is a more proper choice over neural network as classifier.

The representation of video segments from different modalities (audio and visual) raises the question of modality fusion during the model generation of the system. The fusion of different modalities is generally performed at two levels within the field of multimedia content analysis: Feature-level (early fusion) and decision-level (late fusion) [13]. As previous studies (e.g., [115]) on semantic video analysis have experimentally shown that late fusion strategies perform usually better than early ones in most cases, we apply an *SVM-based late fusion strategy* in this work.

For the generation of an audio affective analysis model, learned audio representations are fed into a multi-class SVM. Similarly, a visual affective analysis model is also generated by feeding learned visual representations (RGB or HSV-based ones) into a second multi-class SVM. The probability estimates of the two SVM models are subsequently fed into a third multi-class SVM to generate an audio-visual affective video analysis model. Normally, in a basic SVM, only class labels or scores are output. The class label results from thresholding the score, which is not a probability measure. The scores output by the SVM are converted into probability estimates using the method explained in [131].

In the test phase, mid-level audio and visual representations are created by using the corresponding CNN models for fixed-length music video segments based on MFCC feature vectors and color values in the RGB or HSV color space. The music video segments are then classified by using the affective video analysis models. Final decisions for the classification of music video segments are realized by a plurality voting process where a music video is assigned the label which is most frequently encountered among the set of fixed-length segments constituting the video.

3.5 Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of uni and multi-modal learned representations against handcrafted audio-visual ones. In addition, a comparison of our mid-level audio features against auditory temporal modulations (ATM) features is provided. The ATM features, which describe temporal modulations in the frequency domain, were recently applied for audio content analysis, in particular in the context of music recommendation [117] and genre classification [118]. A direct comparison with the methods which are also tested on the DEAP dataset is limited mainly due to the usage of different subsets of the DEAP dataset in different works. However, our approach is compared against the work presented in [145] whose experimental setup is the closest one to ours. An overview of the DEAP dataset is provided in Section 3.5.1. In Section 3.5.2, the experimental setup is presented. Finally, results and discussions are given in Sections 3.5.4 (uni-modal) and 3.5.5 (multi-modal).

3.5.1 Dataset and ground-truth

The DEAP dataset [73] is a dataset for the analysis of human affective states using electroencephalogram, physiological and video signals. It consists of the ratings from an online self-assessment where 120 highlight extracts of music videos were each rated by 14 to 16 volunteers based on arousal, valence and dominance. We have used all the music video clips whose YouTube links are provided in the dataset and which were available on YouTube at the time when we conducted the experiments (74 music clips). As only the highlight extracts of music videos have been

used during the dataset creation, we perform our evaluations with these extracts. The excerpts of different affective categories downloaded from YouTube equate to 888 fixed-length music video segments.

We perform categorical affective analysis according to two different classification schemes, namely *VA-based* and *wheel-based* classification. For the case of the *VA-based classification*, four affective labels exist. These are *high arousal-high valence (HA-HV)*, *low arousal-high valence (LA-HV)*, *low arousal-low valence (LA-LV)* and *high arousal-low valence (HA-LV)*, each representing one quadrant in the VA-space. Concerning the evaluation of *wheel-based classification*, annotations according to Russell's circumplex model (discussed in detail Section 2.2) are used. The labels for both classification schemes are provided in the dataset and the VA-based classification labels are determined by the average ratings of the participants in the online self-assessment. Table 3.1 summarizes the main characteristics of the DEAP dataset in more detail.

Table 3.1: The characteristics of the DEAP dataset with respect to VA-based category (top) and wheel-based category (bottom). Note that for the *VA-based* classification there are 74 music videos, whereas this number is 70 for the *wheel-based* classification. This is due to one of the videos for which the emotion wheel rating was not properly recorded, and to the fact that there is only 1 clip for each of the *Hope*, *Surprise* and *Contempt* categories.

VA-based Category	**# Music Videos**	**# Segments**
high arousal-high valence (HA-HV)	19	228
low arousal-high valence (LA-HV)	19	228
low arousal-low valence (LA-LV)	14	168
high arousal-low valence (HA-LV)	22	264
Total	**74**	**888**
Wheel-based Category	**# Music Videos**	**# Segments**
Pride	2	24
Elation	3	36
Joy	22	264
Satisfaction	5	60
Interest	4	48
Sadness	18	216
Fear	2	24
Disgust	6	72
Anger	8	96
Total	**70**	**840**

3.5.2 Experimental setup

The length of highlight extracts of music video clips is one minute, whereas each fixed-length piece of the extracts lasts 5 seconds (as suggested in [145]) and corresponds to 125 visual frames. Once the extracts are segmented into fixed-length pieces, the construction of feature representation is performed for each piece separately.

In order to extract the 13-dimensional MFCC features, we employed the MIR Toolbox v1.6.1[1]. Frame sizes of 25 ms at a rate of 10 ms were used. Mean and standard deviation for each dimension of the MFCC feature vectors were computed, which compose the low-level audio representation (*LLR audio*) of music video segments. As mid-level handcrafted audio representation (*MLR handcrafted audio*), vector quantization based Bag-of-Words (BoW) representation of MFCC features was generated. An audio dictionary of size k (k is equal to 512 in this work) was constructed using $400 \times k$ MFCC descriptors and k-means clustering [94] (experimentally determined figure).

In order to generate the RGB-based low-level visual features of music video segments (*LLR colorRGB*), we constructed 64-bin color histograms for each color channel in the RGB color space resulting in 192-dimensional low-level visual feature vectors. In order to generate the HSV-based low-level visual features (*LLR colorHSV*) of video segments, we constructed normalized HSV histograms (of size 16, 4, 4 bins, respectively) in the HSV color space.

We used the Deep Learning toolbox[2] in order to generate mid-level audio and visual representations with CNN models (*MLR audio*, *MLR colorRGB* and *MLR colorHSV*). The learned audio representations (*MLR audio*) are constructed as explained in Section 3.3.1 and are 976-dimensional, whereas the learned color representations (*MLR colorRGB* and *MLR colorHSV*) whose construction is discussed in detail in Section 3.3.2 are 264-dimensional. The sigmoid was used as the activation function. The mean-sampling – in which basically the mean over non-overlapping regions of size N $\times$ N ($N = 2$ in this work) is calculated – was used in the sampling layers of the CNN models. The CNN models were trained with backpropagation and stochastic gradient descent. The batch sizes during CNN model training for audio and visual modalities were 74. A fixed learning rate of 0.2 was used for 500 epochs for both audio and visual modalities. The learning rate and number of epochs were determined empirically according to the classification accuracies of uni-modal SVM models.

We trained the multi-class SVMs with an RBF kernel using libsvm[3] as the SVM implementation. Training was performed using audio and visual features extracted at the music video segment level. SVM hyper parameters were optimized using a grid search and 5-fold cross-validation using the development set. The search

[1] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox`

[2] `https://github.com/rasmusbergpalm/DeepLearnToolbox/`

[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

range for the γ parameter was [3, −15] with a step size of −2, whereas the range for the C parameter was [−5, 15] with a step size of 2. All segmented samples belonging to a specific video were always either in the training set or in the test set during cross-validation. For multi-modal analysis (Section 3.5.5), the fusion of audio and visual features was performed at the decision-level using an SVM (SVM-based fusion is explained in Section 3.4). Due to the limited amount of available music video samples, we used the *leave-one-song-out cross* validation scheme. During model generation, zero mean unit variance normalization was applied on feature vectors of the current development dataset.

In the assessment of the discriminative power of learned audio representations (*MLR audio*), we compare them against ATM features. The reason is that both extract conceptually similar information from the audio data, the major distinction being that MLR audio features are learned, while ATM features do not involve a learning phase but require domain knowledge. As mentioned in Section 3.3.1, the developed MLR audio features capture both the timbre and temporal information of an acoustic signal. Similarly, ATM features constitute a representation of slow temporal modulations of acoustic signals. More specifically, motivated by the human auditory and visual systems [118], they describe the power variation over modulation scales in each primary auditory cortex channel. The extraction of the ATM features was realized as follows: (1) The audio signal of a video segment of 5-second length is converted to a mono signal and down-sampled to 16 kHz, (2) auditory spectrograms are computed as described in [144], (3) the computed auditory spectograms are used to find temporal modulations in the frequency domain through (inverse) discrete Fourier transforms. The implementation details of the ATM feature extraction can be found in [118].

3.5.3 Evaluation metrics

In this chapter, the main metric used to evaluate the performance of the system is classification accuracy. Due to the unbalanced nature of the dataset that we work with according to the wheel-based classification scheme, we also provide macro precision, recall and F-measure metrics which are basically the average of the class-wise precision, recall and F-measure metrics and are computed as illustrated in Equation 3.1 [116].

$$\begin{aligned} Precision_M &= \frac{\sum_{i=1}^{c} \frac{tp_i}{tp_i+fp_i}}{c} \\ Recall_M &= \frac{\sum_{i=1}^{c} \frac{tp_i}{tp_i+fn_i}}{c} \\ F-measure_M &= \frac{Precision_M \cdot Recall_M}{Precision_M + Recall_M} \end{aligned} \tag{3.1}$$

where tp_i, fp_i and fn_i are respectively true positive, false positive and false negative rates for class i; and c is the number of classes.

3.5.4 Evaluation of uni-modal learned representations

The experiments presented in this section aim at discussing the computation times for the representation learning phase and comparing the discriminative power of mid-level audio and visual features which are learned from raw data against handcrafted low and mid-level audio and visual ones.

Computationally, the most expensive phase of the representation learning is the training of the CNNs which takes on average 150 and 350 seconds per epoch for MFCC and color features, respectively. The generation of feature representations using CNNs amounts to 0.5 and 1.2 seconds on average per 5-second video segment for MFCC and color features, respectively. All the timing evaluations were performed with a machine with 2.40 GHz CPU and 8 GB RAM.

Table 3.2: *VA-based classification* accuracies and the macro metrics (macro precision, recall and F-measure) on the DEAP dataset with uni-modal (audio or visual only) representations. The best classification performance is highlighted. (LLR: low-level representation, MacroP: macro precision, MacroR: macro recall, MacroF: macro F-measure, MLR: mid-level representation).

Uni-modal Type	**Accuracy (%)**	**MacroP**	**MacroR**	**MacroF**
LLR audio	37.84	0.391	0.355	0.341
LLR colorHSV	28.38	0.276	0.261	0.239
LLR colorRGB	27.03	0.268	0.250	0.230
MLR audio	**48.65**	**0.509**	**0.466**	**0.464**
MLR handcrafted audio	41.89	0.438	0.397	0.388
MLR colorHSV	43.24	0.454	0.410	0.431
MLR colorRGB	40.54	0.429	0.386	0.379

In Table 3.2, we present the VA-based classification accuracies as well as the macro metrics (macro precision, recall and F-measure) on the DEAP dataset with uni-modal representations. The observations inferred from the results are as follows:

- Mid-level audio and visual representations (learned or handcrafted) outperform the low-level ones in terms of classification accuracy.

- Learned audio and visual representations are superior to the mid-level handcrafted ones (except learned RGB-based visual).

- Color representations based on the HSV color space are superior to the ones of the same level (low or mid-level) based on the RGB color space.

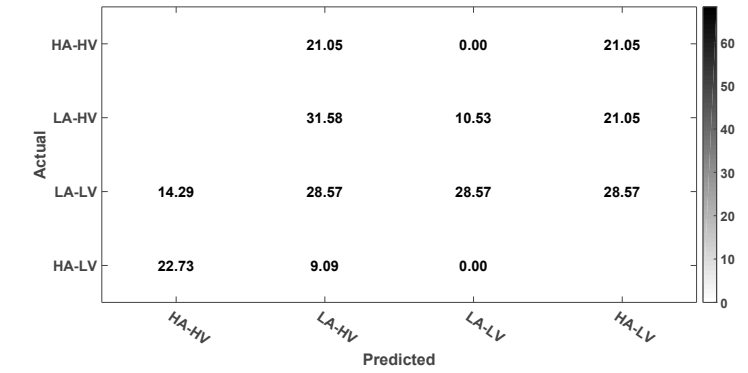
- The aforementioned observations are also confirmed by the macro metrics.

Table 3.3: *Wheel-based classification* accuracies and the macro metrics (macro precision, recall and F-measure) on the DEAP dataset with uni-modal (audio or visual only) representations. The best classification performance is highlighted. (LLR: low-level representation, MacroP: macro precision, MacroR: macro recall, MacroF: macro F-measure, MLR: mid-level representation).

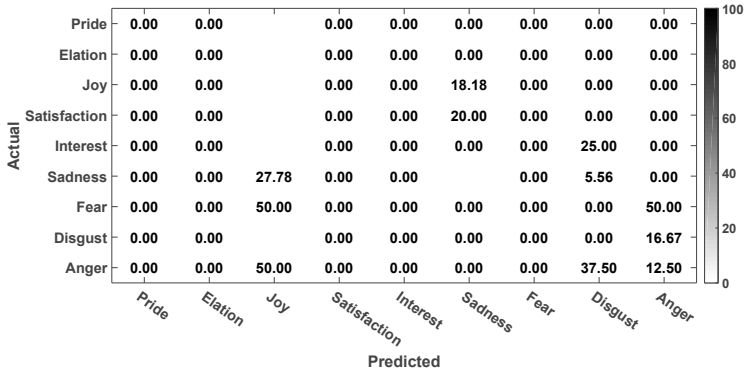
Uni-modal Type	**Accuracy (%)**	**MacroP**	**MacroR**	**MacroF**
LLR audio	27.14	0.107	0.112	0.099
LLR colorHSV	34.29	0.131	0.139	0.123
LLR colorRGB	22.86	0.095	0.095	0.086
MLR audio	41.43	0.151	0.168	0.146
MLR handcrafted audio	40.00	0.145	0.163	0.142
MLR colorHSV	**44.29**	**0.160**	**0.179**	**0.156**
MLR colorRGB	31.43	0.121	0.128	0.113

Table 3.3 reports the wheel-based classification accuracies and the macro metrics on the DEAP dataset with uni-modal representations. The observations concerning the wheel-based classification are similar to the ones derived for the VA-based classification (Table 3.2). However, there are three significant differences in comparison to the results for the VA-based classification. First, HSV-based color information appears to be a more discriminative representation for music videos compared to the audio ones. Second, RGB-based learned color representation lead to poorer results compared to the results for the VA-based classification (e.g., the HSV-based low-level color representation performs better than the learned RGB-based one, although the latter one is learned). Last but not least, the macro metric values are quite low compared to the VA-based classification. This is mainly due to the unbalanced nature of the samples in the dataset according to the wheel-based classification scheme. This is confirmed by the fact that the system has non-zero macro metrics only for the classes *Joy*, *Sadness* and *Anger* which are the first three classes based on the number of samples. As a final remark on Table 3.3, the difference between classification accuracies (Table 3.2 versus Table 3.3) can be explained by the increased risk of confusion due to number of classes.

We present the confusion matrices on the DEAP dataset using the best performing uni-modal representations (according to Tables 3.2 and 3.3) in Figure 3.4 to provide an overview of the misclassification behavior of the system with uni-modal representations for the VA-based and wheel-based classification schemes. For the VA-based classification (Figure 3.4(a)), *HA-LV* (i.e., high arousal and low valence) and *HA-HV* (i.e., high arousal and high valence) are the two classes that can be discriminated at the highest level and also the most confused pair of classes. For the wheel-based classification, Figure 3.4(b) shows the effects of the unbalanced nature of the dataset.



(a)



(b)

Figure 3.4: Confusion matrices for **(a)** the *VA-based classification* and **(b)** the *wheel-based classification*, on the DEAP dataset with the best performing uni-modal audio or visual representation (according to Tables 3.2 and 3.3). Mean accuracy: 48.65% (VA-based), 44.29% (wheel-based). Darker areas along the main diagonal correspond to better discrimination.

As a final evaluation on the uni-modal representations, a comparison between the learned audio representations (MLR audio) and ATM features is performed. The classification accuracy for MLR audio coupled with SVM is 48.65%, whereas this value is 43.24% for ATM coupled with SVM. This shows that our learned audio representations outperform ATM features by 5.4%. In order to verify that the mean of the accuracies obtained using the MLR audio differs significantly from the ones obtained using the ATM features in a statistical sense, a paired Student t-test on classification accuracies from the *leave-one-song-out* cross validation was performed. It was checked that the accuracies being compared follow a normal dis-

tribution and have the same variance using the *Jarque-Bera test* [66] and the *F-test*, respectively. This *t*-test showed that the improvement provided by MLR audio over ATM is statistically significant (5% significance level).

3.5.5 Evaluation of multi-modal learned representations

The experiments presented in this section aim at investigating the extent to which the fusion of the learned audio and static visual features further improves the uni-modal analysis on music video clips. As mid-level static visual representation, we use *MLR colorHSV* (learned from the raw color data in the HSV color space) due to its superior performance compared to the RGB-based one. A comparison against the work presented in [145] is also exposed.

Table 3.4: *VA-based classification* accuracies and the macro metrics (macro precision, recall and F-measure) on the DEAP dataset with multi-modal representations fused by SVM-based fusion. The best classification performance is highlighted. (hc: handcrafted, LLR: low-level representation, MacroP: macro precision, MacroR: macro recall, MacroF: macro F-measure, MLR: mid-level representation).

Multi-modal Type	**Accuracy (%)**	**MacroP**	**MacroR**	**MacroF**
LLR audio & LLR colorHSV	36.49	0.407	0.368	0.372
LLR audio & MLR colorHSV	40.54	0.453	0.404	0.411
MLR audio & LLR colorHSV	44.60	0.490	0.448	0.453
MLR audio & MLR colorHSV	**54.05**	**0.557**	**0.529**	**0.536**
MLR hc audio & LLR colorHSV	41.89	0.476	0.417	0.424
MLR hc audio & MLR colorHSV	50.00	0.517	0.492	0.499

In Table 3.4, we present the VA-based classification accuracies as well as the macro metrics on the DEAP dataset with multi-modal representations fused by SVM (Section 3.4). The observations inferred from the results are as follows:

- The fusion of low-level audio or color representations with representations of either level (low or mid-level – learned or handcrafted) causes a decrease in accuracy (e.g., *LLR audio* combined with *LLR colorHSV*, *MLR colorHSV* or *MLR handcrafted audio* performs poorer compared to the uni-modal results). There is a strong suspicion that this is due to the cascaded classification error introduced by an additional classification layer in the system.

- On the contrary, the fusion of mid-level audio or color representations (either learned or handcrafted) with each other leads to better results in terms of classification accuracy when compared to the results of uni-modal analysis. This is an indication that these features complement each other.

- Last but not least, the best performing representation is the fusion of learned audio and color representations. This concludes that learned representations

are superior to mid-level handcrafted ones in the multi-modal analysis as well.

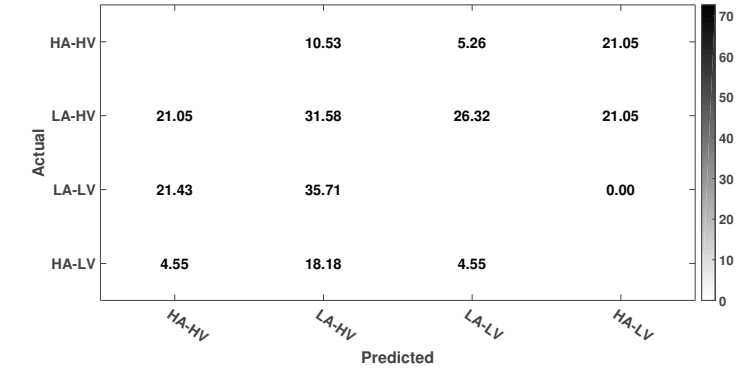
- The aforementioned statements are based on the classification accuracy of the system. When the macro metrics are considered, similar conclusions can be drawn. One difference though is that the macro metrics of the system are slightly better than the uni-modal ones, where we fuse the low-level audio or color representations with the handcrafted representations of either level (low or mid-level).

Table 3.5: *Wheel-based classification* accuracies and the macro metrics (macro precision, recall and F-measure) on the DEAP dataset with multi-modal representations fused by SVM-based fusion. The best classification performance is highlighted. (hc: handcrafted, LLR: low-level representation, MacroP: macro precision, MacroR: macro recall, MacroF: macro F-measure, MLR: mid-level representation).

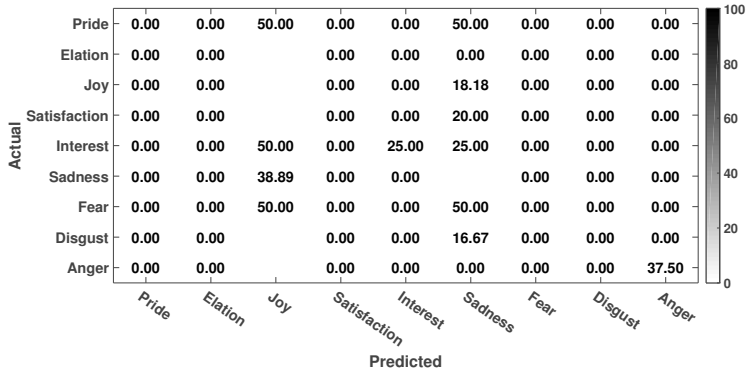
Multi-modal Type	**Accuracy (%)**	**MacroP**	**MacroR**	**MacroF**
LLR audio & LLR colorHSV	30.00	0.084	0.114	0.093
LLR audio & MLR colorHSV	32.86	0.126	0.133	0.117
MLR audio & LLR colorHSV	37.14	0.139	0.150	0.132
MLR audio & MLR colorHSV	**47.14**	**0.185**	**0.198**	**0.178**
MLR hc audio & LLR colorHSV	35.71	0.134	0.145	0.128
MLR hc audio & MLR colorHSV	44.29	0.163	0.178	0.156

Table 3.5 reports the wheel-based classification accuracies and the macro metrics on the DEAP dataset with multi-modal representations. The observations concerning the wheel-based classification are similar to the ones derived for the VA-based classification (Table 3.4), except the increase in the macro metrics for the handcrafted representations in the multi-modal analysis. As in the uni-modal analysis, the difference between classification accuracies (Table 3.4 versus Table 3.5) can be explained by the increased risk of confusion due to the number of classes. Besides, the macro metric values are still quite low compared to the VA-based classification due to the unbalanced nature of the samples in the dataset according to the wheel-based classification scheme as in the uni-modal analysis.

We present the confusion matrices on the DEAP dataset in Figure 3.5 to provide an overview of the misclassification behavior of the system for the VA-based and wheel-based classification schemes. Within the context of the VA-based classification, we observe that *HA-LV* (i.e., high arousal and low valence) and *HA-HV* (i.e., high arousal and high valence) are the two classes that can be discriminated at the highest level. Within the context of the wheel-based classification, there is a tendency of our algorithm to classify video clips belonging to classes nearby the classes *Joy* and *Sadness* according to the ground-truth as *Joy* and *Sadness*. We suspect that this is due to the imbalanced nature of the data in the development set, as there are more video clips labeled as *Joy* and *Sadness*.



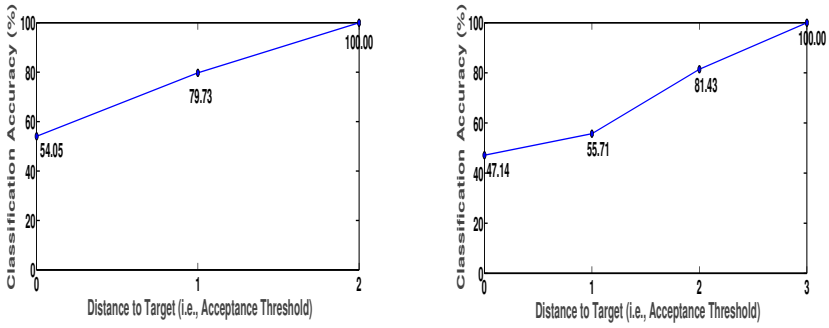
(a)



(b)

Figure 3.5: Confusion matrices for **(a)** the *VA-based classification* and **(b)** the *wheel-based classification*, on the DEAP dataset with the best performing multi-modal audio-visual representation (i.e., MLR audio and colorHSV representations). Fusion method is SVM-based fusion. Mean accuracy: 54.05% (VA-based), 47.14% (wheel-based). Darker areas along the main diagonal correspond to better discrimination.

When looking in more detail at the confusion matrices in Figure 3.5, it appears that the confusion between classes mostly occurs between neighboring classes, i.e., neighboring emotions more likely to resemble each other. Therefore, plotting the *Cumulative Matching Characteristic (CMC)* curves is a more appropriate choice to present the performance of the system as a function of the *distance between classes*, similar to the approach adopted in [148]. The CMC curve plots the accuracy of the system for a range of distance values between classes. We define the distances between classes as follows: (1) The distance between two classes that are on the same



(a) (b)

Figure 3.6: Cumulative Matching Characteristic (CMC) curve for the **(a)** *VA-based classification* and **(b)** *wheel-based classification*, on the DEAP dataset with the best performing multi-modal audio-visual representation (i.e., MLR audio and colorHSV representations). Fusion method is SVM-based fusion.

quadrant of the VA-space is 1; (2) the distances between two classes of different quadrants is defined as $d_q + 1$, where d_q corresponds to the number of quadrants encountered when departing from one class in order to reach the other class. For instance, the distance between *HA-LV* and *HA-HV* is 1, whereas it is 2 for *HA-LV* and *LA-HV*. Similarly, the distance between *Elation* and *Joy*, which are on the same VA-quadrant, is 1; whereas it is 2 between *Sadness* and *Fear* which are on two neighboring VA-quadrants. We provide the CMC curves for the VA-based and wheel-based classifications in Figure 3.6. According to this graph, when relaxing the conditions by taking into account the CMC, the accuracies appear less pessimistic.

As a final evaluation, we provide the VA-based classification accuracies of our approach with the best performing multi-modal representation (i.e., *MLR audio* and *MLR colorHSV*) and the work in [145] in Table 3.6. Learned audio-visual representations perform better in terms of classification accuracy compared to features used in [145]. However, it is worth mentioning the existence of some differences in the experimental setup between our work and [145]. Only a subset of 40 video clips from the DEAP dataset form the basis of the experiments in [145]. Thus, a comparison is biased towards our approach due to the larger dataset (74 clips). On the other hand, the 40 music video clips used in [145] were selected so that only the music video clips which induce strong emotions are used. Therefore, the dataset we used in this paper is more challenging. Another difference with the setup of [145] is that, therein, as the ground-truth, the authors used the user ratings from laboratory experiments instead of the online self-assessment ratings mentioned in Section 4.3.1. As mentioned in the beginning of the experiments, a direct comparison with other methods that use the same dataset is difficult due to different setups and data subsets used in evaluations. We think that the work by Yazdani et al. [145] was the

closest one that we are aware of and that we can compare with to get a rough idea about the performance of our system against the state-of-the-art. The comparison provided in Table 3.6 shows that multi-modal learned audio-visual representations lead to more discriminative classifiers, as the work [145] uses only low-level audio and visual features. More importantly, we show that the fusion of audio and visual representations (the learned ones to be more specific) is better than using unimodal ones for affective content classification, whereas the opposite is the case for the work in [145].

Table 3.6: *VA-based classification* accuracies of our method and the work by Yazdani et al. [145] on the DEAP dataset.

Method	**Accuracy (%)**
Our method *(*MLR audio *and* MLR colorHSV & *SVM-based fusion)*	54.05
Yazdani et al. [145]	36.00

3.6 Conclusions

In this chapter, we presented a direct and categorical affective analysis framework where audio and static visual features are learned directly from raw audio and visual data. As the feature learning method, we applied a CNN-based method which uses MFCC as raw audio data and color values in the RGB and HSV spaces as raw color data. Experimental results on a subset of the DEAP dataset support our assumptions (1) that learned audio-visual representations are more discriminative than handcrafted low-level and mid-level ones, and (2) that HSV color space based learned color features are more discriminative compared to the RGB color space based ones.

The findings of this chapter provide valuable insights for the field of affective content analysis of videos. First, we performed an extensive evaluation of learned audio and visual representations against handcrafted ones at different abstraction levels and showed that the learned ones lead to better classification performance. The second important contribution of this work is to apply 2D CNN modeling – which is usually applied on images – on audio features to capture both timbre and temporal information. The audio representation learned through the CNN modeling is shown to be more discriminative than ATM features which describe temporal modulations in the frequency domain. This has an important impact in the field of emotional content analysis, as we keep the domain knowledge to a minimum with the help of deep learning and still achieve a superior classification performance.

In this chapter, we concentrated on the feature learning perspective of affective video analysis and applied SVM modeling which is the commonplace approach for the modeling part of the analysis. Therefore, in the following chapter (Chapter 4) we will concentrate more on the modeling perspective. We employed only audio and static visual features as representation in this chapter. Motion is also an im-

portant information source for affective video analysis [126, 145]. Therefore, we will also extend our feature set with other sophisticated mid-level visual features in the next chapter (Chapter 4). Evaluations on a more challenging user-generated content dataset in addition the DEAP dataset will also be presented in Chapter 4.

4

ENSEMBLE LEARNING WITH ENRICHED MULTI-MODAL CUES: APPLICATION ON PROFESSIONALLY EDITED AND USER-GENERATED VIDEOS

In the previous chapter, we concentrated on the representation aspects of video affective content analysis and presented a novel method to learn audio and static visual representations from raw video data. In this chapter, we focus primarily on the modeling aspect and model the affective content of videos using ensemble learning instead of SVM modeling which is the dominant classification method used by the most of existing works in the field. Additionally, on the side of representation, we extend our feature set with advanced motion and domain-specific representations to incorporate temporal and domain-specific information about videos into the analysis process. In addition to the DEAP dataset, our method is evaluated on *VideoEmotion*, a more challenging dataset consisting of user-generated videos. The experimental results show that ensemble learning based models are more discriminative on both datasets and that the incorporation of temporal and domain-specific representations further improves the classification performance of the proposed framework. The work presented in this chapter has been published in [6, 8].

4.1 Introduction

In Chapter 3, we presented a direct and categorical affective analysis framework based on audio and static image features which were shown to be promising for the analysis of music videos. In this chapter, we extend our affective analysis framework both from representation and modeling perspectives to analyze user-generated videos in addition to professionally edited videos. User-generated contents are more challenging since they are not edited professionally like music videos, and can be considered as less coherent in terms of content (e.g., more chaotic, absence of a properly defined scenario). In this chapter, the extensions made to the framework address the following research questions:

(RQ1) How does ensemble learning perform in comparison to SVM modeling? The most important extension of the framework tackles the modeling aspect of video affective content analysis. As discussed in detail in Section 2.3, SVM is the most widely used learning method for the generation of analysis models (employed in e.g., [20, 47, 70, 130, 138]). Distancing ourselves from the prior art, here, we apply *ensemble learning* (more specifically, decision tree based bootstrap aggregating) to generate emotional content analysis models and compare its performance to that of SVM modeling on both music video clips and user-generated videos. In machine learning, it is well known that an ensemble of base classifiers is likely to perform better than a single base classifier. However, the application and a comprehensive analysis of ensemble classifiers in the field of affective content analysis have not been investigated yet. Hence, to the best of our knowledge, the proposed system is the first to adopt ensemble learning for emotional characterization of professionally edited and user-generated videos; and we show that the adopted ensemble learning method outperforms SVM modeling in terms of classification accuracy.

(RQ2) What is the effect of dense trajectories and SentiBank representations? Due to the difficulty of analyzing user-generated videos, the feature set should be improved and augmented, which constitutes another research challenge of this chapter. As our work presented in Chapter 3 was limited to learning audio and static visual features only, it could not optimally take into account the motion and temporal coherence exhibited in image sequences compared to a single image. Studies (e.g., [126, 145]) have, indeed, demonstrated that, in addition to audio and static color features, motion plays an important role for affective content analysis in videos. As discussed in detail in Section 2.3.3, the use of motion-based features within the context of affective content analysis is limited to simple motion features such as frame differencing. Recently, a new type of video descriptor has emerged, namely dense feature trajectories. These descriptors, corresponding to points which are densely sampled and tracked using dense optical flow fields, were introduced by Wang et al. [128] for the task of action recognition in videos, and have proven robust for action recognition. However, to the best of our knowledge,

the applicability of these improved dense trajectories to the task of affective content analysis has not been investigated yet. Distinct from the existing works presented in Section 2.3, we propose to use dense trajectory features to derive a mid-level motion representation obtained via a sparse coding based Bag-of-Words method to boost the classification performance of our system.

The above-mentioned audio, static and dynamic visual features that we employ are general-purpose representations within the context of video content analysis. In addition to them, we propose to use SentiBank domain-specific ones which are shown to be particularly effective for emotional analysis of single images [20, 70].

To sum up, in relation to this research question, we first investigate the discriminative power of dense trajectories and SentiBank and then assess the effect of incorporating these representations [20] on the classification performance.

(RQ3) How to optimally combine audio-visual features? One additional question arising when using multiple features, is that of fusion of information. We, therefore, also propose an assessment of fusion mechanisms. Optimal late fusion mechanisms are explored through the analysis of linear and SVM-based fusion for combining the outputs of uni-modal analysis models.

To put it in a nutshell, we significantly extend the affective content analysis method presented in Chapter 3 by addressing problematics relating to both representation and modeling. The important extensions are as follows. First, we apply ensemble learning (decision tree based bootstrap aggregating) in addition to SVM to model the affective content of videos and experimentally show that ensemble learning is superior to SVM modeling. Second, we include advanced motion and domain-specific representations in emotion modeling to boost the classification performance of the framework by diversifying the representation of videos. Third, we report a comprehensive analysis on a more challenging dataset (i.e., VideoEmotion [70]).

The position of our work with respect to existing solutions presented in Section 2.3 can be seen from two different perspectives. The first one is representation. When considering the papers dealing with emotion recognition from videos, the recent works closely related to ours are those by Baveye et al. [16] and Dumoulin et al. [45] where CNNs are used only in the visual feature extraction. Our approach is novel in that deep learning is used not only with visual data but also with audio data. In addition, our work makes use of advanced motion information, which is another novel aspect. Last but not least, we not only concentrate on the representation aspects, but also propose here to comprehensively assess the behavior of two popular classifiers, namely SVM and ensemble learning.

The chapter is organized as follows. In Section 4.2, we introduce our method for the affective classification of videos. We provide and discuss evaluation results on a subset of the DEAP dataset [73] and on the VideoEmotion dataset [70] in Sec-

tion 4.3. Finally, we present concluding remarks and future directions to expand our method in Section 4.4.

4.2 The Video Affective Analysis Framework

In this section, our extended direct and categorical affective analysis framework is presented. We perform the analysis according to two different schemes: *VA-based* and *wheel-based* classification, where emotion categories used in the classification are derived from the VA-space and from an emotion wheel (both discussed in detail in Section 2.2), respectively. The available dataset of user-generated videos (Flickr and YouTube videos) provides annotations according to Plutchik's emotion wheel (discussed in detail in Section 2.2). Concerning professionally edited videos (music video clips), although the dataset we use contains both VA-based and wheel-based annotations, we only report VA-based classification results. The reason is our insights gained in the previous chapter: The low number of samples in this dataset (74) considering the high number of classes to discriminate (12) causes the results to carry little statistical relevance. More details concerning these sets can be found in Section 4.3.

In Figure 4.1, we provide an overview of the system. As shown in Figure 4.1, the system consists of the following steps: (1) Videos (highlight extracts of music video clips or user-generated videos of different length) are first segmented into fixed-length pieces; (2) audio and visual feature extraction; (3) learning mid-level audio and static visual representations *(training phase only)*; (4) generating mid-level audio-visual representations; (5) generating an affective analysis model *(training phase only)*; (6) classifying a fixed-length video segment into one of related emotion categories *(test phase only)*; and (7) classifying a complete video using the results obtained on the fixed-length segments constituting the video *(test phase only)*.

We already discussed the learning of audio and visual features in detail in Chapter 3, to which the interested reader is referred. The only difference for the wheel-based classification case is the use of a different emotion representation model (Plutchik's emotion wheel instead of Russell's circumplex model) in this chapter. The output layer of the trained CNNs, therefore, consists of eight units instead of nine (annotation according to Plutchik's emotion wheel): One unit for each emotion category. The incorporation of temporal information and domain-specific information to the system is explained in Sections 4.2.1 and 4.2.2, respectively. The generation of an affective analysis model is discussed in more detail in Section 4.2.3. This model uses fusion, which is presented in Section 4.2.4.

4.2.1 Deriving mid-level dynamic visual representations

The importance of motion in edited videos such as movies and music video clips, and user-generated videos motivated us to extend the approach presented in Chapter 3 and to incorporate motion information to our analysis framework. To this

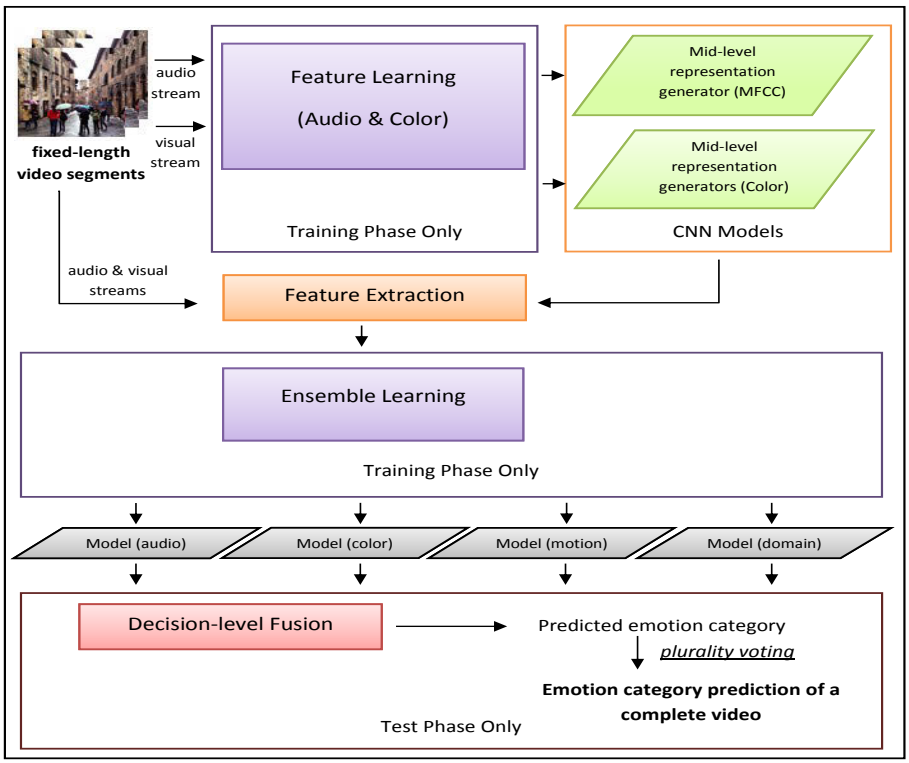


Figure 4.1: A high-level overview of the proposed system. The feature learning and extraction, affective analysis model generation and decision fusion parts are explained in detail in the subsections of Section 4.2.

end, we adopt the work of Wang et al. on improved dense trajectories [128]. Dense trajectories are dynamic visual features which are derived from tracking densely sampled feature points in multiple spatial scales. Although initially used for unconstrained video action recognition [128], dense trajectories constitute a powerful tool for motion or video description, and, hence, are not limited to action recognition only.

Our dynamic visual representation works as follows. First, improved dense trajectories [128] of length L ($L = 15$ in this work) frames are extracted from each video segment. Dense trajectories are subsequently represented by a histogram of oriented gradients (HoG), a histogram of optical flow (HoF) and motion boundary histograms in the x and y directions (MBHx and MBHy, respectively). We learn a separate dictionary for each dense trajectory descriptor (each one of HoG, HoF, MBHx and MBHy). In order to learn the dictionary of size k ($k = 512$ in this work) for sparse coding, $400 \times k$ feature vectors are sampled from the training data (this

figure has experimentally given satisfactory results). The sparse dictionary learning technique presented in [84] is used to learn sparse codes. In the coding phase, we construct the sparse representations of dense trajectory features using the LARS algorithm [46]. Given dense trajectory features and a dictionary as input, the LARS algorithm returns sparse representations for the feature vectors (i.e., sparse mid-level motion representations). In order to generate the final sparse representation which is a set of dense trajectory feature vectors, we apply the *max-pooling* technique [143].

4.2.2 Incorporating domain-specific representations

In addition to general-purpose mid-level audio, static and dynamic visual representations, we incorporate mid-level semantic representations of the visual content of videos for high-level video affective content analysis. More specifically, we use SentiBank [20] which is based on emotion-related concepts. In the first version of SentiBank, there were 1,200 concept detectors; this number increased to 2,089 in the subsequent version[1] (version 1.1). As briefly explained in Section 2.3.1, each emotion-related concept is defined as an Adjective-Noun Pair (ANP) such as "cute baby" or "dark forest". In these ANPs, adjectives (e.g., "funny", "peaceful", "gorgeous", "weird") are strongly connected to emotions, and nouns (e.g., "baby", "dog", "car", "wedding") are usually objects or scenes that can be automatically detected [20]. Each dimension in the SentiBank representation corresponds to the detection score of the corresponding ANP concept detector. In this work, we use both version 1.0 and version 1.1 of the SentiBank representations (1,200 and 2,089 ANP concepts) in order to assess the influence of the number of concepts on the performance of video affective content classification.

4.2.3 Generating the affective analysis model

In addition to SVM modeling [6], we apply *ensemble learning* in order to build affective analysis models. For the generation of the models, mid-level audio, static and dynamic visual, and domain-specific representations are fed into separate classifiers. Mid-level audio and static visual representations are created by using the corresponding CNN models for fixed-length video segments both in the training and test phases. In the training phase, decision trees are combined with bootstrap aggregating (i.e., bagging). In the test phase, the final prediction which corresponds to the prediction score of the ensemble of the trees for the test data is computed as the average of predictions from individual trees. The prediction score generated by each tree is the probability of a test sample originating from the related class computed as the fraction of samples of this class in a tree leaf.

The prediction scores of the models are merged using one of the fusion strategies presented in Section 4.2.4. Once all fixed-length video segments extracted from

[1] `http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html`

a given video are classified, final decisions for the classification of the complete video is realized by a *plurality voting* process. In other words, a video is assigned the label which is most frequently encountered among the set of fixed-length segments constituting the video.

4.2.4 Fusion strategies

When combining results of multiple classifiers, fusion of the results constitutes an important step. As mentioned in Chapter 3, previous studies on semantic video analysis (e.g., [115]) have experimentally shown that late fusion strategies perform usually better than early ones in most cases. Therefore, we investigate in this chapter two distinct late fusion techniques to combine the outputs of the classification models, namely *linear fusion* and *SVM-based fusion*.

4.2.4.1 *Linear fusion*

In linear fusion, probability estimates obtained from the classifiers trained separately with one of the mid-level audio, static and dynamic visual, and domain-specific representations are linearly fused at the decision-level. The classifiers that we adopt are all based on the same ensemble learning algorithm. Hence, we are in the presence of homogeneous "learners" (i.e., all of the same type) according to the terminology of [152]. In such a situation, it is advised to directly fuse the probabilities ($h_i(x_j)$) generated by each of the classifiers (i.e., "learners") using the *weighted soft voting* technique [152]:

$$H(x_j) = \sum_{i=1}^{T} w_i h_i(x_j) \tag{4.1}$$

Classifier-specific weights (w_i in Equation 4.1) are optimized on the training data. The weights assigned to the classifiers are determined in such a way that they always sum up to 1.

4.2.4.2 *SVM-based fusion*

In SVM-based fusion, the probability estimates of the unimodal classifiers are concatenated into vectors and used to construct higher level representations for each video segment. Subsequently, an SVM classifier which takes as input these higher level representations is constructed. This fusion SVM is then used to predict the label of a video segment. When SVMs are used as unimodal classifiers, the scores returned by the SVMs are first converted into probability estimates using the method explained in [50].

4.3 Performance Evaluation

The experiments presented in this section aim primarily at comparing the discriminative power of our method – which is based on mid-level representations learned and derived from audio-visual data coupled with ensemble learning as presented in Section 4.2 – against the works presented in [5, 6, 70, 97, 145]. Accessorily, as in the previous chapter (Chapter 3), our mid-level audio representations are compared against auditory temporal modulations (ATM) features. Section 4.3.1 provides an overview of the DEAP and VideoEmotion datasets used for the evaluation. In Section 4.3.2, we present the experimental setup. Finally, we provide results and discussions in Section 4.3.3.

4.3.1 Dataset and ground-truth

The experiments are conducted on two types of video content: (1) Professionally edited videos (DEAP dataset) and (2) user-generated videos (VideoEmotion dataset). The details of the DEAP dataset such as its annotation process are already discussed in Section 3.5.1. Therefore, here we concentrate on the details of the recently introduced VideoEmotion user-generated video dataset.

The VideoEmotion dataset consists of 1,101 videos collected from Flickr and YouTube. The videos are annotated according to 8 categories, each category corresponding to a basic emotion represented in a section of Plutchik's wheel of emotions (Figure 2.1). In addition to using the whole VideoEmotion dataset, as suggested in [70], we also provide results on a subset of the dataset which contains only four basic emotions more frequently used in the literature ("anger", "fear", "joy", "sadness").

Two different classification schemes are envisaged, one being *VA-based* and the other being *wheel-based*. The evaluation of the *VA-based classification* is performed only on the DEAP dataset. Concerning the evaluation of *wheel-based classification*, we use VideoEmotion, for which the labels are also provided as part of the dataset [70]. Tables 4.1 and 4.2 summarize the main characteristics of the DEAP and of the VideoEmotion datasets.

Table 4.1: The characteristics of the DEAP dataset with respect to VA-based category.

VA-based Category	**# Music Videos**	**# Segments**
high arousal-high valence (HA-HV)	19	228
low arousal-high valence (LA-HV)	19	228
low arousal-low valence (LA-LV)	14	168
high arousal-low valence (HA-LV)	22	264
Total	**74**	**888**

Table 4.2: The characteristics of the VideoEmotion dataset with respect to wheel-based category.

Wheel-based Category	# Flickr Videos	# YouTube Videos	Total
Anger	23	78	101
Anticipation	40	61	101
Disgust	100	15	115
Fear	123	44	167
Joy	133	47	180
Sadness	63	38	101
Surprise	95	141	236
Trust	44	56	100
Total	**621**	**480**	**1,101**

4.3.2 Experimental setup

The length of highlight extracts of music video clips is one minute, whereas the length of user-generated videos is of varying length. Each fixed-length video segment of the videos lasts 5 seconds (as suggested in [145]) and corresponds to 125 visual frames. Once the videos are segmented into fixed-length pieces, the construction of feature representation is performed for each video segment separately. Table 4.3 gives an overview of the features extracted to represent the video segments and the details of feature extraction are discussed in the following paragraphs.

Table 4.3: An overview of extracted audio, static and dynamic visual features in the framework. (LLR: low-level representation, MLR: mid-level representation).

Name	Modality	Dimension#
LLR audio	audio	26
MLR handcrafted audio	audio	512
MLR audio	audio	976
LLR color	static visual	24
MLR color	static visual	264
MLR attribute1200	static visual	1,200
MLR attribute2089	static visual	2,089
MLR motion	dynamic visual	2,048

The setup for the low-level audio (*LLR audio*) and the mid-level handcrafted audio (*MLR handcrafted audio*) representations of video segments was explained in detail in Section 3.5.2. As low-level color representation (*LLR color*), we only constructed normalized HSV histograms (of size 16, 4, 4 bins, respectively) in the HSV color space and exclude the low-level color representation based on the RGB color space in view of poor classification performance results in Chapter 3.

The details about the generation of mid-level audio and color representations with CNN models (*MLR audio* and *MLR color*) were explained in the previous chapter as well (Section 3.5.2). However, there is one point worth mentioning about the setup as a consequence of the additional dataset (VideoEmotion) used in this chapter. The batch sizes during the training of the CNN models for audio and visual modalities were 200 for the VideoEmotion dataset.

Wang's dense trajectory implementation[2] is used to extract improved dense trajectories from video segments. The sampling stride, which corresponds to the distance by which extracted feature points are spaced, is set to 20 pixels due to time efficiency concerns. Subsequently, BoW representations based on the motion features (HoG, HoF, MBHx and MBHy descriptors) of the dense trajectories (*MLR motion*) are generated as explained in Section 4.2. The extraction of dense trajectories takes on average 16 seconds per 5-second video segment. All the timing evaluations were performed with a machine with 2.40 GHz CPU and 8 GB RAM. SentiBank representations are extracted as explained in Section 4.2 using both versions of SentiBank, i.e., 1,200 and 2,089 trained visual concept detectors (*MLR attribute1200* and *MLR attribute2089*).

The multi-class SVMs with an RBF kernel were trained using libsvm [27] as the SVM implementation. Training was performed separately for each audio or visual descriptor extracted at the video segment level. SVM hyper parameters were optimized using a grid search and 5-fold cross-validation using the development set. All segmented samples belonging to a specific video were always either in the training set or in the test set during cross-validation. The search range for the γ parameter was [3, -15] with a step size of -2, whereas the range for the C parameter was [-5, 15] with a step size of 2. In order to determine the number of trees in the bagging modeling, we applied a 5-fold cross-validation on the development dataset and set the number of trees as the one that leads to the best classification accuracy. As in SVM modeling, all samples belonging to a specific video were always either in the training set or in the test set during cross-validation. During model generation, zero mean unit variance normalization was applied on feature vectors of the development dataset. In multi-modal analysis (Section 4.3.3.2), fusion of audio and visual features is performed at the decision-level by linear or SVM-based fusion, as explained in Section 4.2.4. For the DEAP dataset, due to the limited amount of available music video samples, we used the *leave-one-song-out* cross validation scheme; whereas for the VideoEmotion dataset, we used the 10 train-test splits provided as part of the dataset.

As in Chapter 3, ATM features were used to evaluate the discriminative power of the learned audio representations. The extraction process of the ATM features was explained in detail in the previous chapter (Section 3.5.2).

[2] `https://lear.inrialpes.fr/people/wang/improved_trajectories`

4.3.3 Results and discussions

The evaluation of our extended framework is achieved from different perspectives: ***(i)*** Comparison of the discriminative power of the uni-modal representations coupled with ensemble learning and SVM for the VA-based and wheel-based classification; ***(ii)*** comparison of the discriminative power of the multi-modal representations fused at the decision-level and coupled with ensemble learning and SVM for the VA-based and wheel-based classification; ***(iii)*** comparison against the state-of-the-art. The perspective *(i)* is discussed in Section 4.3.3.1, whereas the latter two perspectives *(ii)* and *(iii)* are presented in Section 4.3.3.2. Finally, Section 4.3.3.3 provides valuable insights gained with the experimental results.

4.3.3.1 Evaluation of uni-modal modeling

In this section, we present and discuss the results of uni-modal modeling, where we use only one modality as representation and apply bagging or SVM modeling as the classification method. The aim of these experiments is to investigate the discriminative power of learned audio-visual, motion and SentiBank representations, and also to compare the bagging approach against the SVM modeling (experiments relevant for the research questions *RQ1* and *RQ2*).

MLR audio versus ATM features. The classification accuracy on the DEAP dataset for MLR audio coupled with ensemble learning is 50.00%, whereas it reaches 47.30% for ATM coupled with ensemble learning. The comparison for SVM modeling was already discussed in Chapter 3 and is, therefore, omitted in this section. We continue with the MLR audio-ATM comparison on the VideoEmotion dataset. The findings for VideoEmotion are in accordance with the comparison realized for DEAP. MLR audio features provide an improvement ranging from 1.66% to 3.90% in classification accuracy over ATM. On the entire VideoEmotion set, we obtained for MLR-audio associated with SVM and ensemble learning 34.19% and 40.33%, respectively. For ATM associated with SVM and ensemble learning, the figures are 30.29% and 37.20%, respectively. On the VideoEmotion subset, the results are 42.17%, 50.14%, 40.51%, 46.96%. The experiments on both datasets have shown that our learned audio representations outperform ATM features. To verify that the mean of the accuracies obtained using the MLR audio differs significantly from the ones obtained using the ATM features in a statistical sense, a paired Student t-test on classification accuracies from the cross validation was performed for each dataset separately (DEAP, entire and subset VideoEmotion). As in the previous chapter, it was checked that the accuracies being compared follow a normal distribution and have the same variance using the *Jarque-Bera test* and the *F-test*, respectively. This t-test showed that the improvement provided by MLR audio over ATM is statistically significant (5% significance level).

VA-based Classification – Accuracy Evaluation. The classification accuracies on the DEAP dataset in the case where only one type of descriptor is employed is presented in Figure 4.2. Various observations can be inferred based on the overall results presented in **Figure 4.2**.

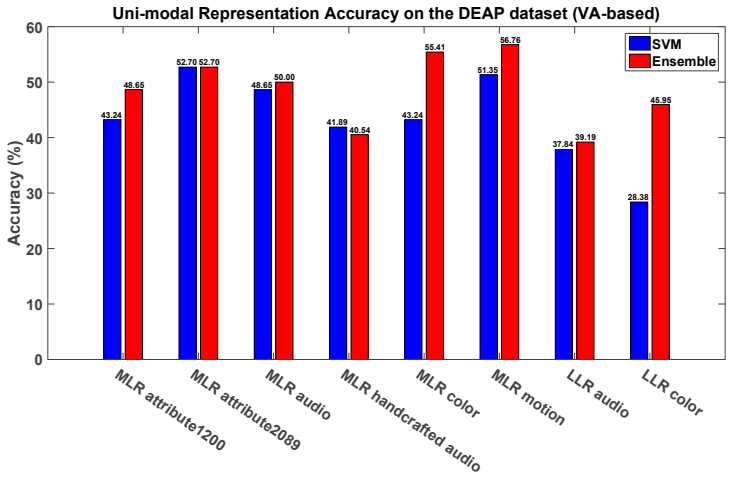


Figure 4.2: *VA-based classification* accuracies on the DEAP dataset with uni-modal (audio or visual-only) representations. (LLR: low-level representation, MLR: mid-level representation)

- Concerning the classification methods, ensemble learning, in general, improves the discrimination power of uni-modal representations over SVM-based learning, independently of the features considered (except for *MLR handcrafted audio*).

- Concerning the features other than domain-specific representations, we first note that the motion representation is the best performing descriptor for both ensemble learning and SVM-based learning. This constitutes evidence that dense motion trajectories are *strong* (i.e., discriminative) features for visual analysis of videos. The superiority of the dynamic visual feature can be explained by the fact that affect present in video clips is often characterized by motion (e.g., camera motion).

- Another observation concerns the performance gain (around 10%) of using learned color features compared to low-level ones. When evaluated together with our previous findings about learning color representations in Chapter 3,

we can conclude that color values in the HSV space lead to more discriminative mid-level representations than color values in the RGB space.

- Concerning domain-specific representations, using 2,089 trained visual concept detectors instead of 1,200 provides a noticeable increase (around 4%) in terms of classification performance.

Wheel-based Classification – Accuracy Evaluation. In the remaining of this section, we address uni-modal modeling on the VideoEmotion dataset. VideoEmotion is a dataset made of user-generated Flickr and YouTube videos, and, hence, is more challenging compared to the DEAP dataset which is made of professionally edited videos. Further, the videos in VideoEmotion are annotated only according to Plutchik's wheel, which implies that the results presented below only cover *wheel-based classification*.

We start evaluations on the VideoEmotion dataset with the classification accuracies in the case where only one type of descriptor is employed (Figures 4.3 and 4.4).

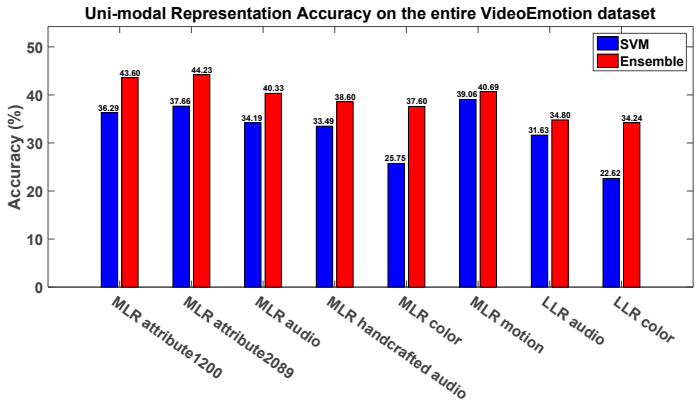


Figure 4.3: *Wheel-based classification* accuracies on the entire VideoEmotion dataset with uni-modal (audio or visual-only) representations. (LLR: low-level representation, MLR: mid-level representation)

Figure 4.3 presents the performance of each descriptor on the entire VideoEmotion dataset which is annotated according to eight emotion categories as explained in Section 4.3.1. When compared to the classification performance reported for the DEAP dataset in Figure 4.2, classification accuracies are lower for the VideoEmotion dataset. This can be explained by the fact that VideoEmotion contains videos which are not necessarily recorded or edited by professionals, and therefore, is more challenging compared to DEAP. The conclusions to which we can

come from Figure 4.3 for VideoEmotion are mostly in concordance with the ones drawn for DEAP.

- Concerning classifiers, ensemble learning improves the discrimination power of uni-modal representations over SVM-based learning in general.

- Concerning features, SentiBank domain-specific representations are the best performing descriptors (especially when 2,089 trained visual concept detectors are used instead of 1,200), followed by mid-level motion and audio descriptors.

- In addition, all mid-level representations (learned or handcrafted) outperform the low-level audio and visual representations. Further, learned audio representations outperform handcrafted mid-level audio ones.

However, there is one significant difference in comparison to the results for DEAP. Although mid-level motion descriptors are still one of the most discriminative descriptors, they are no longer the best performing ones. This can be explained by the fact that motion present in a professionally edited video is "deliberately" present to elicit specific emotions in the audience, whereas this might not be the case for user-generated videos.

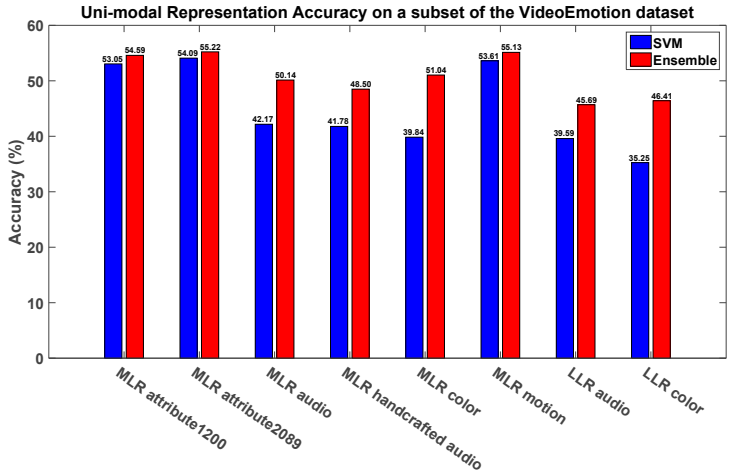


Figure 4.4: *Wheel-based classification* accuracies on the VideoEmotion subset with uni-modal (audio or visual-only) representations. (LLR: low-level representation, MLR: mid-level representation)

Figure 4.4 presents the performance of each descriptor on a subset of the VideoEmotion dataset, where we have four basic emotion categories as explained in Sec-

tion 4.3.1. As a preliminary remark, we note that the results on Figure 4.4 are globally better than those on Figure 4.3. The explanation for this discrepancy lies in the lower number of classes to be discriminated (4 against 8), which means that the risk of confusion is reduced.

The conclusions concerning the subset of VideoEmotion (Figure 4.4) are similar to the ones derived for the whole set (Figure 4.3). Although the improvement it provides is lower on the subset (in comparison to the whole set), ensemble learning still outperforms SVM-based learning. SentiBank and mid-level motion representations are the best performing uni-modal descriptors.

4.3.3.2 Evaluation of multi-modal modeling

In this section, an evaluation of the performance of different combinations of the audio-visual representations with linear and SVM-based fusion is performed on the DEAP and VideoEmotion datasets, where the best performing multi-modal representation and optimal decision-level fusion mechanisms are investigated (relevant for the research questions *RQ2* and *RQ3*).

VA-based Classification – Accuracy Evaluation. In **Table 4.4**, we present the classification performances (according to the *VA-based* annotations) of different combinations of audio-visual representations using linear and SVM-based fusion on the DEAP dataset. The feature combinations we consider involve dense motion trajectories and SentiBank domain-specific representations. This choice is justified by the findings concerning individual features. The experiments have indeed demonstrated that dense motion trajectories and SentiBank domain-specific representations are discriminative features. Further, mid-level learned or handcrafted audio-visual representations are more discriminative than low-level ones. Therefore, we combine the dense motion trajectory and SentiBank representations with mid-level audio and color features in order to evaluate the discriminative power of different multi-modal audio-visual representations.

From Table 4.4, we can derive the following observations which apply both to linear and SVM-based fusion.

- First, the performance gain of combining mid-level motion features with mid-level audio and color ones is higher than the gain of combining them with low-level audio and color ones.

- In addition, the results show that combining low-level color features with mid-level motion and audio ones leads to a decrease in classification accuracy (i.e., leads to more confusion between classes). On the contrary, combining mid-level learned color features with mid-level motion and audio ones leads to less confusion between classes (i.e., increased classification accuracy).

Table 4.4: *VA-based classification* accuracies of multi-modal audio-visual representations using *linear fusion* on the DEAP dataset. (LLR: low-level representation, MLR: mid-level representation).

	Multi-modal Representation	No attribute	MLR attr1200	MLR attr2089
Linear fusion	*MLR motion & LLR audio*	70.27	**74.32**	**74.32**
	MLR motion & MLR audio	75.68	**78.38**	**78.38**
	MLR motion & LLR color	67.57	68.92	**74.32**
	MLR motion & MLR color	70.27	70.27	**75.68**
	MLR motion & LLR audio & LLR color	72.97	74.32	**75.68**
	MLR motion & LLR audio & MLR color	75.68	78.38	**79.73**
	MLR motion & MLR audio & LLR color	72.97	75.68	**77.03**
	MLR motion & MLR audio & MLR color	77.03	78.38	**81.08**
SVM-based fusion	*MLR motion & LLR audio*	67.57	**70.27**	**70.27**
	MLR motion & MLR audio	68.92	71.62	**72.97**
	MLR motion & LLR color	60.81	60.81	**63.51**
	MLR motion & MLR color	64.87	64.87	**66.22**
	MLR motion & LLR audio & LLR color	66.22	66.22	**67.57**
	MLR motion & LLR audio & MLR color	70.27	71.62	**74.32**
	MLR motion & MLR audio & LLR color	66.22	66.22	**70.27**
	MLR motion & MLR audio & MLR color	71.62	71.62	**75.68**

- We also observe that including domain-specific representations in the final decision process improves classification accuracy; as regards this aspect, using 2,089 trained visual concept detectors instead of 1,200 for domain-specific representations increases classification accuracy in general.

- As a final remark, concerning fusion schemes, SVM-based fusion leads to poorer results compared to linear fusion. Like in the previous chapter (Section 3.5.5), we suspect that this is due to the cascaded classification error introduced by an additional classification (model generation) layer in the system.

Wheel-based Classification–Accuracy Evaluation. In the following paragraphs, we present the *wheel-based classification* performances of different combinations of the audio-visual representations using linear and SVM-based fusion (**Table 4.5** for the entire VideoEmotion dataset; **Table 4.6** for the VideoEmotion subset), where the best performing multi-modal representation and optimal decision-level fusion mechanisms are investigated (experiment relevant to the research questions *RQ2* and *RQ3*).

- The first observation (according to **Table 4.5**) is that learned audio-visual representations when combined with motion representations perform better than the combination of motion and SentiBank domain-specific features.

- In addition, combining motion and SentiBank representations with learned audio features achieves better results than combining them with mid-level handcrafted audio ones.

Table 4.5: *Wheel-based classification* accuracies of multi-modal audio-visual representations on the entire VideoEmotion dataset. (hc: handcrafted, MLR: mid-level representation).

	Multi-modal Representation	No attrib	MLR attr1200	MLR attr2089
Linear fusion	*MLR motion & MLR hc audio*	44.60	45.85	46.87
	MLR motion & MLR audio	45.16	46.78	48.05
	MLR motion & MLR color	43.87	45.32	46.32
	MLR motion & MLR hc audio & MLR color	45.50	46.85	47.45
	MLR motion & MLR audio & MLR color	46.66	47.33	**49.19**
	MLR motion & MLR attribute1200	43.08	N/A	N/A
	MLR motion & MLR attribute2089	43.33	N/A	N/A
SVM-based fusion	*MLR motion & MLR hc audio*	43.78	43.87	44.87
	MLR motion & MLR audio	44.32	44.78	46.05
	MLR motion & MLR color	43.24	43.71	45.08
	MLR motion & MLR hc audio & MLR color	44.34	45.17	46.27
	MLR motion & MLR audio & MLR color	45.15	46.39	**47.18**
	MLR motion & MLR attribute1200	42.60	N/A	N/A
	MLR motion & MLR attribute2089	43.14	N/A	N/A

- Concerning domain-specific representations, the use of 2,089 trained visual concept detectors instead of 1,200 provides again an increase in classification accuracy.

- Our final observation is that the best classification performance (highlighted in the related tables) is achieved by combining learned audio-visual representations with motion and SentiBank (*MLR attribute2089*) at decision-level and that simple linear fusion is superior to SVM-based fusion. As evoked earlier, this is likely caused by the cascaded classification error due to an added classification (model generation) layer in the system.

Table 4.6 presents the performance of multi-modal audio-visual representations on the VideoEmotion subset using linear and SVM-based fusion. We can draw conclusions which match those deduced for the entire VideoEmotion dataset. The only important difference is that SentiBank performs much better than learned audio-visual representations when combined with motion representations. This result is actually on par with the results presented in [70], where attribute features that include SentiBank (*MLR attribute1200*) are shown to outperform audio-visual representations. The difference between classification accuracies (when compared to Table 4.5) can again be explained by the increased risk of confusion due to number of classes.

Further Evaluation on VA-based and Wheel-based Classification. In order to give an overview of the misclassification behavior of the system, we present the confusion matrices on the DEAP dataset in Figure 4.5, whereas the confusion matrices on the entire VideoEmotion dataset and on the VideoEmotion subset are illustrated in in Figure 4.6(a) and Figure 4.6(b), respectively. In Figure 4.5, we observe

Table 4.6: *Wheel-based classification* accuracies of multi-modal audio-visual representations on the VideoEmotion subset. (hc: handcrafted, MLR: mid-level representation).

	Multi-modal Representation	No attrib	MLR attr1200	MLR attr2089
Linear fusion	*MLR motion & MLR hc audio*	54.59	57.47	61.81
	MLR motion & MLR audio	55.39	58.53	62.36
	MLR motion & MLR color	55.63	57.01	61.14
	MLR motion & MLR hc audio & MLR color	56.00	59.30	62.16
	MLR motion & MLR audio & MLR color	57.01	60.34	**63.65**
	MLR motion & MLR attribute1200	57.77	N/A	N/A
	MLR motion & MLR attribute2089	60.53	N/A	N/A
SVM-based fusion	*MLR motion & MLR hc audio*	52.25	56.73	60.40
	MLR motion & MLR audio	54.82	57.21	61.77
	MLR motion & MLR color	55.20	56.37	59.91
	MLR motion & MLR hc audio & MLR color	56.50	58.88	61.34
	MLR motion & MLR audio & MLR color	57.58	59.56	**62.16**
	MLR motion & MLR attribute1200	55.59	N/A	N/A
	MLR motion & MLR attribute2089	59.67	N/A	N/A

that *HA-LV* (high arousal and low valence) is the class that can be discriminated at the highest level. The common characteristic of misclassified classes is that the number of instances of those classes in both the development and test sets are limited.

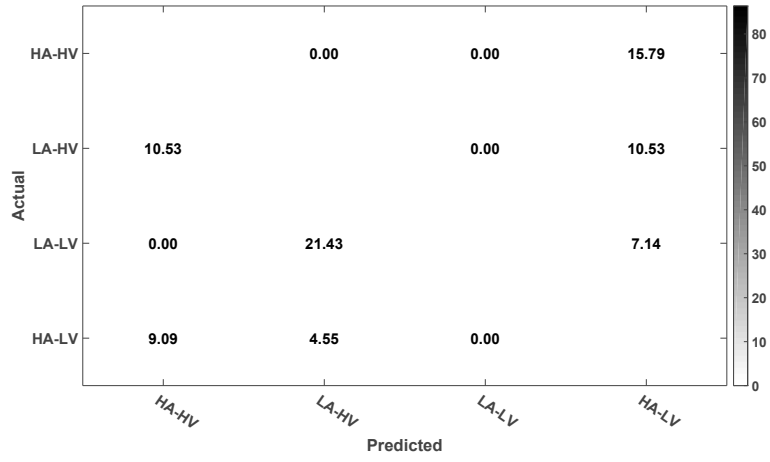


Figure 4.5: Confusion matrices for the *VA-based classification* on the DEAP dataset with the best performing multi-modal audio-visual representation (MLR audio, motion, color and domain-specific representations). Fusion method is linear fusion. Mean accuracy: 81.08%. Darker areas along the main diagonal correspond to better discrimination.

In **Figure 4.6(a)**, we observe that *Surprise* is the class that can be discriminated the most. The *Anticipation* and *Trust* classes are difficult to differentiate; it seems that these classes do not contain clear audio-visual cues. In **Figure 4.6(b)** (where only four basic emotions are considered), we observe that the *Joy* and *Fear* classes can be very well discriminated from other classes, whereas *Sadness* is the class with the lowest recognition rate.

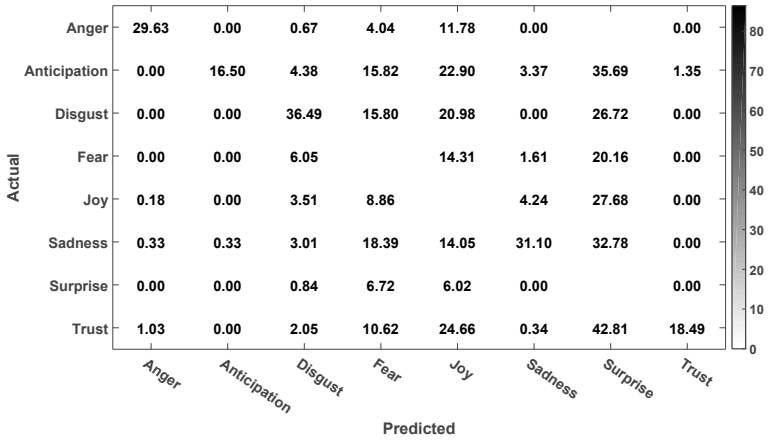
When looking in more detail at the confusion matrices in Figure 4.5 and Figure 4.6, it appears that the confusion between classes is mostly between neighboring classes, i.e., neighboring emotions more likely to resemble each other. Therefore, the *Cumulative Matching Characteristic (CMC)* curves are plotted to present the performance of the system as a function of the *distance between classes* as in the previous chapter (Section 3.5.5). We provide the CMC curve on the DEAP dataset in **Figure 4.7**. According to this graph, when relaxing the conditions by taking into account the CMC, the accuracies appear less pessimistic.

As Plutchik's wheel is used for the VideoEmotion dataset, the distances between classes are defined in a slightly different manner than for the DEAP dataset. Here, the distance between any two classes is defined as the minimum number of emotion leaves encountered when going from one class to the other in the emotion wheel (Figure 2.1). For instance, the distance between *Fear* and *Anger*, which are opposite emotions, is 4. Similarly, the distance between *Sadness* and *Anticipation* is 3. We provide the CMC curves on the entire VideoEmotion dataset in **Figure 4.8(a)** and on the VideoEmotion subset in **Figure 4.8(b)**. As stated earlier, with the use of these distances, classification accuracies achieved on both datasets are revealed to be higher, when the acceptable threshold for predictions is set to 2.

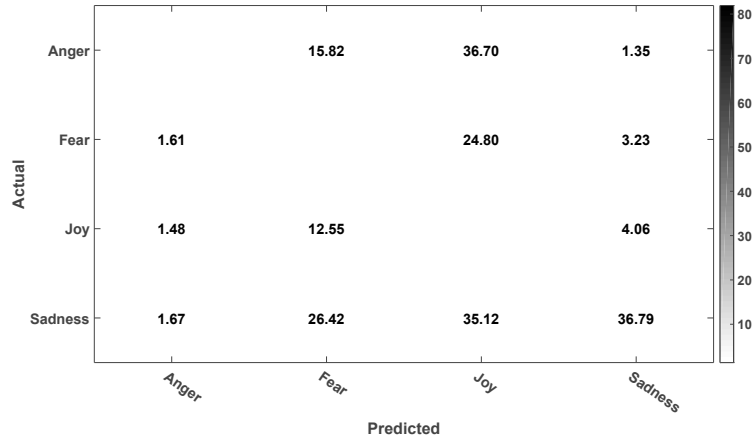
Comparison against the State-of-the-Art. As a final evaluation for the VA-based and wheel-based classification, we compare the classification accuracy of the framework against the state-of-the-art. **Table 4.7** provides the VA-based classification accuracies of our method (ensemble learning using MLR audio, motion, color and domain-specific representations linearly fused at the decision-level) compared to the works [5], [6] and [145] to position our approach in relation to these prior approaches (evaluation pertaining to the research question *RQ1*). Our method outperforms these works by achieving 81.08% accuracy. The paired Student t-test on classification accuracies from the cross validation showed that the improvement over the works [5] and [6] is statistically significant at the 5% significance level. As before, normal distribution and equal variance checks were performed using the *Jarque-Bera test* [66] and the *F-test*, respectively.

The results in Table 4.7 demonstrate the potential of our approach for video affective content analysis. As already discussed in detail in Section 3.5, the experimental setup employed in this chapter differs from the one of the work presented in [145] (e.g., different subset from the DEAP dataset).

Table 4.8 provides the wheel-based classification accuracies of our method compared to the works [70] and [97] to position our approach in relation to these prior



(a)



(b)

Figure 4.6: Confusion matrices for the *wheel-based classification* (a) on the entire VideoEmotion dataset and (b) on the VideoEmotion subset with the best performing multi-modal audio-visual representation (MLR audio, motion and domain-specific representations) using linear fusion. Darker areas along the main diagonal correspond to better discrimination. Mean accuracy: (a) 49.19%, (b) 63.65%.

approaches (addressed research question *RQ1*). Our method outperforms them by achieving 49.19% and 63.75% accuracies for the entire VideoEmotion dataset and the VideoEmotion subset, respectively.

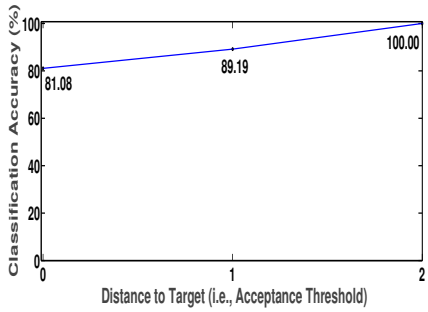
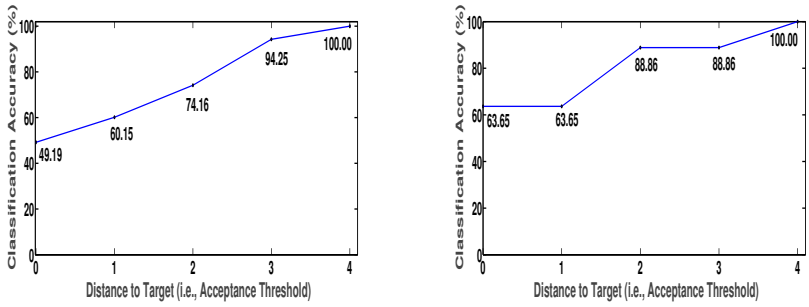


Figure 4.7: Cumulative Matching Characteristic (CMC) curve for the *VA-based classification* on the DEAP dataset with the best performing multi-modal audio-visual representation (MLR audio, motion, color and domain-specific representations). Fusion method is linear fusion.



(a) (b)

Figure 4.8: Cumulative Matching Characteristic (CMC) curves for the *wheel-based classification* (a) on the entire VideoEmotion dataset and (b) on the VideoEmotion subset with the best performing multi-modal audio-visual representation (MLR audio, motion and domain-specific representations) using linear fusion.

Table 4.7: *VA-based classification* accuracies on the DEAP dataset with audio-visual representations. *MLR audio-visual features*: MLR audio, motion, color and domain-specific.

Method	**Accuracy (%)**
Our method *(MLR audio-visual features & ensemble learning & linear fusion)*	81.08
Acar et al. (MLR audio, color and motion & SVM learning and fusion) [6]	66.22
Acar et al. (MLR audio and color & SVM learning and fusion) [5]	54.05
Yazdani et al. [145]	36.00

Table 4.8: *Wheel-based classification* accuracies of our method and of the works by Jiang et al. [70] and Pang et al. [97] on the VideoEmotion dataset with audio-visual representations.

Method	Accuracy – Entire (%)	Accuracy – Subset (%)
Our method *– MLR audio-visual features & ensemble learning & linear fusion*	49.19	63.75
Jiang et al. [70]	46.10	60.50
Pang et al. [97]	37.10	-

4.3.3.3 Summary of evaluation results: The bottom line

In this section, we identify the common denominator of the evaluation results on both professionally edited and user-generated videos, i.e., the findings consistent across both datasets. This also enables us to answer the research questions posed in Section 4.1. First of all, similar to the experimental results in Chapter 3, the experiments on both datasets have shown that learned audio and color representations are more discriminative than handcrafted low and mid-level representations for the user-generated videos as well.

- Regarding *RQ1* and *RQ3*, we explored the modeling perspective of video affective content analysis. Our findings from the perspective of modeling are as follows:

  1. As a result of the extensive experiments conducted on both datasets, ensemble learning (decision tree based bagging) is superior to SVM-based learning in emotion modeling of videos.
  2. Fusing the outputs of ensemble learning models using a simpler fusion method (linear fusion) has proven to be more effective than an advanced fusion mechanism (SVM-based fusion).

- Regarding *RQ2*, we investigated the discriminative power of uni and multimodal representations. Our findings from the perspective of feature representation are as follows:

  1. When considering unimodal representations, we observe that SentiBank and dense trajectory-based motion representations are the most discriminative features for emotional content analysis of both profesionnaly edited and user-generated videos.
  2. For SentiBank, using 2,089 trained visual concept detectors instead of 1,200 provides a noticeable increase in terms of classification performance. We closely looked at the importance of individual ANPs to check if some of them were consistently playing an important role across all videos. There appears to be no ANP which clearly stands out and is represented in all videos of both datasets. Therefore, we can only conclude

that increasing the number of ANPs helps increasing classification accuracies.

4.4 Conclusions

In this chapter, we presented a promising approach for the affective labeling of professionally edited and user-generated videos using mid-level multi-modal representations and ensemble learning.

We concentrated both on the representation and modeling aspects. Concerning the aspects relating to representation, higher level representations were learned from raw data using CNNs and fused with dense trajectory based motion and SentiBank domain-specific features at the decision-level. As a basis for feature learning, MFCC was employed as audio feature, while color values in the HSV space formed the static visual features. Concerning the aspects relating to modeling, we applied ensemble learning, *viz.* decision tree based bagging, to classify each video into one of the predefined emotion categories.

Experimental results on the VideoEmotion dataset and on a subset of the DEAP dataset support our assumptions (1) that learned audio-visual representations are more discriminative than handcrafted low-level and mid-level ones, (2) that including dense trajectories and SentiBank representations contribute increasing the classification performance, and (3) that ensemble learning is superior to multi-class SVM for video affective content analysis. In addition, we have demonstrated that fusing the outputs of ensemble learning models using a simpler fusion method (linear fusion) is more effective than an advanced fusion mechanism (SVM-based fusion).

As regards the modeling aspects, one potential improvement is to experiment more advanced classifiers instead of decision trees in the ensemble learning framework, such as SVMs [106].

The works we presented in Chapters 3 and 4 concentrated on affective content analysis based on the actual audio-visual content that is being analyzed (an analysis from a data perspective). Adopting an analysis from a user perspective such as collaborative user interaction and context evaluation (a scenario in which users can indicate their audio-visual content preferences and rate them) could bring adaptability to the framework, and constitutes, therefore, another important future direction.

In this part of the dissertation, the focus was on affective content analysis of multimedia data items in general. In the second part, we address a specific case, namely the detection of violent (disturbing) scenes in movies and user-generated videos.

Part II

Violent Content Detection: A Special Case of Affective Content Analysis

5

DETECTING VIOLENT CONTENT IN MOVIES: A PERSPECTIVE ON REPRESENTATION

As emotions play an important role in multimedia content selection and consumption, enabling search for movies (or videos in general) containing a particular expected emotion is a decisive functionality. In this chapter, we address the problem of detecting "violent" (i.e., "disturbing") scenes in movies at the video shot level, which can be regarded as a special case of emotional analysis of videos. The solution to the addressed problem constitutes an important video content understanding tool e.g., for providing automated youth protection services. Violent video content detection is highly challenging due to the complex nature of videos which contain both audio and visual information, and due to the possible diverse forms under which the concept of violence can be expressed. Therefore, combining features from different modalities is an effective strategy for violent content analysis in videos. In this chapter, we focus on the video representation part of the violence detection problem and perform a comprehensive analysis of audio-visual representations at different abstraction levels (low and mid-level). To this end, we employ audio, color, texture, motion and attribute representations for violent content analysis. The use of multiple features inevitably raises the question of fusion, which we also address. Experiments are performed on the MediaEval 2013 Violent Scenes Detection (VSD) dataset. Concerning descriptors, the results show that the best performing uni-modal descriptor is the dense trajectory based motion descriptor.

Concerning fusion, combining features from different modalities (audio, color, texture and motion) at the decision level using linear fusion helps improving the classification performance of the system. Without any post-processing such as temporal violence score smoothing or segment merging, the system achieves promising results when compared to the state-of-the-art results. The work presented in this chapter has been published in [3, 4, 86].

5.1 Introduction

As the amount of available multimedia content becomes more and more abundant, the use of automatic multimedia analysis solutions in order to find relevant semantic search results or to identify illegal content present on the World Wide Web has reached a critical importance. In addition, the advances in digital media management techniques have facilitated delivering digital videos to consumers. Therefore, accessing online movies through services such as VOD has become extremely easy. As a result, parents are not able to constantly and precisely monitor what their children watch. Children are, consequently, exposed to movies, documentaries, or reality shows which have not necessarily been checked by parents, and which might contain inappropriate content. Violence constitutes one example of such inappropriate content. Psychological studies have shown that violent content in movies has harmful impacts, especially on children [23]. As a consequence, there is a need for automatically detecting violent scenes in videos, where the legal age ratings are not available.

Like any other research challenge, tackling the problem of violence detection begins with establishing a framework, in particular adopting a definition of violence to work with. Since the concept of violence is highly subjective (i.e., person-dependent) – not everybody would indeed evaluate a particular scene of a movie as violent, one of the challenges within the context of multimedia violent content detection is to properly delimit the boundaries of what can be designated as a "violent" scene. In our work, we aim at sticking to the definition of violence as described in [37]: The subjective point of view. *Subjective violent* scenes are "those which one would not let an 8 years old child see because they contain physical violence".

Next to the issue of definition, another important step in the task of video violent content detection is the representation of video segments. In order to represent the complex heterogeneous content of movies (or videos in general), it is necessary to use multi-modal features of different abstraction levels describing different aspects of the movies. Within this context, solutions using mid-level feature representations have recently gained popularity. These solutions shifted away not only from the traditional approaches which represented videos using low-level features (e.g., [28, 53]) but also from the use of state-of-the-art detectors designed to identify high-level semantic concepts (e.g., "a killing spree"). The earlier solutions could not carry enough semantic information, and the latter ones have not reached a sufficient level of maturity. Hinted by these recent developments, inferring mid-

level abstract representations is more suitable than directly using low-level features in order to bridge the semantic gap between the features and the high-level human perception of violence. The use of mid-level representations, therefore, may help modeling video segments one step closer to human perception. In this chapter, we perform a comprehensive study of uni-modal features at different abstraction levels (including mid-level ones) which are used to represent videos from different perspectives. These are ***sound*** (through low and mid-level audio), ***action*** (through low and mid-level motion) and ***scene characteristics*** (through color, texture and attribute features). As a basis for the mid-level audio and motion representations, we employ MFCC and dense trajectory features, respectively. As attribute features, we use Classemes [121] which basically correspond to automatic detection scores of 2,659 semantic concepts (mainly objects and scenes) trained on images from the Web. To sum up, in this chapter, our aim is to answer the following research questions:

(RQ1) What is the discriminative power of uni-modal low-level and mid-level audio and visual representations? The experimental results of this study shed light on the significance of selected uni-modal representations in order to model violence in movies.

(RQ2) To which extent does the fusion of the uni-modal audio and visual representations improve the classification performance of the system? The results also shed light on the best feature combinations for the problem of violent content detection. We experimentally show that different combinations of uni-modal descriptors with linear fusion always improves the classification performance of the system and the best performance is achieved when audio and visual representations at different abstraction levels (low-level or mid-level) are linearly fused. Besides, the best classification performance of the introduced framework – without any post-processing such as temporal violence score smoothing or segment merging – is promising when compared to the state-of-the-art results.

The chapter is organized as follows. Section 5.2 gives an overview of the framework. In Section 5.3, we present audio representations constructed at different levels. We discuss static and dynamic visual features that we evaluate in Sections 5.4 and 5.5, respectively. The violence modeling aspect of the system is introduced in Section 5.6. We provide and discuss evaluation results obtained on the MediaEval 2013 VSD dataset consisting of Hollywood movies in Section 5.7. Concluding remarks and future directions to expand our current approach are presented in Section 5.8.

5.2 Framework Overview

We first present an overview of the framework to evaluate the performance of uni-modal and multi-modal analysis for violent scenes detection in movies. As shown in Figure 5.1, the first phase of the system is feature generation. In this phase, audio

and visual features are extracted for each video shot of movies. For audio feature construction, the whole audio signal of the video shots is used. Only the keyframes of the video shots are processed for the extraction of static visual features, whereas all frames constituting the video shots are used in the dynamic visual feature extraction. The details of the feature generation are explained in Sections 5.3 (audio), 5.4 (static visual) and 5.5 (dynamic visual). In the modeling phase of the system, audio and visual analysis models are built using the uni-modal features and the built models are used in the test phase to predict the violence scores of the video shots. The details of model generation and prediction are discussed in Section 5.6.

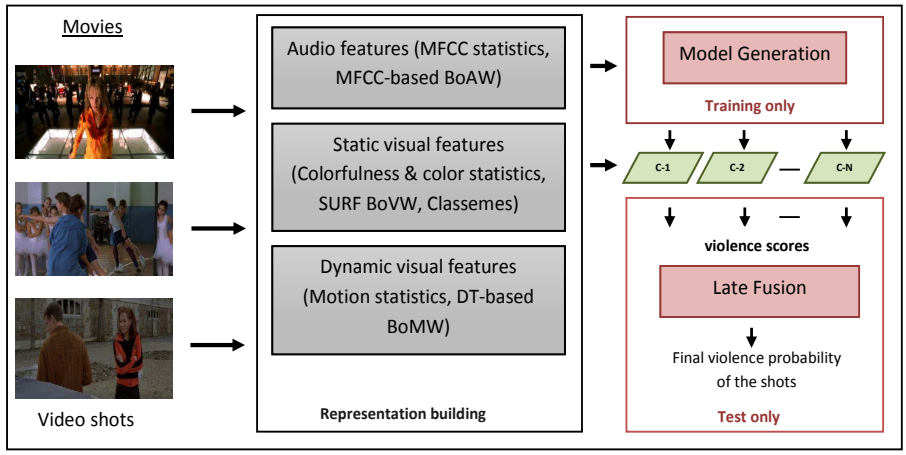


Figure 5.1: An overview of the violent scenes detection framework. One classifier is built for each uni-modal feature, i.e., N = 7. (BoAW: Bag-of-Audio-Words, BoMW: Bag-of-Motion-Words, BoVW: Bag-of-Visual-Words, C: classifier, DT: dense trajectory, SURF: Speeded-Up Robust Features)

5.3 Audio Features

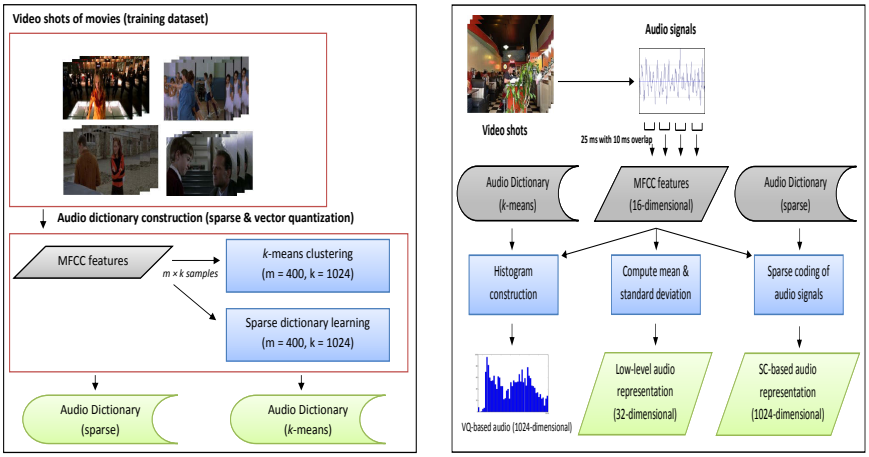
Sound effects and background music in movies are essential elements used by film editors for stimulating people's perception [126]. Therefore, audio signals are an important data source for the representation of movies. Among the plurality of available audio descriptors, MFCC features are shown to be indicators of the excitement level of video segments [137]. For this reason, we employ MFCC extracted from the audio signal of video shots as base audio features as illustrated in Figure 5.2. We represent the audio content of videos at two different levels: Low-level and mid-level, which we discuss in detail in Sections 5.3.1 and 5.3.2, respectively.

5.3.1 Low-level representation

Due to the variability in duration of video shots annotated as violent or non-violent, each shot comprises a different number of MFCC feature vectors. Aiming at constructing low-level audio representations having the same dimension, mean and standard deviation for each dimension of the MFCC feature vectors are computed. The resulting statistics of the MFCC features compose the low-level audio representations of the shots.

5.3.2 Mid-level representation

In order to construct mid-level audio representations for video shots, we apply an abstraction process which uses an MFCC-based Bag-of-Audio-Words (BoAW) approach. We implement this approach using two different coding schemes, namely vector quantization (VQ) and sparse coding (SC). The feature generation process is illustrated in Figure 5.2. The details of the VQ-based method is presented in Section 5.3.2.1, whereas Section 5.3.2.2 discusses the SC-based one.



(a) (b)

Figure 5.2: (a) The generation of two different audio dictionaries: One by using vector quantization (VQ) and another dictionary for sparse coding (SC). (b) The generation process of VQ-based, SC-based and low-level audio representations for video shots of movies.

5.3.2.1 Vector quantization

The vector quantization based method consists of two phases: (1) Dictionary construction and (2) feature encoding. Figure 5.2 illustrates the construction of the VQ-

based audio dictionary. An unsupervised way is followed for the construction of the dictionary. First, MFCC feature vectors extracted from video shots in the development dataset are clustered with a k-means clustering [94], in which the centroid of each of the k clusters is treated as an audio word. For the dictionary construction, we sampled $400 \times k$ MFCC feature vectors from the development data. The number of feature vectors to sample was determined with preliminary evaluations.

In the feature encoding phase of the method, once an audio vocabulary of size k ($k = 1024$ in this work) is built, each MFCC feature vector is assigned to the closest audio word in terms of Euclidean distance. Subsequently, a BoAW histogram representing the audio word occurrences is computed for each video shot in the development and test datasets.

5.3.2.2 Sparse coding

Similarly to the VQ-based method, the sparse coding based method also contains the two phases of dictionary learning and feature encoding. Figure 5.2 presents the construction of the SC-based audio dictionary, which is the first phase of the method. We employ the dictionary learning technique presented in [84]. The advantage of this technique is its scalability to very large datasets with millions of training samples which makes the technique well suited for our work. In order to learn the dictionary of size k ($k = 1024$ in this work) for sparse coding, $400 \times k$ MFCC feature vectors are sampled from the training data. The figures were again determined experimentally with preliminary evaluations.

In the feature encoding phase, we construct the sparse representations of audio signals by using the LARS algorithm [46]. Given an audio signal and a dictionary, the LARS algorithm returns sparse representations for MFCC feature vectors. In order to generate the final sparse representation of video shots which are a set of MFCC feature vectors, we apply the *max-pooling* technique [143].

5.4 Static Visual Features

In addition to sound effects, scene characteristics of movies are also an important element for scenes in the movies. In order to represent those scene characteristics, color, texture and attribute features which basically describe static visual properties of the scenes are used in this work. We discuss these features in detail in Sections 5.4.1 and 5.4.2. The construction of static visual features is summarized in Figure 5.3.

5.4.1 Color and texture representation

Color properties constitute an effective and computationally efficient representation of the video frames for violence analysis. For instance, in movies, many violent scenes usually picture a dark environment or contain a lot of blood. As we regard

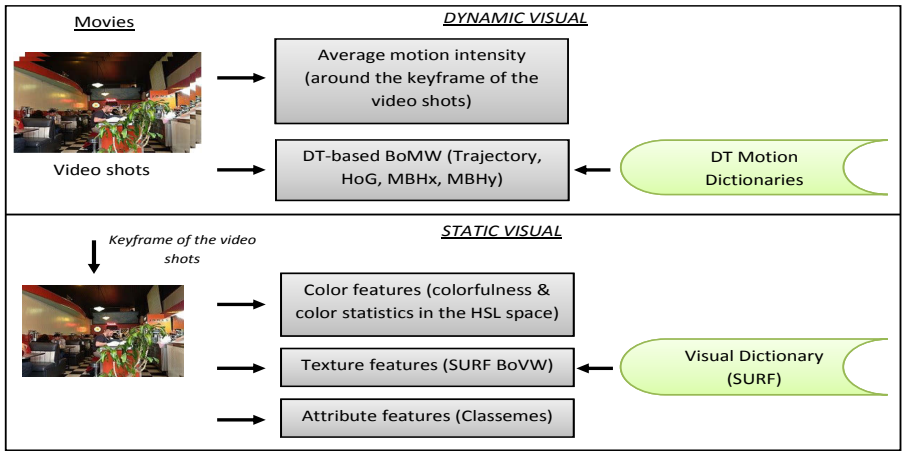


Figure 5.3: The generation process of static and dynamic visual representations for the video shots of movies. (BoMW: Bag-of-Motion-Words, BoVW: Bag-of-Visual-Words, DT: dense trajectory, HoG: Histogram-of-Oriented-Gradients, HSL: Hue-Saturation-Lightness, MBHx & MBHy: motion boundary histograms in the x and y directions, SURF: Speeded-Up Robust Features)

violence analysis as a special case of affective content analysis, we use colorfulness [58] and color statistics in the HSL color space (mean and standard deviation of each color channel) which are commonly used features for the emotional analysis of images and videos (e.g., [83, 151]).

In addition to color, texture properties are also an important cue to describe the content of the videos. The dense Speeded-up robust features (SURF) method [18] is used in this work to describe the textural properties of scenes. Bag-of-Words representations of dense SURF features (Bag-of-Visual-Words – BoVW) are generated using vector quantization. The dictionary construction and feature encoding phases are performed similarly to the audio case (Section 5.3.2.1). A visual dictionary of size k (k is equal to 1024) is constructed using $400 \times k$ SURF descriptors and k-means clustering. In the feature encoding phase, the SURF feature vectors of the keyframe of video shots are assigned to the closest visual word in terms of Euclidean distance. Subsequently, a BoVW histogram representing the visual word occurrences is computed for each video shot in the development and test datasets.

5.4.2 Attribute representation

In order to present scene information of videos at a higher level (higher when compared to the color statistics and the BoVW representation presented in Section 5.4.1), we use *Classemes* [121] attribute features. Classemes is a global attribute descriptor

generated by models trained on images retrieved by the *Bing Image Search*[1] engine. Since it is a global descriptor, a frame is treated and processed as a whole to determine the semantic concepts existing in the frame. Basically, the Classemes descriptor consists of the detection scores of 2,659 semantic concepts (mainly objects and scenes), where each dimension of the descriptor corresponds to one of these semantic categories[2].

5.5 Dynamic Visual Features

Motion captures the temporal variation of videos and is another important visual element for film editors to elicit some particular perception in the audience. The importance of motion in edited videos (e.g., movies and Web videos) motivated us to incorporate motion information to our framework. The dynamic visual content of videos provides complementary information and is commonly used for the detection of violence in videos (e.g., [42, 53, 75]). For instance, the amount of activity is proven to be strongly correlated with the existence of violence [53]. Similarly to the audio case, we represent the dynamic visual content of videos at two different levels: Low-level and mid-level, which we discuss in detail in Sections 5.5.1 and 5.5.2, respectively. The construction of dynamic visual features is also summarized in Figure 5.3.

5.5.1 Low-level motion representation

Average motion intensity is a commonly used low-level motion feature to analyze violence in videos (e.g., [29, 53]). In order to characterize the degree of motion of video shots, we compute the normalized average motion intensity [81] of the video shots based on motion vectors. The motion vectors are computed by block-based estimation and then the normalized average motion intensity for the k^{th} frame of a video shot is deduced as illustrated in Equation 5.1.

$$avgMotion_k = \frac{\sum_{i=1}^{I} mv_k(i)}{I \cdot MV_k} \tag{5.1}$$

where $mv_k(i)$ is the magnitude of the motion vector of the i^{th} block, I is the total number of motion vectors, and MV_k is the magnitude of the largest motion vector of the frame. In this work, this process is performed only in the vicinity of the keyframe of the video shots. More specifically, average motion values are computed between a given keyframe and its preceding and succeeding frames, respectively.

[1] `http://www.bing.com/?scope=images`

[2] `http://www.cs.dartmouth.edu/~lorenzo/projects/classemes/classeme_keywords.txt`

5.5.2 Mid-level motion representation

As mid-level motion representation, we adopt the work of Wang et al. on dense trajectories [127]. Improved dense trajectories are dynamic visual features which are derived from tracking densely sampled feature points in multiple spatial scales. Although initially used for unconstrained video action recognition [127], dense trajectories constitute a powerful tool for motion or video description, and, hence, are not limited to action recognition.

Our dynamic visual representation works as follows. First, dense trajectories of length L ($L = 15$ in this work) frames are extracted from each video shot. Dense trajectories are subsequently represented by a Trajectory, a histogram of oriented gradients (HoG) and motion boundary histograms in the x and y directions (MBHx and MBHy). The histogram of oriented optical flow (HoF) descriptor of dense trajectories – part of the "bundle" of features presented in [127] – is excluded, since it demonstrated a poor performance in our preliminary evaluations. The sparse dictionary learning and coding phases are performed similarly to the audio case. For each dense trajectory descriptor (Trajectory, HoG, MBHx and MBHy), we learn a separate dictionary of size k ($k = 1024$) by sampling $400 \times$ k feature vectors from the development data. Finally, sparse representations are constructed using the LARS and *max-pooling* algorithms.

5.6 Violence Modeling and Prediction

Two-class SVMs are trained for each uni-modal descriptor presented in Sections 5.3, 5.4 and 5.5. More specifically, we train a two-class SVM for each of the following descriptors: **(1)** Low-level audio descriptor (MFCC statistics), **(2)** mid-level audio descriptor (MFCC-based BoAW), **(3)** colorfulness and color statistics, **(4)** mid-level static visual descriptor (SURF-based BoVW), **(5)** attribute descriptor (Classemes), **(6)** low-level motion descriptor (average motion intensity), and **(7)** mid-level motion descriptor (dense trajectory based BoMW).

In the learning phase, the main issue is the problem of imbalanced data. In the development dataset, the number of non-violent video shots is much higher than the number of violent ones. This results in the learned boundary being too close to the violent instances. Consequently, the SVM tends to classify every sample as non-violent. Different strategies to "push" this decision boundary towards the non-violent samples exist. Although more sophisticated methods dealing with the problem of imbalanced data have been proposed in the literature (see [60] for a comprehensive survey), we choose, in the current framework, to perform random undersampling to balance the number of violent and non-violent samples (with a balance ratio of 1:2). This method proposed by Akbani et al. [10] appears to be particularly adapted to the application context of our work. In [10], different under- and oversampling strategies are compared. According to the results, SVM with the undersampling strategy provides the most significant performance gain over stan-

dard two-class SVMs. In addition, the efficiency of the training process is improved as a result of the reduced training data. Hence, this method is scalable to large datasets similar to the ones used in the context of our work.

In the test phase of the framework, the fusion of the predictions of uni-modal analysis models is performed at the decision level with a linear combination of uni-modal violence scores applied with weights optimized on the development set.

5.7 Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of audio and visual features at different levels (low and mid-level) for the detection of violent content in movies. We also investigate the extent to which the fusion of the uni-modal features further improves the classification performance of the system. A direct comparison of our results with other works discussed in Section 2.4 could be misleading due to the discrepancies in the definition of "violence" in published works. However, we can reliably compare our method with the methods of the participants to the the MediaEval VSD task of 2013 as they also stick to the same "violence" definition and are evaluated on the same dataset.

5.7.1 Dataset and ground-truth

The MediaEval 2013 VSD dataset[3] consists of 46,525 video shots from 24 Hollywood movies of different genres (ranging from extremely violent movies to movies without violence), where each shot is labeled as violent or non-violent. The data set is divided into a development set consisting of 35,280 shots from 17 movies and a test set consisting of 11,245 shots from the remaining 7 movies. The movies of the development and test set were selected in such a manner that both development and test data contain movies of variable violence levels (extreme to none). On average, around 11.57% and 20.24% of the shots are annotated as violent in the development and test sets, respectively. Tables 5.1 and 5.2 give more details about the main characteristics of the datasets.

The ground-truth[4] was generated by 7 human assessors, partly by developers, partly by possible users. Violent movie segments are annotated at the frame level (i.e., violent segments are defined by their starting and ending frame numbers). Automatically generated shot boundaries with their corresponding key frames are also provided for each movie. A detailed description of the dataset and the ground-truth generation are given in [40].

[3] `http://www.technicolor.com/en/innovation/research-innovation/scientific-data-sharing/violent-scenes-dataset`

[4] Annotations were made available by *Fudan University*, *Vietnam University of Science*, and *Technicolor*.

Table 5.1: The characteristics of the MediaEval 2013 VSD development dataset (movie length in seconds, the number of video shots and the percentage of violent segments per movie).

#	Video Title	Length (sec)	Video Shot#	Violence (%)
1	*Armageddon*	8,680	3,562	07.77
2	*Billy Elliot*	6,349	1,236	02.45
3	*Dead Poets Society*	7,413	1,583	00.58
4	*Eragon*	5,985	1,663	13.25
5	*Fight Club*	8,005	2,335	15.82
6	*Harry Potter V*	7,953	1,891	05.43
7	*I am Legend*	5,779	1,547	15.64
8	*Independence Day*	8,875	2,652	13.12
9	*Leon*	6,344	1,547	16.34
10	*Midnight Express*	6,961	1,677	07.11
11	*Pirates of the Caribbean*	8,239	2,534	18.14
12	*Reservoir Dogs*	5,712	856	30.40
13	*Saving Private Ryan*	9,751	2,494	33.96
14	*The Bourne Identity*	6,816	1,995	07.18
15	*The Sixth Sense*	6,178	963	02.00
16	*The Wicker Man*	5,870	1,638	06.44
17	*The Wizard of Oz*	5,859	908	01.02
–	**Total**	120,769	35,280	11.57

Table 5.2: The characteristics of the MediaEval 2013 VSD test dataset (movie length in seconds, the number of video shots and the percentage of violent segments per movie).

#	Video Title	Length (sec)	Video Shot#	Violence (%)
1	*Fantastic Four I*	6,094	2,002	35.81
2	*Fargo*	5,646	1,061	24.13
3	*Forrest Gump*	8,177	1,418	16.78
4	*Legally Blond*	5,523	1,340	00.00
5	*Pulp Fiction*	8,887	1,686	29.42
6	*The God Father I*	10,195	1,893	10.46
7	*The Pianist*	8,567	1,845	20.11
–	**Total**	53,089	11,245	20.24

5.7.2 Experimental setup

We first start by explaining the pre-processing steps performed before any video processing was applied. This begins with a conversion of the DVD content to mpg

videos in MPEG-2 format by using DVDFab version 8[5] and VOB2MPG version 3[6]. Once the mpg videos were generated, we resized the videos to a width of 640 pixels and scaled the height to keep the same aspect ratio. Feature extraction is performed on these pre-processed videos. Table 5.3 gives an overview of the features extracted to represent video shots and the details of feature extraction are discussed in the following paragraphs.

Table 5.3: An overview of extracted audio, static and dynamic visual features in the framework. (LLR: low-level representation, MLR: mid-level representation, SC: sparse coding, VQ: vector quantization).

Name	**Modality**	**Dimension#**
LLR audio	audio	32
MLR audio-VQ	audio	1,024
MLR audio-SC	audio	1,024
LLR color	static visual	7
MLR texture	static visual	1,024
MLR attribute	static visual	2,659
LLR motion	dynamic visual	2
MLR motion	dynamic visual	4,096

The MIR Toolbox v1.6.1[7] was employed to extract 16-dimensional MFCC feature vectors. Frame sizes of 25 ms with a step size of 10 ms were used. MFCC statistics (referred hereafter as *LLR audio*) and MFCC BoAW representations with the two different coding schemes (referred hereafter as *MLR audio*) were computed. The details of the audio feature extraction are explained in Section 5.3.

The keyframes provided by the MediaEval VSD task organizers were used to extract static visual features (color statistics including colorfulness, SURF BoVW and Classemes). The details of the static visual feature extraction are explained in Section 5.4. The VLG extractor[8] was used to extract Classemes features of the frames. We refer to the color statistics, SURF BoVW and Classemes hereafter as *LLR color*, *MLR texture* and *MLR attribute*, respectively.

Improved dense trajectories were extracted using the software[9] provided by Wang et al. [127]. The sampling stride, which corresponds to the distance by which extracted feature points are spaced, is set to 20 pixels due to time efficiency concerns. The remaining details of the dynamic visual feature extraction are provided in Section 5.5. Average motion intensity and dense trajectory based BoMW representations are hereafter referred to as *LLR motion* and *MLR motion*, respectively.

[5] http://de.dvdfab.cn/

[6] http://www.svcd2dvd.com/VOB2MPG/

[7] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[8] http://vlg.cs.dartmouth.edu/projects/vlg_extractor/vlg_extractor/

[9] https://lear.inrialpes.fr/people/wang/improved_trajectories

We employed the SPAMS toolbox[10] in order to compute sparse codes which are used for the generation of mid-level audio and dynamic visual representations. The VLFeat[11] open source library was used to perform *k*-means clustering for vector quantization.

Once audio, static and dynamic visual features were extracted at the video shot level, two-class SVMs with a Radial Basis Function (RBF) kernel were trained using libsvm[12] as the SVM implementation. Training was performed separately for each audio or visual descriptor extracted at the video shot level. The details of model generation are provided in Section 5.6. SVM hyper parameters were optimized using a grid search and 5-fold cross-validation using the development set. The search range for the γ parameter was [3, −15] with a step size of −2, whereas the range for the C parameter was [−5, 15] with a step size of 2. All segmented samples belonging to a specific movie were always either in the training set or in the test set during cross-validation. Our approach was evaluated using a training-test split using the split explained in detail in Section 5.7.1. In order to account for the problem of imbalanced training data, we performed undersampling by choosing random non-violent samples (Section 5.6).

5.7.3 Evaluation metrics

Precision and recall are metrics based on the results obtained for the whole list of video shots of the movies. Metrics other than precision and recall are, however, required to compare the performance of the representations, since the ranking of violent shots is more important for the use case described in Section 5.1 (providing a ranked list of violent video shots to the user). As evaluation metrics, therefore, we use the *mean average precision at 100* (MAP at 100 – MAP@100) which is also the official metric used in the MediaEval 2013 VSD task. Basically, the metric MAP at *N* is the mean of average precision over the top *N* best ranked violent shots (with *N* = 100 for MAP@100) for each movie in the test set. The AP@100 metric is basically the average precision at the 100 top-ranked violent shots over the movies in the test dataset [40]. The value 100 for the computation of the MAP at *N* is reasonable, since a user will only have a look at the video shots that are presented in the first few pages of the returned list.

In addition to the official metric, we use the detection error trade-off (DET) curves that enable us to avoid the sole comparison of the systems at given operating points. The uni-modal and multi-modal analysis used in this work are compared by plotting P_{fa} as a function of P_{miss} given a segmentation, where P_{fa} and P_{miss} are the estimated probabilities of false alarm (false positive) and missed detection (false negative) given the system's output and the ground-truth. The false alarm and miss probabilities are calculated on a per shot basis.

[10] `http://spams-devel.gforge.inria.fr/`
[11] `http://www.vlfeat.org/`
[12] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

5.7.4 Results and discussions

The evaluation of our framework is performed from different perspectives: *(i)* Uni-modal violence modeling; *(ii)* multi-modal violence modeling; and *(iii)* comparison to MediaEval 2013 participants. These perspectives are discussed in detail in Sections 5.7.4.1, 5.7.4.2 and 5.7.4.3, respectively. In Section 5.7.4.4, we provide a summary of the evaluation results.

5.7.4.1 Uni-modal analysis

This section provides an evaluation of our framework from perspective *(i)*, i.e., uni-modal violence modeling. More specifically, we present and discuss the discriminative power of MFCC-based statistics, MFCC-based BoAWs, color statistics, SURF-based BoVW, Classemes and dense trajectory based BoMW features. The comparison of the uni-modal representations is presented in Table 5.4.

Table 5.4: The MAP@100 for the uni-modal representations on the MediaEval 2013 VSD test dataset. ***LLR audio***: MFCC-based audio statistics, ***MLR audio-VQ***: vector quantization based MFCC BoAW, ***MLR audio-SC***: sparse coding based MFCC BoAW, ***LLR color***: color statistics, ***MLR texture***: SURF-based BoVW, ***MLR attribute***: Classemes, ***LLR motion***: average motion intensity, ***MLR motion (Trajectory, HoG, MBHx & MBHy)***: dense trajectory based BoMW with Trajectory, HoG and $MBH_{x,y}$ descriptors. The best performing uni-modal descriptors are highlighted in bold for each feature category (audio, static visual and dynamic visual). (MAP@100: mean average precision at 100, Dynamic: dynamic visual features, Static: static visual features).

	Uni-modal Representation	MAP@100 (%)
Audio	*LLR audio*	35.00
	MLR audio-VQ	45.43
	MLR audio-SC	**46.57**
Static	*LLR color*	41.57
	MLR texture	28.76
	MLR attribute	**42.65**
Dynamic	*LLR motion*	34.81
	MLR motion (Trajectory, HoG, $MBH_{x,y}$)	**48.89**

Various observations can be inferred based on the overall results presented in Table 5.4.

- The first observation is that the best performing uni-modal representation is the *MLR motion* (these basically are dense trajectory based BoMW features using trajectory, HoG and $MBH_{x,y}$ descriptors), whereas the *MLR texture* (dense SURF-based BoVW) has the poorest classification performance.

- When we analyze the best classification performances according to feature categories (audio, static visual and dynamic visual), the dynamic visual descriptors provide the best performance; this category is followed by the audio and static visual categories respectively.

- The abstraction process applied on MFCC feature vectors provides an important increase in classification performance. Besides, the sparse coding scheme leads to better results compared to the vector quantization one.

Table 5.5: The AP@100 (%) for the best performing uni-modal representations per feature category (audio, static visual and dynamic visual) on each movie of the MediaEval 2013 VSD test dataset. (AP@100: average precision at 100, A: audio representing *SC-based MLR audio*, SV: static visual representing *MLR attribute*, DV: dynamic visual representing *MLR motion*).

Movie	**A – AP@100**	**SV – AP@100**	**DV – AP@100**
Fantastic Four I	53.70	62.68	65.69
Fargo	31.90	20.56	45.79
Forrest Gump	38.40	28.47	48.99
Legally Blond	00.00	00.00	00.00
Pulp Fiction	84.05	88.20	73.39
The God Father I	90.70	26.77	43.18
The Pianist	27.24	71.87	65.19

In order to have a closer look at the results and to evaluate the performance of uni-modal representations for movies of different violence levels, average precision at 100 (AP@100) for the best performing uni-modal representations per feature category (audio, static visual and dynamic visual) on each movie of the MediaEval 2013 VSD test dataset are presented separately in Table 5.5. When we evaluate the AP@100 values per movie, we see that the performance of the framework is the worst, when there is no violence in the movie that is being analyzed (*Legally Blond* in our case). This result is actually on par with the results of the teams participating to the MediaEval VSD task of 2013 [88]. For some movies (e.g., *The God Father I*), the audio modality is a better perspective than the visual ones (static or dynamic) to describe violence in the movies, whereas the opposite is true for some others (e.g., *The Pianist*). In addition, we observe that for some movies (e.g., *Pulp Fiction*), both audio and visual modalities lead to promising results. This variance over movies suggests that the expression of violence in movies is diverse and may

vary for different movies. Therefore, considering multi-modal features to represent movie segments is a more judicious choice.

5.7.4.2 Multi-modal analysis

This section evaluates our framework from perspective *(ii)*, i.e., multi-modal violence modeling. In other words, we analyze the classification performance of the framework when multi-modal features are used to describe video content. Table 5.6 presents the classification performances of the fusion of different audio and visual features.

Table 5.6: The MAP@100 for the multi-modal modeling on the MediaEval 2013 VSD test dataset. ***MLR audio-SC***: sparse coding based MFCC BoAW, ***MLR attribute***: Classemes, ***MLR motion (Trajectory, HoG, MBHx & MBHy)***: dense trajectory based BoMW with Trajectory, HoG and $MBH_{x,y}$ descriptors. (MAP@100: mean average precision at 100, SC: sparse coding).

Multi-modal Representation	**MAP@100 (%)**
All features (in Table 5.4)	62.58
Best of all feature categories (according to Table 5.4)	58.32
MLR attribute, MLR motion (Trajectory, HoG, $MBH_{x,y}$)	55.21
MLR audio-SC, MLR motion (Trajectory, HoG, $MBH_{x,y}$)	51.09
MLR audio-SC, MLR attribute	48.20

From Table 5.6, we can derive the following observations concerning the different combinations of uni-modal representations.

- The first observation is that the performance of the framework is at its best when we fuse the violence scores of all uni-modal models, regardless of their individual performance in uni-modal classification.

- In addition to using all the uni-modal models, we evaluated the different binary combinations of the best performing descriptors per feature category (audio, static and dynamic visual). From the results, we observe that every fusion of pairs of multi-modal features performs better than each of the corresponding two uni-modal modelings of the pairs.

- The first two observations suggest that uni-modal features representing a video from different perspectives (audio, color, texture, attribute and motion) provide complementary information about the content of the video.

As in the previous section (Section 5.7.4.1), a closer look at the AP@100 values is provided in Table 5.7, where the results of the two best performing multi-modal representations on each movie of the MediaEval 2013 VSD test dataset are

presented. When Tables 5.5 and 5.7 are assessed together, we observe the improvements in classification performance in terms of AP@100 values. Except for one movie (*The God Father I*), the AP@100 values for each movie in the test dataset are improved. This figure suggests that considering different perspectives of movies such as sound (through MLR audio), scene characteristics (through MLR attribute) and action (through MLR motion) to represent them during violent content analysis lead to more discriminative systems regardless of the violence level of the movies.

Table 5.7: The AP@100 (%) for the two best performing multi-modal representations on each movie of the MediaEval 2013 VSD test dataset. (AP@100: mean average precision at 100, All: multi-modal analysis with all features, BestPerCategory: multi-modal analysis with all best performing features per feature category).

Movie	**All – AP@100**	**BestPerCategory – AP@100**
Fantastic Four I	72.75	68.04
Fargo	60.22	58.13
Forrest Gump	65.73	61.30
Legally Blond	06.12	05.01
Pulp Fiction	92.80	89.23
The God Father I	61.55	50.52
The Pianist	78.89	76.01

As another evaluation of the performance of the uni-modal and multi-modal modeling, Figure 5.4 provides the Detection Error Trade-off (DET) curves of the best performing uni-modal features of each feature category (audio, static visual and dynamic visual) and multi-modal features (all audio-visual features and best of all feature categories to be more specific). When we evaluate the curves of the representations in Figure 5.4, the two best solutions are (1) multi-modal modeling where we fuse all features presented in Table 5.4, and (2) multi-modal modeling where only the best performing feature per feature category is used in the classification process. In addition, the best performing uni-modal representation (in terms of DET) is the mid-level audio representation constructed using sparse coding (*MLR audio-SC*).

5.7.4.3 MediaEval comparisons

As a final evaluation of the framework, we compare the performance of our system against the best run of the MediaEval 2013 participants in terms of the MAP@100 official task metric (perspective *(iii)*).

Our approach takes the fourth place in terms of the MAP@100 official task metric, after the LIG, VIREO and Fudan teams. When we analyze the systems proposed by these three teams, we observe that all of them applied post-processing steps such as temporal re-ranking or score smoothing. This improves the performance

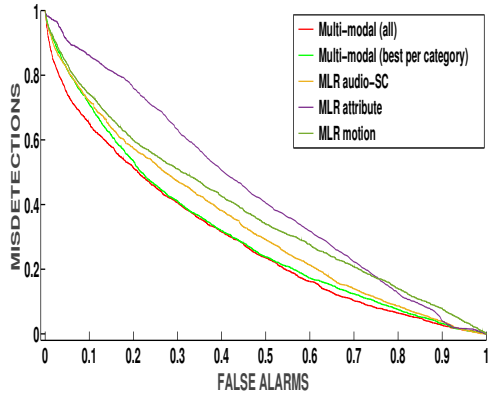


Figure 5.4: Detection Error Trade-off (DET) curves for the uni-modal (best performing ones of each feature category according to Table 5.4), all features (in Table 5.4) and best of all feature categories (according to Table 5.4). (MLR: mid-level representation, SC: sparse coding)

Table 5.8: The MAP@100 for the best run of teams in the MediaEval 2013 VSD Task [88] (the two best runs of the Fudan team are presented for a detailed comparison) and our best performing method on the MediaEval 2013 VSD test dataset. Teams with at least one participant member of the MediaEval VSD organizing team are marked by *.

Team	**MAP@100 (%)**
LIG (run 1) [42]	69.04
VIREO (run 4) [120]	68.96
Fudan (run 5) [33]*	68.16
Our approach with all features (in Table 5.4)	**62.58**
NII-UIT (run 1) [75]*	59.59
Fudan (run 4) [33]*	58.70
Technicolor/INRIA (run 1) [100]*	53.59

of the systems. For instance, the *run 4* of the Fudan team where no post-processing is applied reaches a MAP@100 value of 58.70% (which is below our score), whereas the *run 5* of the Fudan team is generated by applying a temporal score smoothing on the *run 4* and achieves 68.16%. This amounts to a performance gain of around 10%. As our aim was to concentrate on the discriminative power of uni-modal and multi-modal representations for violence detection in videos, we apply no post-processing on our results in this chapter and address this issue of post-processing in the following chapter (Chapter 6).

5.7.4.4 Summary of evaluation results: The bottom line

In this work, violent scenes were defined as "those which one would not let an 8 years old child see because they contain physical violence". Scenes falling within the scope of this "violence" definition are quite diverse. For example, both explosions and fight scenes are labeled as violent according to this definition. Hence, in order to differentiate such diverse scenes, it is necessary to represent the content of a video from different perspectives such as sound, action and scene characteristics. We, therefore, evaluated different audio and visual representations at different abstraction levels (low-level and mid-level) which match the aforementioned perspectives.

Regarding *RQ1*, we observed that the best performing uni-modal feature is the dense trajectory based BoMW representation which describes the action perspective of a video.

Regarding *RQ2*, the extensive evaluations in the previous two sections confirmed that multi-modal representations are necessary to achieve state-of-the-art performance for violent content detection in movies. Our system which makes use of multi-modal representations (audio, color, texture, attribute and motion) and no post-processing such as temporal violence score smoothing or segment merging achieved highly promising results when compared to the best systems of the MediaEval 2013 participants.

5.8 Conclusions

In this chapter, we concentrated on the representation perspective of violence modeling and presented a framework where an extensive set of audio and visual features at different abstraction levels are used to detect violence in movies at the video shot level. As audio features, we used MFCC statistics, MFCC-based BoAW with two different coding schemes (vector quantization and sparse coding). As static visual features, color statistics, dense SURF-based BoVW and Classemes have been extracted. Finally, the average motion intensity and dense trajectory based BoMW representations were constructed as dynamic visual characteristics of video shots. We evaluated the performance of these uni-modal features and the extent to which the fusion of the uni-modal audio and visual representations improves the classification performance of the system. Although we concentrated solely on the representation perspective, the system achieved a promising classification performance compared to the best systems of the MediaEval 2013 participants. The impact of these results is two-fold: The effective detection of violent content requires the use of advanced motion features, such as dense trajectory based motion representations. It also necessitates combining multiple features, e.g., in a fusion fashion.

Incited by the promising results of the framework presented in this chapter, we investigate the modeling perspective of the system (in Chapter 6) to consider the diverse nature of videos from a modeling perspective. In addition, our approach is

extended to user-generated videos. Different from Hollywood movies, these videos are not professionally edited, e.g., in order to enhance dramatic scenes. Therefore, they are usually more challenging for the task of violence analysis.

6

DETECTING VIOLENT CONTENT IN MOVIES AND USER-GENERATED VIDEOS: A PERSPECTIVE ON CONCEPT MODELING

Chapter 5 addressed the choice of discriminative features which is one key issue concerning automatic multimedia content analysis. In this chapter, we concentrate on two other important aspects of video content analysis: Time efficiency and modeling of concepts (in this case, violence modeling). Traditional approaches to violent scene detection build on audio or visual features to model violence as a single concept in the feature space. Such modeling does not always provide a faithful representation of violence in terms of audio-visual features, as violence is not necessarily located compactly in the feature space. Consequently, in this chapter, we target to narrow this gap. To this end, we present a solution which uses the audio-visual features which were shown to be promising according to the extensive evaluations in Chapter 5 and propose to model violence by means of multiple so-called (sub)concepts using a feature space partitioning approach where we use SVMs and stacked auto-encoders as base classifiers. To cope with the heavy computations induced by the use of attribute and motion features, we perform a coarse-to-fine analysis, starting with a coarse-level analysis with time efficient audio-visual features and pursuing with a fine-level analysis with advanced features when necessary. As the presented framework provides a data-driven solution – no need to de-

scribe explicitly violence-related concepts – to violence analysis in videos, it can be extended to other complex video concepts. The presented framework is also extensible from the representation (by enriching coarse or fine-level features) and modeling aspects (by the use of other base classifiers). We perform experiments to evaluate both the effectiveness and efficiency of the framework. The experimental results on the standardized datasets of the latest editions of the *MediaEval Affect in Multimedia: Violent Scenes Detection (VSD) task* of 2014 and 2015 demonstrate that the presented framework is superior compared to the best systems of the MediaEval 2014 and 2015 participants. The work presented in this chapter has been published in [2, 7, 9].

6.1 Introduction

For the reasons we stated in the introductory section of Chapter 5, an effective violence detection solution, which is designed to automatically detect violent scenes in movies (or in videos in general), is highly desirable. Such an automated solution requires working with a proper representation of data which is an essential processing step. We performed an extensive evaluation of audio-visual features at different abstraction levels (low-level and mid-level) in Chapter 5. In this chapter, we use the audio and visual features which revealed to be promising according to these evaluations.

Based on an in-depth analysis of the literature, we identified two research questions (RQs) that we aim to address with the work presented in this chapter.

(RQ1) How to model the concept of violence given the data? This question has been qualified as an interesting research direction by the organizers of the VSD challenge [39]. We remarked that, in nearly all of the works mentioned in Section 2.4, the emphasis is on the feature engineering part of violence modeling and a single model is employed to model violence using audio or visual features at different abstraction levels. In other words, the samples constituting the development set are taken as a whole to train a unique "violence" classifier. Two types of improvements can be envisaged to better model violence. Both are based on the fact that violence can be expressed in very diverse manners. For instance, two distinct events (e.g., "explosion" and "fight") may be both intensely violent in the eyes of a consumer and, nevertheless, be located in different regions of the feature space (i.e., "violence" might not be expressed in a "compact" manner in the feature space). In addition, two events of the same type (e.g., "fight") might be characterized by distinct audio or visual features (fight between two individuals *vs.* brawl of 50 people). The first type of improvement is using manually designed multiple violence models, i.e., one model for each possible type of violence (e.g., one for "explosion", one for "fight"). The work by Ionescu et al. [61] is an example. However, such approaches do not solve the latter problem and are hardwired to the violence concept. The second type is deriving subconcepts from the data to build multi-

ple models. Partitioning the feature space to build the multiple models that correspond to the same concept might help in properly recognizing a given concept. In this type of approach, instead of incorporating domain knowledge (violence-related concepts in our case) into the system, the decision boundary is simplified by the use of multiple classifiers which are experts for different regions of the feature space. Therefore, instead of building a unique model to detect violence, we use feature space partitioning [152], where each "partition" (cluster) is used to build a separate violence subconcept model. This presents several advantages. It enables a faithful modeling of "violence". It also constitutes a data-driven operation, as it does not require defining manually several "violence" concepts (e.g., no need to have a separate concept for "explosion", "fire" or other similar concepts), as it directly builds on the data. Finally, this aspect is not hardwired to "violence" only, but can be extended to other complex concepts.

To the best of our knowledge, the sole work on violent scene detection performing feature space partitioning is the one by Goto and Aoki [55]. Feature space partitioning is achieved through mid-level violence clustering in order to implicitly learn mid-level concepts. However, their work is limited in two aspects. First, they cluster violent samples only. The inclusion of non-violent samples in the training process is done by a random selection. Such an approach presents the drawback of not taking into account the proximity of some violent and non-violent samples. For instance, if a violent sample and a non-violent one are closely located in the feature space, this is an indication that they are difficult to discriminate. Therefore, in order to obtain optimal classification boundaries, such particularities should be considered when building the models. Second, they use motion features, which are computationally expensive. This does not pose a problem for training. However, such a solution might introduce scalability issues, and might hinder the execution of violence detection in a real-world environment.

(RQ2) How to efficiently use computationally expensive features such as advanced motion and attribute features? In many of the existing works, next to audio or static visual cues, motion information is also used in the detection of violent scenes. Employed motion features range from simplistic features such as motion changes, shot length, camera motion or motion intensity [29, 53, 80, 99], to more elaborated descriptors such as STIP, ViF [36, 59] or dense trajectories, which have recently enjoyed great popularity. Dense trajectory features [127] have indeed received attention even among the VSD participants (e.g., [14, 34, 75]). Both types of motion approaches have drawbacks and advantages. Simplistic ones do not induce heavy computations but are likely to fail when it comes to efficacy; elaborated ones constitute powerful representations but result from computationally expensive processes. In addition to the advanced motion features, we also present the applicability of attribute features which are higher level static visual representations and which, to the best of our knowledge, have not been investigated for violence analysis in videos yet. To cope with the heavy computations induced by the use of

motion and attribute features, we perform a coarse-to-fine analysis, starting with coarse-level analysis with time efficient audio-visual features and pursuing with fine-level analysis with advanced features when necessary. Although MFCC-based mid-level audio features are effective and efficient representations, we decide to employ time efficient static visual representations in addition to the audio ones in the coarse-level analysis to keep the decrease in classification accuracy to a minimum due to this coarse-to-fine setup. In addition to time efficiency, the coarse-to-fine setup can be used for designing scalable solutions, i.e., adjustable depending on the processing power or accuracy requirements. To the extent of our knowledge, none of the studies addressing the detection of violent scenes in videos solve the issue of computational expense using such a staged approach.

To sum up, the contributions of this chapter can be summarized as follows: **(1)** A modeling of violence with feature space partitioning, which reliably models violence in the feature space without extensive supervision, and can be easily transposed for the detection of other concepts; and **(2)** a coarse-to-fine analysis approach which paves the road for time efficient and scalable implementations.

The chapter is organized as follows. Section 6.2 gives a detailed explanation of the proposed method including the representation of video segments, the model generation and the post-processing steps of the framework. We provide and discuss evaluation results obtained on the MediaEval VSD datasets of 2014 and 2015 consisting of Hollywood movies, Web videos and movies shared under CC licenses in Section 6.3. Concluding remarks and future directions to expand our current framework are presented in Section 6.4.

6.2 The Proposed Method

In Chapter 5, we performed violence detection at the video shot level. In this chapter, we address the problem of violence detection at the segment level. This means that we do not have at our disposal the video shot boundaries of the videos that we analyze. Therefore, we start our analysis by partitioning all videos into fixed-length segments.

As can be seen in Figure 6.1, our violence detection approach consists of training and testing phases. The training task involves the extraction of features from the raw video data, which are used to build audio and visual representations, the set of which constitutes a feature space. As indicated earlier, we do not wish to construct a single model obtained with these features, but a plurality of models (plurality of "subconcepts"), as we argue that "violence" might not be expressed in a "compact" manner in the feature space.

For most machine learning methods, testing would follow training, i.e., it would involve extraction of both audio and visual features followed by classification. However, because the extraction of motion and attribute features is computationally heavy, our testing does not exactly follow training. Instead, for the execution of our

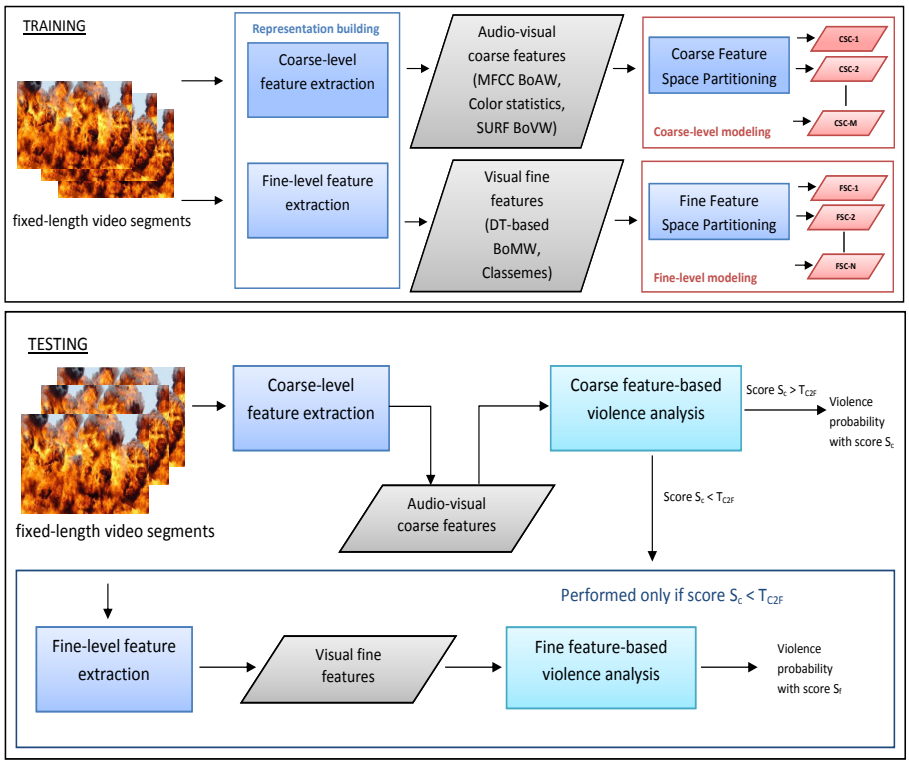


Figure 6.1: The general overview of our approach illustrating the two main phases of the system. The upper part of the figure gives the main steps performed during training (coarse and fine-level model generation), while the lower part shows the main steps of testing (execution). (BoAW: Bag-of-Audio-Words, BoMW: Bag-of-Motion-Words, BoVW: Bag-of-Visual-Words, CSC: coarse subconcept, DT: dense trajectory, FSC: fine subconcept, SURF: Speeded-Up Robust Features)

system, a *coarse-to-fine approach* is adopted, which explains the parallel construction of the testing scheme in Figure 6.1, where coarse-level violence detection is always performed while fine-level analysis is optional.

Therefore, in order to present in more detail each of these steps, the remainder of this section is organized as follows. Section 6.2.1 deals with video representation. Feature space partitioning is explained in Section 6.2.2. Section 6.2.3 details the generation of the subconcepts subsequent to partitioning. The audio or visual violence detections are based on combining the outputs of the models, which is presented in Section 6.2.4. Temporal smoothing and merging which are designed to further enhance performance are introduced in Section 6.2.5. Section 6.2.6 presents the coarse-to-fine analysis.

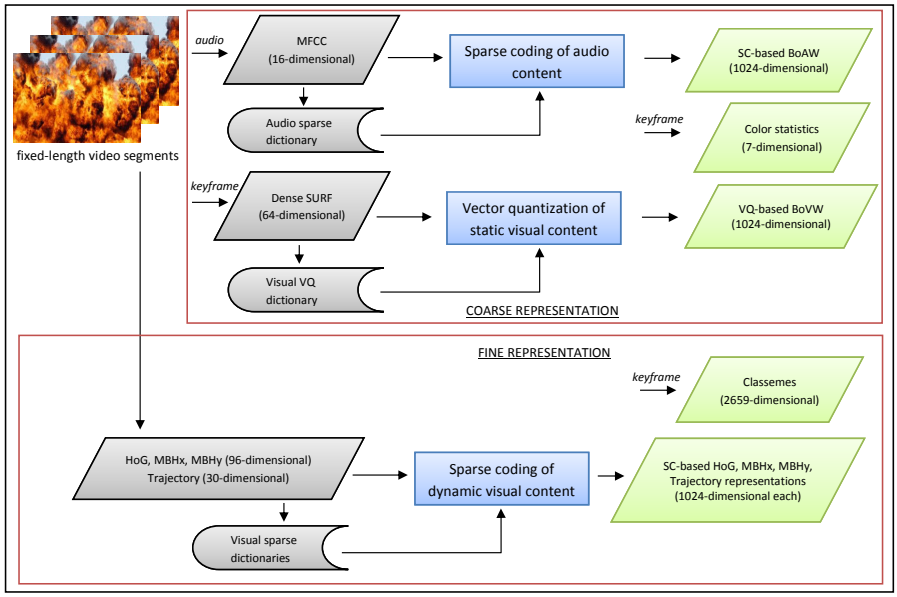


Figure 6.2: The generation process of audio and visual representations for video segments (upper part: coarse-level analysis features, lower part: fine-level analysis features). Separate dictionaries are constructed and used for MFCC, Dense SURF, Trajectory, HoG, MBHx and MBHy to generate 1024-dimensional representations. MFCC feature vectors are 16-dimensional and dense SURF ones are 64-dimensional. Each HoG, MBHx and MBHy descriptor is 96-dimensional, whereas the Trajectory descriptor is 30-dimensional. (BoAW: Bag-of-Audio-Words, BoVW: Bag-of-Visual-Words, SC: sparse coding, VQ: vector quantization)

6.2.1 Representation of video segments

In this section, we start introducing our framework by outlining the representation of audio-visual content of videos and present features that we use for *coarse-level* and *fine-level* video content analysis. We use audio, static and dynamic visual representations at different abstraction levels (low and mid-level) to represent fixed-length video segments from different perspectives in order to assess if they contain the concept of "violence". We determine the audio-visual representations used in this chapter according the performance evaluations of the feature descriptors in Chapter 5. An overview of the feature generation phase of the framework is presented in Figure 6.2.

For *coarse-level violence analysis*, the audio and static visual representations discussed in Chapter 5 are employed. More specifically, we use sparse coding based MFCC BoAW as the audio, and color statistics and SURF-based BoVW as the static

visual representations. We prefer sparse coding over vector quantization for the audio modality as, in this case, sparse coding was shown to provide more discriminative representations [3]. The construction of the audio and static visual representations are discussed in detail in Sections 5.3 and 5.4.1, respectively.

For the *fine-level analysis* phase of the framework, the dynamic visual and attribute representations presented in Chapter 5 are used. Dense trajectory based BoMW feature descriptors are employed as the dynamic visual representation (discussed in detail in Section 5.5), whereas Classemes feature descriptors are extracted as the attribute representation (discussed in detail in Section 5.4.2).

The static visual representations of the video segments both in the coarse-level and fine-level stages are extracted from the keyframe of the videos which is the frame in the middle of a fixed-length video segment, whereas the whole segment is processed for the audio and dynamic visual representations.

6.2.2 Feature space partitioning

As discussed before, "violence" is a concept which can be expressed in diverse manners. For instance, in a dataset, both explosions and fight scenes are labeled as violent according to the definition that we adopted. However, these scenes might highly differ from each other in terms of audio-visual appearance depending on their characteristics of violence. Instead of learning a unique model for violence detection, learning multiple models constitutes a more judicious choice. Therefore, in the learning phase, a "divide-and-conquer" (*"divide-et-impera"*) approach is applied by performing feature space partitioning. The first step of learning multiple violence subconcept models is to partition the feature space into smaller portions. We perform partitioning by clustering the set of fixed-length video segments in our development dataset. Moreover, we employ the Approximate Nearest Neighbor (ANN) k-means algorithm [94] which is a variant of Lloyd's algorithm [82] particularly suited for very large clustering problems [125]. The ANN algorithm computes approximated nearest neighbors to speed up the instance-to-cluster-center comparisons. We use the Euclidean metric for distance computations, initialize cluster centers with the k-means++ algorithm [12] and repeat k-means clustering several times before determining data clusters. By applying feature space partitioning, we infer (sub)concepts in a data-driven manner as opposed to approaches (e.g., [61]) which use violence-related concepts as a mid-level step.

6.2.3 Model generation for subconcepts

For the generation of subconcept models, we apply the following procedure. After clusters are generated (Section 6.2.2), they are distinguished as *atomic* and *non-atomic* clusters as in [103]. A cluster is defined as atomic, if it contains patterns of the same type, i.e., patterns which are all either "violent" or "non-violent". No model generation is realized for atomic clusters and their class labels are stored for

further use in the test phase. Non-atomic clusters are clusters which include patterns from both the violent and non-violent classes. For those non-atomic clusters, a different model is built for each violence subconcept (cluster). As base classifier in order to learn violence models, we use two different classifiers, namely the two-class SVM and the stacked auto-encoder (SAE). An overview of the generation of the violence models is presented in Figure 6.3. For the SAE-based approach, we train a deep neural network with two hidden layers. For each hidden layer, we employ an auto-encoder and train each auto-encoder in an unsupervised manner. As a final step, fine tuning is performed on the whole neural network (i.e., training the deep neural network with the development data in a supervised manner).

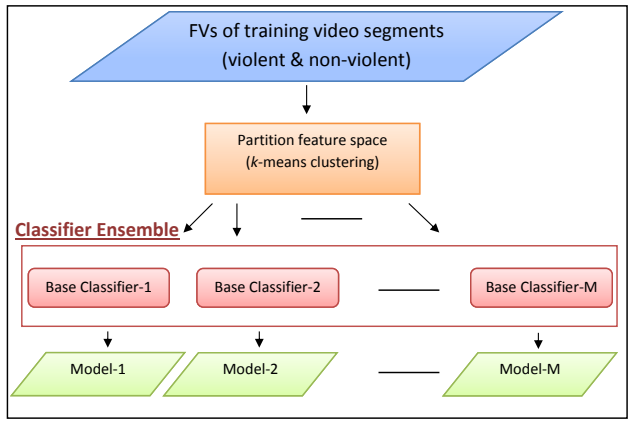


Figure 6.3: The generation of violence detection models with feature space partitioning through k-means clustering. Feature vectors (either coarse or fine features) of training segments are given as input to the combination process. Two different base classifiers are used in this work: SVM and stacked auto-encoder. (FV: feature vector)

In the learning step of the framework, the main issue is the problem of imbalanced data. This is caused by the fact that the number of non-violent video segments is much higher than the number of violent ones in the development dataset. When the two-class SVM is used as the base classifier, this phenomenon causes the learned boundary being too close to the violent instances. Consequently, the SVM has a natural tendency towards classifying every sample as non-violent. Our undersampling strategy to cope with this bias is discussed in detail in the previous chapter (Section 5.6). When the base classifier is the SAE, we use the same development data as the two-class SVM. As a result of the undersampling strategy used in this work, which causes a reduction of the development data, the efficiency of the training process is improved; hence, training is easily scalable to large datasets similar to the ones used in the context of our work.

6.2.4 Combining predictions of models

In the test phase, the main challenge is to combine the classification results (probability outputs) of the violence models. In order to achieve the combination of classification scores of base classifiers, we perform one of two different combination methods, namely *classifier selection* and *classifier fusion*, which are alternative solutions. Normally, in a basic SVM, only class labels or scores are output. The class label results from thresholding the score, which is not a probability measure. The scores output by the SVM are converted into probability estimates using the method explained in [131].

An overview of the combination methods is presented in Figure 6.4. Both methods follow the main canvas of Figure 6.4, i.e., they get feature vectors as input and return a violence score; the frames highlight the specificities of each of them.

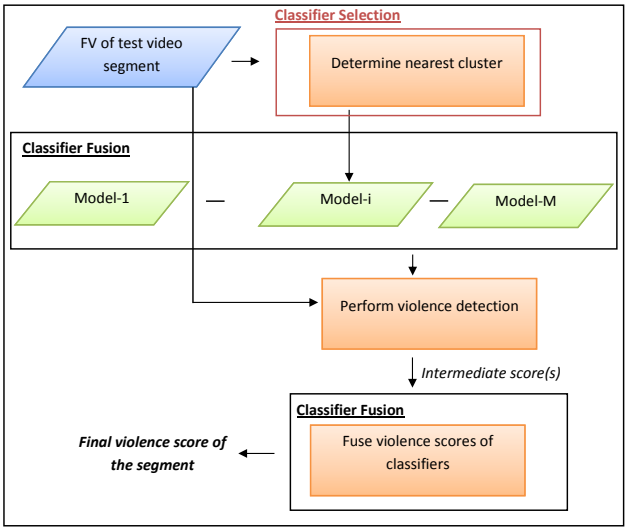


Figure 6.4: An overview of the classifier decision combination phase of our method. Feature vector (either coarse or fine features) of the test segment is given as input to the combination process. (FV: feature vector)

6.2.4.1 Classifier selection

In the *classifier selection* method, we first determine the nearest cluster to a test video segment using the Euclidean distance measure, i.e., the cluster minimizing the following Euclidean distance is identified:

$$d(c_i, x_j) = \|c_i - x_j\| \tag{6.1}$$

where c_i represents a given cluster center, and x_j a video segment. If the nearest cluster is an atomic cluster, then the test video segment is labeled with the unique label of the cluster and the probability measure is set to 1.0. For the non-atomic cluster case, once the best fitting classifier for the video sample is determined, the probability output of the corresponding model is used as the final prediction for that video sample.

6.2.4.2 Classifier fusion

In the *classifier fusion* method, we combine the probability outputs of all cluster models (both atomic and non-atomic clusters). As in the classifier selection method, the probability measure is set to 1.0 for atomic clusters. The classifiers that we adopt are all either SVMs or SAEs. Hence, we are in the presence of homogeneous "learners" (all of the same type) according to the terminology of [152]. In such a situation, it is advised to directly fuse the violence probabilities ($h_i(x_j)$) generated by each of the classifiers ("learners") using the *weighted soft voting* technique [152]:

$$H(x_j) = \sum_{i=1}^{T} w_{ij} h_i(x_j) \tag{6.2}$$

As shown in Equation 6.2, a classifier-specific weight (w_{ij}) is dynamically assigned to each classifier for each test video segment (x_j) using the Euclidean distance of the segment to the cluster centers. The weights assigned to the clusters are determined such that they always sum up to 1.

6.2.5 Temporal smoothing and merging

As mentioned earlier, we split videos into small fixed-length segments. However, a violent scene may be composed of several continuous segments. While some of these segments can easily be identified as violent, others might be more challenging to detect. Previous findings support that if a segment contains violence, its neighboring segments are likely to contain violence as well and that, consequently, temporal score smoothing is likely to boost the performance. Therefore, we perform the post-processing steps of (1) temporal smoothing and (2) segment merging, in order to further improve the performance of the system as suggested in [34].

The temporal smoothing technique that we adopt consists in applying a simple yet efficient score smoothing, where the smoothed violence prediction score of a segment is determined as the average of the scores in a three-segment window (i.e., window of three consecutive segments).

Our segment merging method is based on the use of a threshold value ($T_{violence}$) which is determined experimentally on the development dataset. We merge two neighboring segments, if they are both identified as violent or non-violent (i.e., their violence scores are above or below $T_{violence}$) and set the new violence prediction score as the average of the related segments.

6.2.6 Coarse-to-fine violence analysis

The inclusion of a coarse-to-fine analysis in the whole process originated from the observation that the extraction of dense trajectory features or of Classemes is a computationally expensive process. The former involves computing and tracking features over several frames, while the latter requires getting responses from a large number of weakly trained object category classifiers.

Various precautions could be taken to cope with this issue. Concerning motion features, some straightforward solutions include for instance: Resizing frames; tuning parameters, e.g., using an increased step size; considering only a subset of the dense trajectory features. Concerning Classemes, one possibility could be to use a more compact binary version of the Classemes descriptor. However, the gain in computation time obtained through such measures comes at the expense of decreased accuracy. We therefore developed an alternative solution in the form of coarse-to-fine classification.

We observed that, for the task of violence detection, audio is an extremely discriminative feature. A violence detection approach for video analysis based solely on audio features (e.g., MFCC) will normally fail only if the video contains no sound or if the volume is low. In order to address this issue, we decide to include time-efficient static visual features such as color and texture in the coarse classification phase, as these features provide complementary information (e.g., for violent scenes with a low sound volume, but with a dark nature). When, according to audio and static visual features, a segment is classified as violent, we can realistically assume that this "violent" label is correct. However, if it is classified as non-violent, then a verification by the use of "advanced" visual features (dense trajectory based Bag-of-Motion-Words and Classemes) would be necessary to confirm the absence of violence.

From a practical point of view, the implementation of violence detection (i.e., execution of the system during test) follows the scheme under the lower part of Figure 6.1 (testing). First of all, coarse detection based on audio and static visual features is performed. MFCC features are extracted and converted to mid-level representations, and static visual features are computed as described under Section 6.2.1. Coarse-level analysis is further performed in line with the teachings of Sections 6.2.4 and 6.2.5. During coarse analysis, a segment is assigned a score (S_c) which is compared to a threshold T_{C2F} (coarse-to-fine analysis threshold). If S_c exceeds this threshold T_{C2F}, the segment is labeled as violent with score S_c. If not, fine analysis based on advanced visual features is initiated. The fine-level visual features are extracted and converted to mid-level features, and visual feature analysis is run. The outcome is a score S_f, which is compared to the threshold $T_{violence}$ (threshold mentioned in Section 6.2.5) to determine if the segment is violent.

6.3 Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of our method based on feature space partitioning (referred hereafter as "FSP") against the method employing a unique model and also at highlighting the advantages brought by the coarse-to-fine analysis. We also evaluate the performance of audio and visual features at the different levels (coarse and fine) described in Section 6.2.1. Because of potential differences in the definition of "violence" adopted in published works discussed in Section 2.4, a direct comparison of our results with those works is not always meaningful. Nevertheless, we can compare our approach with the methods which took part to the MediaEval 2014 and 2015 VSD tasks (segment-level detection for 2014, and clip-level detection for 2015), where the same "violence" definition and datasets are employed. In addition, our FSP approach is compared with the fundamental ensemble learning methods, namely Bagging and Mixture-of-Experts (MoE).

Using the evaluation framework provided by the MediaEval VSD task is an opportunity to test our algorithms in a standardized manner on a standardized data corpus. Although running since 2011, the MediaEval VSD task reached a certain level of maturity only in 2014, when the organizers opted for the subjective definition of violence, and for the use of the mean average precision metric. The same definition and evaluation metric were kept for the 2015 edition. For these reasons, we show our results on the latest two editions of the the MediaEval VSD task (2014 and 2015). The 2015 dataset differs from the 2014 data in difficulty, as can be seen from the results (see below).

The MediaEval 2014 VSD challenge is structured as two separate tasks: A *main task* which consists in training on Hollywood movies only and testing on Hollywood movies only; and a *generalization task* which consists in training on Hollywood movies only and testing on Web videos only. The MediaEval 2015 VSD structure is slightly different than the task of 2014 in terms of dataset and consists of only one task. In addition, violence annotation is performed at the clip-level instead of at the frame-level as in 2014.

6.3.1 Dataset and ground-truth

For the evaluation within the context of the **MediaEval 2014 VSD task**, we used two different types of dataset in our experiments: (1) A set of 31 movies which were the movies of the MediaEval 2014 VSD task (referred hereafter as the "Hollywood movie dataset"), and (2) a set of 86 short YouTube Web videos under Creative Commons (CC) licenses which were the short Web videos of the MediaEval 2014 VSD task (referred hereafter as the "Web video dataset").

A total of 24 movies from the Hollywood set are dedicated to the development process: *Armageddon, Billy Elliot, Eragon, Harry Potter V, I am Legend, Leon, Midnight Express, Pirates of the Caribbean I, Reservoir Dogs, Saving Private Ryan, The Sixth Sense, The Wicker Man, The Bourne Identity, The Wizard of Oz, Dead Poets So-*

ciety, Fight Club, Independence Day, Fantastic Four I, Fargo, Forrest Gump, Legally Blond, Pulp Fiction, The God Father I and *The Pianist*. The remaining 7 movies – *8 Mile, Brave Heart, Desperado, Ghost in the Shell, Jumanji, Terminator II* and *V for Vendetta* – and the Web video dataset serve as the test set for the main and generalization task, respectively.

Each movie and short Web video is split in a multitude of fixed-length video segments as exposed in Section 6.2. The development set (24 movies) consists of 57,933 video segments, whereas the movie test set (7 movies) consists of 16,668 video segments and the Web video dataset consists of 3,149 such short video segments, where each segment is labeled as *violent* or *non-violent*. The characteristics of the MediaEval 2014 VSD development dataset are provided in Table 6.1; Table 6.2 provides the details on the test set of the same year. The movies of the development and test sets were selected in such a manner that both development and test data contain movies of variable violence levels (ranging from extremely violent movies to movies without violence) from different genres and production years ranging from 1939 (*The Wizard of Oz*) to 2007 (*I am Legend*). On average, around 14.45% of segments are annotated as violent in the whole dataset (the Hollywood movie and Web video datasets combined).

The ground-truth of the Hollywood dataset was generated by 9 human assessors, partly by developers, partly by potential users. Violent movie and Web video segments are annotated at the frame level. For the generalization task, the ground-truth was created by several human assessors[1] who followed the subjective definition of violence as explained in [37]. A detailed description of the Hollywood and the Web video datasets, and the ground-truth generation are given in [113].

In addition to the datasets provided by the MediaEval 2014 VSD task, we also performed our evaluations on the dataset provided by the **MediaEval 2015 VSD task** (referred hereafter as the "VSD 2015 movie dataset"). The development and test sets are completely different from the dataset of the MediaEval 2014 VSD task. The VSD 2015 movie dataset consists of 10,900 video clips (each clip lasting from 8 to 12 seconds approximately) extracted from 199 movies, both professionally edited and amateur movies. The movies in the dataset are shared under CC licenses that allow redistribution. The VSD 2015 development dataset contains 6,144 video clips, whereas the test set has 4,756 clips. As in the MediaEval 2014 VSD evaluation, each movie clip is split in a multitude of video segments, where each video segment has a duration of 3 seconds. On average, around 4.61% of segments are annotated as violent in the whole dataset (development and test sets combined). The details of the VSD 2015 development and test set are presented in Table 6.3. The ground-truth generation process of the VSD 2015 movie dataset is similar to the VSD task of 2014. The details of the ground-truth generation are explained in [112].

[1] Annotations were made available by *Fudan University, Vietnam University of Science*, and *Technicolor*.

Table 6.1: The characteristics of the Hollywood movie development dataset (The number of violent and non-violent video segments and the percentage of violent segments).

#	Video Title	Violent	Non-violent	Overall
1	*Armageddon*	251 (8.68%)	2,642	2,893
2	*Billy Elliot*	71 (3.36%)	2,045	2,116
3	*Dead Poets Society*	19 (0.77%)	2,452	2,471
4	*Eragon*	284 (14.24%)	1,711	1,995
5	*Fantastic Four I*	435 (21.42%)	1,596	2,031
6	*Fargo*	296 (15.73%)	1,586	1,882
7	*Fight Club*	462 (17.32%)	2,206	2,668
8	*Forrest Gump*	235 (8.62%)	2,490	2,725
9	*Harry Potter V*	169 (6.38%)	2,482	2,651
10	*I am Legend*	311 (16.15%)	1,615	1,926
11	*Independence Day*	403 (13.68%)	2,542	2,945
12	*Legally Blond*	0 (0.00%)	1,841	1,841
13	*Leon*	371 (17.55%)	1,743	2,114
14	*Midnight Express*	183 (7.89%)	2,137	2,320
15	*Pirates of the Caribbean I*	526 (19.15%)	2,221	2,747
16	*Pulp Fiction*	753 (25.43%)	2,208	2,961
17	*Reservoir Dogs*	594 (31.20%)	1,310	1,904
18	*Saving Private Ryan*	1,113 (34.25%)	2,137	3,250
19	*The Bourne Identity*	179 (7.88%)	2,093	2,272
20	*The God Father I*	209 (6.15%)	3,189	3,398
21	*The Pianist*	460 (16.11%)	2,395	2,855
22	*The Sixth Sense*	51 (2.48%)	2,008	2,059
23	*The Wicker Man*	141 (7.21%)	1,815	1,956
24	*The Wizard of Oz*	27 (1.38%)	1,926	1,953
–	**Total**	**7,543 (13.02%)**	50,390	**57,933**

6.3.2 Experimental setup

As mentioned in the method section (Section 6.2), videos were first segmented into fixed-length pieces before any further processing. Each video segment lasts 3 seconds and corresponds to 75 visual frames. This duration of 3 seconds was determined according to the pre-evaluation runs. We experimentally verified that the 3-second time window was short enough to be computationally efficient and long enough to retain sufficient relevant information. Once the videos are segmented into fixed-length pieces, the construction of feature representation is performed for each piece separately.

The setup for audio and visual feature extraction process of the framework was already explained in the previous chapter (Section 5.7.2). The only difference in the

Table 6.2: The characteristics of the Hollywood movie test dataset and the Web video dataset (The number of violent and non-violent video segments and the percentage of violent segments).

#	Video Title	Violent	Non-violent	Overall
1	*8 Mile*	108 (5.10%)	2,010	2,118
2	*Brave Heart*	771 (22.62%)	2,637	3,408
3	*Desperado*	658 (32.82%)	1,346	2,004
4	*Ghost In The Shell*	174 (10.51%)	1,481	1,655
5	*Jumanji*	150 (7.51%)	1,848	1,998
6	*Terminator II*	769 (26.12%)	2,174	2,943
7	*V For Vendetta*	402 (15.81%)	2,140	2,542
8	*The Web video dataset*	1,122 (35.63%)	2,027	3,149
–	**Total**	**4,154 (20.96%)**	15,663	**19,817**

Table 6.3: The characteristics of the VSD 2015 development and test sets (The number of violent and non-violent video segments and the percentage of violent segments).

Dataset Type	Violent	Non-violent	Overall
Development	738 (4.41%)	15,999	16,737
Test	634 (4.88%)	12,367	13,001
Total	**1,372 (4.61%)**	28,366	**29,738**

setup is that we use the frame in the middle of a video segment as the keyframe of that segment, since the keyframes of videos are not provided for the datasets.

The VLFeat[2] open source library was used to perform k-means clustering (k ranging from 5 to 40 in this work). In the clustering phase, the number of iterations represents a trade-off between clustering quality and time efficiency. We observed that after 8 iterations stable clusters were obtained; increasing this number did not improve significantly the quality of the obtained clusters.

Training was performed using audio and visual features extracted at the video segment level. As base classifiers, we used two-class SVMs and SAEs. The two-class SVMs were trained with an RBF kernel using libsvm[3] as the SVM implementation. The hyper-parameter optimization of the SVMs were explained in the previous chapter (Section 5.7.2). When the SAE was the base classifier, we used the scaled conjugate gradient backpropagation algorithm [92] due to its computational efficiency and sigmoid as the activation function. Hidden layer sizes of the network were one fourth and one eighth of the number of dimensions of coarse-level or fine-level representations. Unsupervised training of the deep neural network was performed using the development datasets of the MediaEval 2014 and 2015 VSD tasks.

[2] `http://www.vlfeat.org/`

[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Intermediate violence scores of video segments – provided by the base classifiers – were combined to determine the violence score of the segments using either the classifier selection or fusion mechanism as explained in detail in Section 6.2.4. Subsequently, post-processing was applied on the violence score of the video segments as explained in detail in Section 6.2.5.

In order to generate the MoE models that are used for comparison with the FSP approach, we employed the Mixture-of-Experts Matlab toolbox[4] and applied a cooperative approach for model generation. For the MoE model generation, the number of experts was determined according to the number of clusters that leads to the best performance for the FSP approach. The number of experts for the MoE modeling was set to 10 for the MediaEval 2014 VSD datasets, and to 5 for the MediaEval 2015 VSD dataset. For the bagging modeling, the number of trees was 100 (experimentally determined figure).

The evaluation metric is *average precision (AP)* which is the official metric of the MediaEval 2014 and 2015 VSD tasks. The mean of AP (MAP) values on the MediaEval 2014 and 2015 datasets are provided in the results. The details regarding the computation of MAP on the MediaEval 2014 dataset are provided in [107], whereas the MAP values on the MediaEval 2015 VSD dataset are computed using the tool provided by NIST[5].

6.3.3 Results and discussion

The evaluation of our approach is achieved from different perspectives: ***(i)*** Comparison to unique concept modeling; ***(ii)*** comparison to fundamental ensemble learning methods (i.e., Bagging [21] and Mixture-of-Experts [65]); ***(iii)*** comparison to MediaEval 2014 participants; ***(iv)*** comparison to MediaEval 2015 participants; and ***(v)*** added-value of coarse-to-fine analysis.

Experiment *(i)* reviews the gain in classification performance brought by FSP. In this case, the baseline method for comparison is the approach using a single SVM or SAE classifier trained with the same data resulting in a single model for violence. The results are summarized in Tables 6.4, 6.5 and 6.6 which provide a comparison of the FSP method against the unique violence detection model (no FSP) in terms of MAP metric on the Hollywood movie dataset (Table 6.4), on the Web video dataset (Table 6.5) and on the VSD 2015 movie dataset (Table 6.6), respectively. For the FSP method, evaluated k values for the MediaEval 2014 datasets are 10, 20 and 40, whereas these values range from 5 to 20 for the MediaEval 2015 dataset. The conclusions which can be drawn from **Table 6.4** are as follows:

- The highest MAP values of 0.55 and 0.60 (for SVM-based and SAE-based FSP, respectively) are achieved with a fine-level FSP analysis for the situation where

[4] `https://goker.wordpress.com/2011/07/01/mixture-of-experts/`
[5] `http://trec.nist.gov/trec_eval/`

Table 6.4: The MAP of the FSP method with coarse-level and fine-level representations, k clusters (k = 10, 20 and 40) and different classifier combination methods (selection and fusion) and an SVM-based/SAE-based unique violence detection model (no feature space partitioning) on the **Hollywood movie dataset**. The highest MAP values for the SVM-based and SAE-based FSP method are highlighted. (MAP: mean average precision, SAE: stacked auto-encoder).

Method	MAP (SVM)	MAP (SAE)
FSP method (fine, fusion, k = 10)	0.42	**0.60**
FSP method (fine, fusion, k = 20)	**0.55**	0.54
FSP method (fine, fusion, k = 40)	0.41	0.44
FSP method (coarse, fusion, k = 10)	0.38	0.48
FSP method (coarse, fusion, k = 20)	0.45	0.42
FSP method (coarse, fusion, k = 40)	0.35	0.37
FSP method (fine, selection, k = 10)	0.45	0.47
FSP method (fine, selection, k = 20)	0.34	0.38
FSP method (fine, selection, k = 40)	0.32	0.35
FSP method (coarse, selection, k = 10)	0.37	0.42
FSP method (coarse, selection, k = 20)	0.32	0.35
FSP method (coarse, selection, k = 40)	0.30	0.33
Unique model (fine)	0.35	0.42
Unique model (coarse)	0.32	0.37

the feature space is split in k clusters (k = 20 for the SVM-based FSP and k = 10 for the SAE-based FSP) and classifier scores are fused.

- Another observation is that the MAP values of the FSP method are usually higher regardless of the type of base classifier (SVM or SAE), when the adopted score combination method is classifier fusion.

- One final observation is about the choice of the number of clusters (i.e., the optimal number of clusters) for the FSP method. For instance, decreasing the number of clusters from 20 to 10 reduces performance for the SVM-based FSP approach. Similarly, increasing the number of clusters to 40 (SVM-based FSP) and 20 (SAE-based FSP) does not help in obtaining a better accuracy either. This shows that a non-optimal number of clusters fails to provide a faithful representation and modeling of the development dataset.

The important observations from **Table 6.5** are summarized as follows:

- The highest MAP value of 0.75 is achieved with a fine-level SAE-based FSP analysis for the situation where the feature space is split in k = 10 clusters and classifier selection is used as the score combination method.

Table 6.5: The MAP of the FSP method with coarse-level and fine-level representations, k clusters (k = 10, 20 and 40) and different classifier combination methods (selection and fusion) and an SVM-based/SAE-based unique violence detection model (no feature space partitioning) on the **Web video dataset**. The highest MAP values for the SVM-based and SAE-based FSP method are highlighted. (MAP: mean average precision, SAE: stacked auto-encoder).

Method	**MAP (SVM)**	**MAP (SAE)**
FSP method (fine, fusion, k = 10)	0.64	0.65
FSP method (fine, fusion, k = 20)	0.65	0.67
FSP method (fine, fusion, k = 40)	0.63	0.64
FSP method (coarse, fusion, k = 10)	0.67	0.69
FSP method (coarse, fusion, k = 20)	0.63	0.65
FSP method (coarse, fusion, k = 40)	0.61	0.62
FSP method (fine, selection, k = 10)	**0.69**	**0.75**
FSP method (fine, selection, k = 20)	**0.69**	0.71
FSP method (fine, selection, k = 40)	0.66	0.67
FSP method (coarse, selection, k = 10)	**0.69**	0.72
FSP method (coarse, selection, k = 20)	0.64	0.65
FSP method (coarse, selection, k = 40)	0.58	0.60
Unique model (fine)	0.59	0.61
Unique model (coarse)	0.44	0.45

- In addition, we observe that generally the results show little variability. For Web videos, FSP methods outperform unique modeling solutions regardless of the type of base classifier, when comparing approaches using the same coarse-level or fine-level representations.

- More importantly, for Web videos, selection based combinations perform better than fusion based combinations, in general. This differs from the results obtained on Hollywood movies (Table 6.4), where fusion based combinations provided better outcomes. A plausible explanation is that, in movies, violent scenes are long, and often correspond to several subconcepts, while user-generated Web videos are short, and, therefore, are likely to correspond to a single subconcept.

The conclusions which can be drawn from **Table 6.6** are as follows:

- The highest MAP value of 0.3413 is achieved with a fine-level SAE-based FSP analysis method for the situation where the feature space is split in k = 5 clusters and classifier selection is used as the score combination method.

- Globally, for the videos in the VSD 2015 movie dataset, selection-based combinations perform better than fusion-based combinations as in the Web videos

Table 6.6: The MAP of the FSP method with coarse-level and fine-level representations, k clusters (k = 5, 10 and 20) and different classifier combination methods (selection and fusion) and an SVM-based/SAE-based unique violence detection model (no feature space partitioning) on the **VSD 2015 movie dataset**. The highest MAP values for the SVM-based and SAE-based FSP method are highlighted. (MAP: mean average precision, SAE: stacked auto-encoder).

Method	**MAP (SVM)**	**MAP (SAE)**
FSP method (fine, fusion, k = 5)	0.1419	0.2104
FSP method (fine, fusion, k = 10)	0.2032	0.1798
FSP method (fine, fusion, k = 20)	0.1118	0.1189
FSP method (coarse, fusion, k = 5)	0.1341	0.1512
FSP method (coarse, fusion, k = 10)	0.1238	0.1375
FSP method (coarse, fusion, k = 20)	0.1092	0.1168
FSP method (fine, selection, k = 5)	0.2682	**0.3413**
FSP method (fine, selection, k = 10)	**0.3173**	0.3256
FSP method (fine, selection, k = 20)	0.1438	0.2103
FSP method (coarse, selection, k = 5)	0.1851	0.2548
FSP method (coarse, selection, k = 10)	0.2215	0.1912
FSP method (coarse, selection, k = 20)	0.1128	0.1213
Unique model (fine)	0.2265	0.2427
Unique model (coarse)	0.2048	0.2167

(Table 6.5). This differs from the results obtained on Hollywood movies (Table 6.4). A possible explanation is that, in the VSD 2015 movies, violent scenes correspond to single events (e.g., a fight) rather than complicated events, and, therefore, are likely to correspond to a single subconcept.

- When compared to the MAP values reported in Table 6.4 and 6.5, the MAP values are globally lower. This characteristic is independent from our approach. While the best runs of the VSD 2014 participants could reach a MAP of 0.6, those of 2015 could not exceed 0.3. The 2015 dataset is, hence, more difficult. We think that the extra difficulty of the 2015 dataset is caused by two factors. First, the VSD 2015 movie dataset is not originally selected for violence concept analysis; the violence level of the dataset is, therefore, quite low (around 4% of the clips). Second, the violence concept is less "emphasized" by the film editing rules which are usually used in the 2014 Hollywood movies.

As presented in this section, the optimal number of clusters that leads to maximum classification performance in terms of MAP varies among the MediaEval 2014 and 2015 datasets. We have a closer look at the video segments in clusters which are used to generate FSP violence models for the VSD 2015 movie dataset in order

to evaluate whether the clusters correspond to a meaningful subconcept from a "human-perspective" such as fights or explosions, or whether they just simplify the decision boundary of classifiers. It is observed that the segments within the clusters have similar audio and visual scene characteristics. For instance, in one of the clusters, the video segments mainly have grayish color tones and melancholic music in the background. However, there seems to be no apparent "human-perspective" subconcept. This finding suggests that the FSP approach leads to simplified local decision boundaries which do not match "human-perspective" violence-related subconcepts. In addition, the dependence between the number of atomic clusters and the optimal number of clusters for the MediaEval 2014 and 2015 datasets is evaluated. We see that on both datasets, the classification performance in terms of MAP starts decreasing with the introduction of atomic clusters in the FSP approach.

Experiment *(ii)* aims at comparing our FSP approach with the fundamental ensemble learning methods (Bagging and MoE). Table 6.7 provides a comparison of our best performing SVM-based and SAE-based FSP methods using coarse-level and fine-level representations with these ensemble learning methods (in terms of MAP) on the Hollywood movie, Web video and VSD 2015 movie datasets.

Table 6.7: The MAP for MoE and Bagging ensemble learning methods, and our best performing FSP method on the Hollywood movie, Web video and VSD 2015 movie datasets. The details relating to bagging and MoE model generation are explained in Section 6.3.2. The highest MAP values per MediaEval dataset are highlighted. (cf: classifier fusion, cs: classifier selection, MAP: mean average precision, MoE: Mixture-of-Experts, SAE: stacked auto-encoder).

Method	**MAP - Movies'14**	**MAP - Web**	**MAP - Movies'15**
FSP (SVM, coarse)	0.45 (cf, k=20)	0.69 (cs, k=10)	0.2215 (cs, k=10)
FSP (SVM, fine)	0.55 (cf, k=20)	0.69 (cs, k=10)	0.3173 (cs, k=10)
FSP (SAE, coarse)	0.48 (cf, k=10)	0.72 (cs, k=10)	0.2548 (cs, k=5)
FSP (SAE, fine)	**0.60** (cf, k=10)	**0.75** (cs, k=10)	**0.3413** (cs, k=5)
Bagging (coarse)	0.34	0.30	0.13
MoE (fine)	0.38	0.42	0.21

The important observations from **Table 6.7** can be summarized as follows:

- The fine-level SAE-based FSP analysis method outperforms the bagging and MoE ensemble learning methods in terms of MAP on the MediaEval 2014 and 2015 VSD datasets. In addition, the FSP method regardless of the type of base classifier (SVM or SAE) performs better than the bagging and MoE fundamental ensemble learning methods.

- Another interesting observation is that the MoE modeling performs better when fine-level representations are used, whereas the bagging modeling per-

forms better with coarse-level representations. This might be due to the fact that the bagging modeling is applied on the decision tree method.

Experiment *(iii)* aims at comparing our approach with the MediaEval submissions of 2014 [89]. Table 6.8 provides a comparison of our best performing SVM-based and SAE-based FSP methods with the best run of participating teams (in terms of MAP for the main and generalization tasks) in the MediaEval 2014 VSD task. If we look at the results, we notice that there is a pattern: All the solutions perform better for the generalization task, except Fudan-NJUST. For a fair evaluation, our results are compared only against the teams who submitted results for both tasks; in this case, the best systems were the ones from Fudan-NJUST and FAR. Fudan-NJUST and FAR ranked first for the main task (with a MAP of 0.63) and for the generalization task (with a MAP of 0.664), respectively.

Table 6.8: The MAP for the best run of teams in the MediaEval 2014 VSD Task [89] and our best performing SVM-based and SAE-based FSP methods on the Hollywood movie and Web video datasets. Teams with at least one participant member of the MediaEval VSD organizing team are marked by *. The runs provided by the organizing team are marked by **. (NA: Not Available).

Team	**MAP - Movies**	**MAP - Web**
*Fudan-NJUST** (run4*) [34]*	**0.63**	0.5
FSP (SAE, fine, fusion, k = 10)	**0.60**	**0.65**
NII-UIT [76]*	0.559	NA
FSP (SVM, fine, fusion, k = 20)	0.55	0.65
FSP (SAE, fine, selection, k = 10)	**0.47**	**0.75**
*Fudan-NJUST** (run2*) [34]*	0.454	0.604
FAR (run1*) [114]*	0.451	0.578
FSP (SVM, fine, selection, k = 10)	0.45	0.69
MIC-TJU [149]	0.446	0.566
RECOD [14]	0.376	0.618
FAR (run3*) [114]*	0.25	**0.664**
ViVoLab-CVLab [25]	0.178	0.43
TUB-IRML [2]	0.172	0.517
*Random run (baseline) **	0.061	0.364
MTM [44]	0.026	NA

The following observations can be inferred from **Table 6.8**. For the main task, our SAE-based FSP solution (fine-level representations with classifier fusion and $k = 10$) achieves results very close to *run4* of Fudan-NJUST, i.e., our results are competing with state-of-the-art results (0.60 versus 0.63). However, a closer look at the short paper describing the Fudan-NJUST system [89] reveals that the Fudan-NJUST team used more features than our system: Not only the 3 visual and 1 audio features that we used here, but also 6 additional ones (STIP; Fisher-encodings of HoG, HoF,

MBHx, MBHy and trajectory shape). Hence, we argue that the performance difference can be explained by the inclusion of more features, which also results in larger complexity; there also seems to be strong evidence in the literature supporting our belief that the performance difference is caused by the different feature set. In a paper by members of the Fudan-NJUST team [123], the accuracy obtained with larger feature sets is reported to be better. In addition, the MAP of 0.63 of *run4* was achieved by fusing SVM with deep learning methods (the one-classifier-type *runs 1* and *2* were below 0.454). In contrast, we used only one type of classifiers (SVM). Finally, our FSP method also presents the main advantage of being time efficient (see below for coarse-to-fine analysis).

For the generalization task, our SVM-based and SAE-based FSP solutions (fine-level representations with classifier selection and $k = 10$) outperform the *run3* of the FAR team, which ranked first. We achieved 0.69 and 0.75 with the SVM-based and SAE-based solutions respectively, while their *run3* achieved a MAP of 0.664.

The comparison to other MediaEval participants outlined above clearly demonstrates excellent results already. However, when considering the **aggregated** results of a given run (i.e., the MAP on the main task and the MAP on the generalization task **for a given run)**, our SAE-based FSP solution that uses fine-level analysis, classifier fusion and $k = 10$ **(0.47 - 0.75)** outperforms *runs 2* **(0.454 - 0.604)** and *4* **(0.63 - 0.5)** of Fudan-NJUST and *runs 1* **(0.45 - 0.578)** and *3* **(0.25 - 0.664)** of FAR. The same observation can be inferred for the best performing SVM-based FSP method (fine-level analysis, classifier fusion and $k = 20$), where we achieve a MAP of **(0.55 - 0.65)** for the main and generalization tasks, respectively. The conclusion which can be drawn from this aggregated comparison is that our best performing setup (regardless of the type of FSP base classifier) provides more stable results, i.e., the performance in real-world scenarios will be more predictable.

Experiment *(iv)* aims at comparing our approach with the MediaEval submissions of 2015 [90]. Table 6.9 provides a comparison of our best performing FSP method with the best run of participating teams (in terms of MAP) in the MediaEval 2015 VSD task (i.e., Affective Impact of Movies – including Violence – task). The following conclusions can be drawn from the results in **Table 6.9**.

- Our SVM-based and SAE-based FSP solutions where only fine-level representations are used outperform the *run5* of Fudan-Huawei, which ranked first. The Fudan-Huawei team employs learned spatio-temporal and violence specific representations using convolutional neural network architectures.

- The SVM-based and SAE-based FSP solutions where only coarse-level representations are used achieve also very promising results (MAP of 0.2215 and 0.2548, respectively) compared to the participating teams of the MediaEval 2015 VSD task. In addition, all solutions are above the baseline methods (ran-

Table 6.9: The MAP for the best run of teams in the MediaEval 2015 VSD Task [90] and our best performing SVM-based and SAE-based FSP methods on the VSD 2015 movie dataset. Teams with at least one participant member of the MediaEval VSD organizing team are marked by *. The runs provided by the organizing team are marked by **.

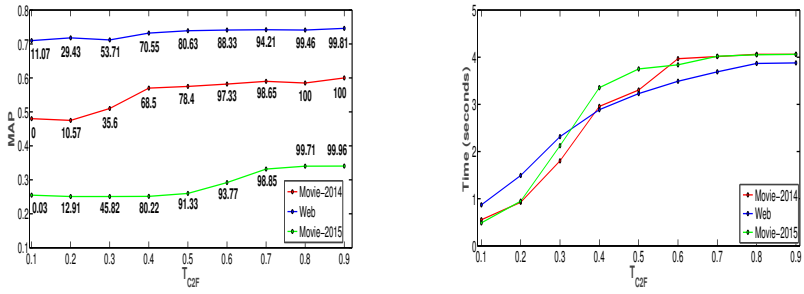
Team	**MAP**
FSP method (SAE, fine, selection,* k=*5)	**0.3413**
FSP method (SVM, fine, selection,* k=*10)	**0.3173**
Fudan-Huawei [35]	**0.2959** (*run5*)
MIC-TJU [147]*	0.2848 (*run1*)
NII-UIT [77]*	0.2684 (*run5*)
FSP method (SAE, coarse, selection,* k=*5)	**0.2548**
FSP method (SVM, coarse, selection,* k=*10)	**0.2215**
RUCMM [71]	0.2162 (*run4*)
ICL-TUM-PASSAU [122]	0.1487 (*run4*)
RFA [91]*	0.1419 (*run4*)
KIT [87]	0.1294 (*run5*)
RECOD [93]	0.1143 (*run1*)
UMons [110]	0.0967 (*run1*)
TCS-ILAB [26]	0.0638 (*run2*)
*Random (baseline)***	0.0511
*Trivial (baseline)***	0.0486

dom and trivial) provided by the MediaEval VSD organizing team in terms of MAP.

Experiment *(v)* provides results for the coarse-to-fine analysis. For a given segment, coarse-level analysis is run, and if the violence score for that segment is below a threshold (T_{C2F}), fine-level analysis is also run. We repeat the experiment with different values for the threshold, ranging from 0.1 to 0.9 (a threshold of 0 is equivalent to coarse-level analysis only, while a threshold of 1 is equivalent to fine-level analysis only). The best detectors in terms of MAP values according to the experiment *(i)* are selected (Tables 6.4, 6.5 and 6.6).

Our findings regarding the coarse-to-fine analysis are as follows:

- On the Hollywood movie dataset, setting a threshold value equal to 0.4 provides a MAP of 0.57. In other words, running the coarse-level analysis with a threshold of 0.4, helps in drastically reducing the total computations for the fine-level analysis; for such a threshold, we indeed observe that the fine-level analysis is executed on only 68.5% of the segments. Such a result means that running the coarse-level and fine-level analysis on 68.5% of the segments, provides results which are almost as good as the fine-level analysis running



(a) (b)

Figure 6.5: **(a)** Plot of the coarse-to-fine analysis threshold (T_{C2F}) vs MAP. The numbers indicated next to the points in the graph correspond to the percentage of segments for which the fine-level analysis is performed. **(b)** Average computational time per video segment of coarse-to-fine analysis with respect to the threshold (T_{C2F}), where the computational time includes raw feature extraction, feature encoding and classification. All the timing evaluations were performed with a machine with 2.40 GHz CPU and 8 GB RAM. (Movie-14: the Hollywood movie dataset, Web: the Web video dataset, Movie-15: the VSD 2015 movie dataset).

independently. Threshold values between 0.5 and 0.7 also return very accurate classifications, but for these values the gain in computation time is rather limited.

- For Web videos, the gain is even more pronounced: From Figure 6.5(a), setting a threshold value equal to 0.1 provides a MAP of 0.71 on the Web video dataset, while the fine-level analysis is performed only on 11.07% of the segments.

- Concerning the results on the VSD 2015 movie dataset (Figure 6.5(a)), in order to achieve MAP values closer to the highest MAP value of the proposed framework, the fine-level analysis is run in addition to the coarse-level analysis on more than 90% of the segments. We can realistically assume that this is due to the expression of violence in the movies of the VSD 2015 dataset which is mainly in terms of advanced visual features other than special audio effects which are usually used in the Hollywood movies, or color and textural properties of video scenes.

- Finally, we observe from Figure 6.5(b) that, on average, coarse-to-fine analysis can provide a significant gain in execution time, especially for Web videos (with a threshold of 0.1, 5 times faster and quasi-identical performance, when compared to fine analysis).

6.3.4 Computational time complexity

In this section, an evaluation of the time complexity and computational time of our system is provided. We present measures for the two main components of the system: (1) Feature generation, and (2) model learning. All the timing evaluations were performed with a machine with 2.40 GHz CPU and 8 GB RAM. For feature generation, Figure 6.6 presents average computational times calculated on the MediaEval 2014 and 2015 development datasets for raw feature extraction, sparse dictionary learning and feature encoding. The raw feature extraction part is the most time-consuming part of the whole component. Especially, the extraction time of dense trajectory and Classemes feature descriptors is respectively 14 and 7 times higher than that of MFCC descriptors.

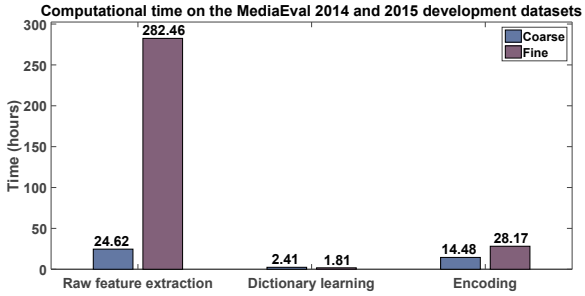


Figure 6.6: Computation times (in hours) of coarse-level (MFCC-based BoAW, color statistics and SURF-based BoVW) and fine-level (DT-based BoMW and Classemes) features on the development datasets of MediaEval VSD 2014 and 2015. *Raw feature extraction*: extraction of raw descriptors from audio and visual signals. *Dictionary learning*: the generation of vector quantization and sparse coding dictionaries. *Encoding*: the feature encoding phases for coarse and fine features. (BoAW: Bag-of-Audio-Words, BoMW: Bag-of-Motion-Words, BoVW: Bag-of-Visual-Words, DT: dense trajectory, SURF: Speeded-Up Robust Features)

Concerning model learning, Figure 6.7 provides a computational time comparison of the FSP method against unique modeling whose classification performances are discussed in detail in the previous section (Section 6.3.3). The best detectors in terms of MAP values according to the experiment *(i)* are selected (Tables 6.4 and 6.6). As presented in Figure 6.7, the model generation time is drastically reduced by using the FSP method. In simple words, this shows that training multiple base classifiers (SVM with RBF kernel or SAE) using different parts of the training data is much faster than training a single one with the whole data. In a nutshell, besides improved accuracies, FSP provides an advantage in training time.

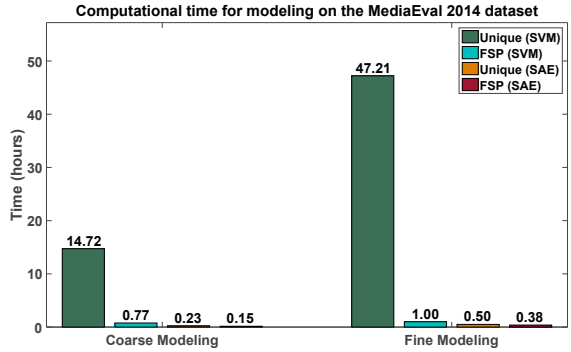


Figure 6.7: Computation times (in hours) of coarse-level and fine-level model generation with unique concept modeling and FSP, using the development dataset of MediaEval VSD 2014. The computation times of the unsupervised learning phase of the SAE are not included in the figure. These amount to 16 and 28 hours for coarse-level and fine-level modeling, respectively. (FSP: feature space partitioning, SAE: stacked auto-encoder)

6.3.5 Summary of evaluation results: The bottom line

In this section, we identify the findings consistent across the MediaEval 2014 and 2015 VSD datasets (across both movies and Web videos). This also enables us to answer the research questions posed in Section 6.1.

- Regarding *RQ1*, the discriminative power of feature space partitioning using two different base classifiers (SVM and SAE) was investigated. Our conclusions from the perspective of violence modeling are as follows:

    1. The feature space partitioning approach results in more discriminative models for violence analysis in videos regardless of the type of the base classifier. In addition, it outperforms the fundamental ensemble learning methods such as bagging and MoE.
    2. When "violence" is expressed in terms of several subconcepts, using classifier fusion as the score combination method of the feature space partitioning leads to better results in terms of MAP. On the other hand, classifier selection is a better choice as the score combination method when "violence" is more likely to correspond to a single subconcept.
    3. The optimal number of clusters when modeling violence with the feature space partitioning approach is different for different datasets. Thus, further work on automating the selection of the number of clusters is required.

- Regarding *RQ2*, we explored the coarse-to-fine analysis from the effectiveness and efficiency perspectives. Our conclusions regarding the coarse-to-fine setup are as follows:

    1. The most important conclusion is that the coarse-to-fine analysis provides an important gain in computation time without sacrificing too much accuracy. Besides, this enables a scalable solution which is adjustable depending on the processing power or accuracy requirements.
    2. From the overall results, we can also conclude that the gain in computational time for model generation is more emphasized, when the base classifier is SVM. On the other hand, the gain in computational time in the test phase of the framework seems to be similar when different base classifiers are used for the feature space partitioning.

6.4 Conclusions and Future Work

In this chapter, an effective and efficient approach for the detection of violent content in movies and Web videos was introduced. We adopted audio, static and dynamic visual representations at different abstraction levels (low-level and mid-level) which, in Chapter 5, were identified as promising for the analysis of the "violence" concept in videos. To boost performance, feature space partitioning was introduced, where we used SVM and SAE as base classifier. The partitioning was applied on coarse-level and on fine-level feature spaces, for different number of clusters, and each cluster was used to build a model designed to detect a particular violence subconcept. The combination of the classification results was realized under the classifier selection and fusion mechanisms.

The results obtained on the standardized MediaEval dataset of 2014 demonstrated that our system is on par with the state-of-the-art for the main and generalization tasks taken separately. When evaluating a given solution on both tasks simultaneously, we outperform state-of-the-art detectors, which implies that our solutions constitute more stable violence detectors (i.e., their performance can be better predicted in real world situations). The results obtained on the latest MediaEval dataset of 2015 also demonstrated that our FSP approach outperforms the state-of-the-art detectors regardless of the type of base classifier. In addition, we experimentally showed that the feature space partitioning approach outperforms the fundamental ensemble learning methods (Bagging and MoE) both on MediaEval 2014 and 2015 datasets.

Finally, in an attempt to develop a solution which can execute promptly, a coarse-to-fine analysis was introduced. This has shown that an important gain in computation time can be achieved, without sacrificing accuracy.

The results presented in this chapter impact the modeling and the practical design of concept detection. Concerning the modeling, although only demonstrated for SVM and SAE classifiers, one can foresee a gain in performance with FSP, inde-

pendently of the type of classifier used. Concerning the design, the coarse-to-fine strategy paves the road for scalable solutions with a trade-off between processing speed and accuracy. In other words, any researcher in concept detection can benefit from these findings.

We plan to further investigate the feature space partitioning and coarse-to-fine analysis components to enhance the classification performance. An interesting and important research direction is automating the selection of the number of clusters in the modeling based on feature space partitioning. Finally, we intend to evaluate the performance of classifiers other than SVM and SAE as base classifiers.

7

AFFECTIVE CONTENT ANALYSIS OF VIDEOS: CONCLUSIONS AND OUTLOOK

The novel approaches presented in this thesis were basically focused on the automatic affective content analysis of videos of various genres (music video clips, movies and Web videos to be more specific) in order to tag the videos according to their emotions. The main objectives were to close the emotional gap between raw video data and high-level emotions, and to enable the structuring of mostly unstructured or ill-structured video data from an emotional point of view. The thesis approached this goal from two different perspectives. The first part (Chapters 3 and 4) concentrated on the analysis of emotions in videos in general, whereas the second part (Chapters 5 and 6) dealt with the concept of violence in videos as a special case of video emotional analysis.

In this chapter, we review the main goals that we set for our research, and analyze to what extent these goals have been met in detail (Section 7.1). Subsequently, an overview of the possible expansions of our work is given (Section 7.2) and finally the thesis is concluded with remarks about the work presented in the thesis (Section 7.3).

7.1 Summary of the Thesis and Contributions

In the introductory chapter (Chapter 1), we defined our ultimate goal as the development of multimedia analysis solutions capable of inferring the **emotions** present in audio-visual contents, i.e., affective content analysis (Part I). As mentioned in the background chapter (Chapter 2), we followed the definition of "expected emotion" – emotion either intended to be communicated toward an audience by professionals or likely to be elicited from the majority of an audience. As a special case of affective content analysis, we considered the detection of **violent scenes** (Part II). Towards achieving this ultimate goal, we identified two research objectives: **(1)** The **representation** of audio-visual data items, and **(2)** the **modeling** of classifiers for affective content analysis. In the background chapter (Chapter 2), the basic concepts of affective analysis used in the thesis were introduced. In addition, a comprehensive literature review on affective content analysis and on the special case of violence detection in videos was provided from different perspectives such as feature representation and modeling. For the violence detection part, the MediaEval VSD task and the best performing systems which participated in the challenge since the inception of the task were presented and discussed in detail. While working towards answering these two research objectives, we directed our efforts to keep the domain knowledge about the problem or application at hand to a minimum. Besides the impacts of the various conclusions drawn in the individual chapters, keeping the domain knowledge to a minimum can be seen as the main impact of this thesis.

Concerning the ***objective (1)***, for both general affective analysis and violence detection, we considered various audio and visual features. In the general emotion analysis case, as explained in Chapter 3, the majority of existing methods either use low-level audio-visual features or generate handcrafted higher level representations. The handcrafted higher level representations are usually problem specific, i.e., hardwired for a given task. For instance, a problem specific approach would consist in detecting laughter to infer the presence of funny scenes in sitcoms or spotting horror sounds to detect horror scenes in a thriller movie. In order to effectively learn representations instead of performing feature engineering, we considered the raw audio-visual data from which, using deep learning methods, mid-level audio and static visual representations which proved to offer better discrimination than low-level and handcrafted mid-level ones were constructed; the superiority of those learned mid-level features was demonstrated both individually (i.e., when compared to other features of the same modality) and in combination. More specifically, we created convolutional neural network (CNN) architectures both for MFCC and the color data to learn a hierarchy of audio and static visual features and to construct higher level video representations. The color CNN architecture was used to encode the higher level color structure of video keyframes, whereas the MFCC CNN architecture was designed to captured both the timbre and temporal information of audio signals through 2D CNN modeling.

In Chapter 4, two major additions to the contributions in Chapter 3 were presented. First, we extended the feature set with advanced motion features – i.e., we exploited the coherence among successive frames – in a Bag-of-Words fashion. The applicability of those features for affective content analysis had not yet been investigated. Second, we experimentally verified that emotion-related learned attribute detectors help in further improving the classification performance of our system. It was also experimentally shown that learned audio representations have superior performance compared to auditory temporal modulation features which describe temporal modulations in the frequency domain.

In the special case of violence detection, a comprehensive analysis of feature representations for the same modalities as in the general case – i.e., audio, static visual and dynamic visual – of different abstraction levels – i.e., low-level and mid-level – were performed (Chapter 5) using the evaluation methodology provided by the well-established MediaEval VSD task. Mid-level representations, especially advanced motion features based on dense trajectories, were found to provide the best discrimination. In addition, combining all features yielded maximal classification performance, on par with the results of the best teams which took part to the latest MediaEval editions. These results, which were achieved without the need for any post-processing such as temporal smoothing or clip merging, are therefore at the state-of-the-art level.

Concerning the ***objective (2)***, for the general emotion analysis case, in Chapter 3, we employed a traditionally used classifier – SVM – using the representations that we derived with deep CNNs; this already provided promising results in view of existing emotional analysis algorithms (e.g., the work [145]). This chapter also saw a contribution from a fusion point of view. Because multiple feature types were used, multiple classification results were produced, which had to be aggregated. This (late) fusion was achieved with an additional SVM layer combining the outputs of the SVM classifiers. Then, in Chapter 4, those contributions were extended. First, we compared SVMs against ensemble learning in the form of decision-tree bootstrap aggregating; the outcome was that – apart from a few cases – ensemble learning revealed to be a better classifier choice for emotional analysis. Second, we assessed simple linear fusion instead of an advanced mechanism such as SVM-based fusion; the conclusion was that a simple fusion technique provides better results for the ensemble learning approach, as it does not add another modeling layer unlike SVM-based fusion.

For the special case of violence detection, we focused on the issue of modeling a given unique concept. The concept "violence" was a particularly good candidate, as annotated violent video clips are abundant thanks to the MediaEval datasets. The existing works for violent content analysis in videos model violence as a unique concept. However, due to its definition, "violence" can be expressed in diverse manners. To this end, we demonstrated in Chapter 6 that partitioning the feature space of violent data samples and training a dedicated detector for each partition provided a high gain in classification performance. As the results attest to it, this

additional modeling step transformed results (obtained in Chapter 5) which are "good" into results superior to the best MediaEval submissions. The last contribution of Chapter 6 was the coarse-to-fine processing which was designed to cope with the demanding computations caused not only by the use of multiple detectors induced by partitioning but also caused by the computations resulting from the use of advanced motion and attribute features. As an extra advantage, the coarse-to-fine setup also enables a scalable solution which is adjustable depending on the processing power or accuracy requirements.

The main intersection point of the two aforementioned objectives was to keep external input or domain knowledge about video data that is being analyzed at the affective-level to a minimum during feature representation and modeling parts of the system.

7.2 Outlook

Within the scope of this thesis, novel approaches of representation and modeling were presented to perform automatic affective video content analysis based solely on the audio-visual content of videos. Although, in this thesis, we answered the research questions which we identified, there still are possibilities for improvement. We identify below potential research directions that can be followed in the future.

In this thesis, we concentrated on the audio and visual modalities of videos. The textual modality is another source of valuable information present in videos. By textual modality, we not only refer to text that might appear in a video in the form of e.g., subtitles, but to all forms of text that can be derived from a video. For instance, the lyrics of a song are as important as the audio-visual content of the related music video clip produced for that song. Similarly, the script of a movie is as important as audio and visuals we hear and see, especially in dramas. Web videos usually have user comments attached to them. Despite the noise that they can contain, these comments can also be helpful in better understanding the content of a Web video. Hence, taking the textual modality of videos into account through the application of advanced natural language processing techniques would enable having a nearly complete "view" of a video.

In addition to feature-level fusion, weighted linear fusion and SVM-based fusion techniques have been studied in the thesis in an attempt to evaluate one simple and one advanced machine learning based decision-level fusion of multiple audio-visual features. Hybrid fusion strategies can be further studied to benefit from the advantages of both early and late fusion strategies.

Within the specific context of violent content detection in videos (Part II), we introduced a data-driven approach in Chapter 6 to model violence by partitioning the feature space into subspaces, without the manual intervention of the user, and achieved superior results. The incorporation of violence specific knowledge (i.e., violence-related concepts such as fight, explosions, etc.) into the system would allow a finer, more accurate classification. Typically, the samples stemming from the

various subconcepts could be presented to the user who, in a few clicks, could label the subconcepts, if possible, as corresponding to a given violence-related concept (e.g., "subconcept 1 corresponds to fights", "subconcept 2 corresponds to explosions", etc.) without much effort. Such a construction would provide a tremendous gain in time when modeling concepts, in comparison to traditional approaches.

The application scenarios of the system principally concern the structuring of mostly unstructured or ill-structured video data from an emotional perspective. However, the approaches presented in the thesis can be easily applied in real-world scenarios such as filtering Web content based on emotions.

The future research directions discussed above are the ones that can be performed using only the content of videos. The incorporation of information about users into the system can open up a new research path. Although, in this thesis, we discussed scenarios where unstructured data needs to be structured, the presented system can be applied in recommending videos to users according to the affect(s) of the videos based only on the audio-visual content of the videos or in a cold-start situation. This basic scenario can be personalized by including emotional preferences of users. This means that a mapping can be created between the "objective" emotional content of videos (since the content of the videos themselves are used to infer their affective meaning) and the emotional reactions of a user. Accordingly, personalized recommendations based on the emotional preference of the user can be realized.

Studying the effects of the time and social contexts on the personalization of recommendations based on emotional analysis may also be an interesting direction to follow. The emotional preferences of a user can indeed change based on the context. For instance, the user may be watching a romantic comedy alone on a Sunday night, whereas the very same user prefers to watch a science fiction movie with her friends on a Friday night. In this scenario, emotionally preferred movie suggestions should be updated accordingly.

7.3 Final Remarks

The majority of affective computing studies is exclusively focused on signals conveyed by humans (i.e., human machine interaction) and not on the multimedia items that are being consumed. We believe that multi-disciplinary research where both the fields of affective computing and multimedia content analysis are combined will lead to the construction of "interesting" perspectives for both fields. Analyzing the content of multimedia data at the affective-level is one of the two basic content analysis perspectives and is complementary to the analysis at the cognitive-level. Providing solid solutions for affective content analysis is crucial as emotions play an important role in multimedia selection and consumption. We provided robust solutions for automatic affective content analysis of videos and evaluated them with standard datasets and metrics. The solutions presented in this thesis were solely based on the audio-visual content of the videos. We strongly

believe that the incorporation of context during emotional analysis of multimedia items is the most important future direction within the field of affective content analysis.

BIBLIOGRAPHY

[1] Esra Acar. Learning representations for affective video understanding. In *ACM International Conference on Multimedia (ACMMM)*, pages 1055–1058, October 2013.

[2] Esra Acar and Sahin Albayrak. TUB-IRML at MediaEval 2014 violent scenes detection task: Violence modeling through feature space partitioning. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[3] Esra Acar, Frank Hopfgartner and Sahin Albayrak. Detecting violent content in Hollywood movies by mid-level audio representations. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 73–78, June 2013.

[4] Esra Acar, Frank Hopfgartner and Sahin Albayrak. Violence detection in Hollywood movies by the fusion of visual and mid-level audio cues. In *ACM International Conference on Multimedia (ACMMM)*, pages 717–720, October 2013.

[5] Esra Acar, Frank Hopfgartner and Sahin Albayrak. Understanding affective content of music videos through learned representations. In *International Conference on MultiMedia Modelling (MMM)*, pages 303–314, January 2014.

[6] Esra Acar, Frank Hopfgartner and Sahin Albayrak. Fusion of learned multimodal representations and dense trajectories for emotional analysis in videos. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2015.

[7] Esra Acar, Frank Hopfgartner and Sahin Albayrak. Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies. *Neurocomputing*, 208:225–237, 2016.

[8] Esra Acar, Frank Hopfgartner and Sahin Albayrak. A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material. *Multimedia Tools and Applications*, pages 1–29, 2016.

[9] Esra Acar, Melanie Irrgang, Dominique Maniry and Frank Hopfgartner. Detecting violent content in hollywood movies and user-generated videos. In *Smart Information Systems*, pages 291–314. Springer International Publishing, 2015.

[10] Rehan Akbani, Stephen Kwek and Nathalie Japkowicz. Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004*, pages 39–50. Springer, 2004.

[11] Stuart Andrews, Ioannis Tsochantaridis and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 561–568, December 2002.

[12] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, January 2007.

[13] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010.

[14] Sandra Avila, Daniel Moreira, Mauricio Perez, Daniel Moraes, Isabela Cota, Vanessa Testoni, Eduardo Valle, Siome Goldenstein and Anderson Rocha. RECOD at MediaEval 2014: Violent scenes detection task. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[15] Yoann Baveye, Jean-Noël Bettinelli, Emmanuel Dellandréa, Liming Chen and Christel Chamaret. A large video database for computational models of induced emotion. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 13–18, September 2013.

[16] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret and Liming Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 77–83, September 2015.

[17] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret and Liming Chen. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, January 2015.

[18] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[19] Yoshua Bengio, Aaron Courville and Pierre Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1798–1828, August 2013.

[20] Damian Borth, Tao Chen, Rongrong Ji and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM International Conference on Multimedia (ACMMM)*, pages 459–460, October 2013.

[21] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[22] Bree Brouwer. YouTube Now Gets Over 400 Hours Of Content Uploaded Every Minute. Accessed: 2016-04-04.

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

[24] Luca Canini, Sergio Benini and Riccardo Leonardi. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):636–647, April 2013.

[25] Diego Castán, M Rodrıguez, Alfonso Ortega, Carlos Orrite and Eduardo Lleida. ViVoLab and CVLab-MediaEval 2014: Violent scenes detection affect task. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[26] Rupayan Chakraborty, Avinash Kumar Maurya, Meghna Pandharipande, Ehtesham Hassan, Hiranmay Ghosh and Sunil Kumar Kopparapu. TCS-ILAB-MediaEval 2015: Affective impact of movies and violent scene detection. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[27] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, May 2011.

[28] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang and Chih-Wen Su. Horror video scene recognition via multiple-instance learning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1325–1328, May 2011.

[29] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang and Chih-Wen Su. Violence detection in movies. In *International Conference on Computer Graphics, Imaging and Visualization (CGIV)*, pages 119–124, August 2011.

[30] Tao Chen, Damian Borth, Trevor Darrell and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *Computing Research Repository (CoRR)*, abs/1410.8586, 2014.

[31] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia (ACMMM)*, pages 367–376, November 2014.

[32] Yue Cui, Jesse S. Jin, Shiliang Zhang, Suhuai Luo and Qi Tian. Music video affective understanding using feature importance analysis. In *Proc. ACM International Conference on Image and Video Retrieval*, pages 213–219, July 2010.

[33] Qi Dai, Jian Tu, Ziqiang Shi, Yu-Gang Jiang and Xiangyang Xue. Fudan at MediaEval 2013: Violent scenes detection using motion features and part-level attributes. In *Working Notes Proceedings of the MediaEval Workshop*, October 2013.

[34] Qi Dai, Zuxuan Wu, Yu-Gang Jiang, Xiangyang Xue and Jinhui Tang. Fudan-NJUST at MediaEval 2014: Violent scenes detection using deep neural networks. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[35] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu and Yu-Gang Jiang. Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

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

[37] Claire-Hélène Demarty, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl and Cédric Penet. Benchmarking violent scenes detection in movies. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2014.

[38] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier and Mohammad Soleymani. The MediaEval 2012 Affect Task: Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval Workshop*, October 2012.

[39] Claire-Hélène Demarty, Cédric Penet, Bogdan Ionescu, Guillaume Gravier and Mohammad Soleymani. Multimodal violence detection in hollywood movies: State-of-the-art and benchmarking. In *Fusion in Computer Vision*, pages 185–208. Springer, 2014.

[40] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Bogdan Ionescu, Vu Lam Quang and Yu-Gang Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval Workshop*, October 2013.

[41] Nadia Derbas and Georges Quénot. Joint audio-visual words for violent scenes detection in movies. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 483–486, April 2014.

[42] Nadia Derbas, Bahjat Safadi and Georges Quénot. LIG at MediaEval 2013 affect task: Use of a generic method and joint audio-visual words. In *Working Notes Proceedings of the MediaEval Workshop*, October 2013.

[43] Xinmiao Ding, Bing Li, Weiming Hu, Weihua Xiong and Zhenchong Wang. Horror video scene recognition based on multi-view multi-instance learning. In *Computer Vision–ACCV 2012*, pages 599–610. Springer, 2013.

[44] Bruno do Nascimento Teixeira. MTM at MediaEval 2014 violence detection task. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[45] Joël Dumoulin, Diana Affi, Elena Mugellini, Omar Abou Khaled, Marco Bertini and Alberto Del Bimbo. Affect recognition in a realistic movie dataset using a hierarchical approach. In *First International Workshop on Affect & Sentiment in Multimedia (ASM)*, pages 15–20, October 2015.

[46] Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[47] Jana Eggink and Denise Bland. A large scale experiment for mood-based classification of tv programmes. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 140–145, July 2012.

[48] Joseph G Ellis, W Sabrina Lin, Ching-Yung Lin and Shih-Fu Chang. Predicting evoked emotions in video. In *IEEE International Symposium on Multimedia (ISM)*, pages 287–294, December 2014.

[49] Florian Eyben, Felix Weninger, Nicolas Lehment, Björn Schuller and Gerhard Rigoll. Affective video retrieval: Violence detection in hollywood movies by large-scale segmental feature extraction. *PloS one*, 8(12):e78506, 2013.

[50] Ting fan Wu, Chih-Jen Lin and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2003.

[51] Gerald Friedland and Ramesh Jain. *Multimedia Computing*. Cambridge University Press, 2014.

[52] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou and Sergios Theodoridis. Violence content classification using audio features. *Advances in Artificial Intelligence*, 3955:502–507, May 2006.

[53] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. *Artificial Intelligence: Theories, Models and Applications*, 6040:91–100, May 2010.

[54] Yu Gong, Weiqiang Wang, Shuqiang Jiang, Qingming Huang and Wen Gao. Detecting violent scenes in movies by auditory and visual cues. In *Advances in Multimedia Information Processing – PCM*, pages 317–326. Springer, 2008.

[55] Shinichi Goto and Terumasa Aoki. Violent scenes detection using mid-level violence clustering. In *International Conference on Computer Science & Information Technology (CCSIT)*, February 2014.

[56] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image and Vision Computing*, 31(2):120–136, February 2013.

[57] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, 2005.

[58] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging*, pages 87–95, January 2003.

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

[60] Haibo He and E.A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009.

[61] Bogdan Ionescu, Jan Schlüter, Ionut Mironica and Markus Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 215–222, April 2013.

[62] Go Irie, Kota Hidaka, Takashi Satou, Akira Kojima, Toshihiko Yamasaki and Kiyoharu Aizawa. Latent topic driving model for movie affective scene classification. In *ACM International Conference on Multimedia (ACMMM)*, pages 565–568, October 2009.

[63] Go Irie, Kota Hidaka, Takashi Satou, Toshihiko Yamasaki and Kiyoharu Aizawa. Affective video segment retrieval for consumer generated videos

based on correlation between emotions and emotional audio events. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 522–525, July 2009.

[64] Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki and Kiyoharu Aizawa. Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE Transactions on Multimedia*, 12(6):523–535, November 2010.

[65] Robert A Jacobs, Michael I Jordan, Steven J Nowlan and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[66] Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163–172, 1987.

[67] Sylvie Jeannin and Ajay Divakaran. Mpeg-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):720–724, June 2001.

[68] Shuiwang Ji, Wei Xu, Ming Yang and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(1):221–231, January 2013.

[69] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia (ACMMM)*, pages 675–678, November 2014.

[70] Yu-Gang Jiang, Baohan Xu and Xiangyang Xue. Predicting emotions in user-generated videos. In *The AAAI Conference on Artificial Intelligence (AAAI)*, July 2014.

[71] Qin Jin, Xirong Li, Haibing Cao, Yujia Huo, Shuai Liao, Gang Yang and Jieping Xu. RUCMM at MediaEval 2015 affective impact of movies task: Fusion of audio and visual cues. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[72] Orit Kliper-Gross, Tal Hassner and Lior Wolf. The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(3):615–621, March 2012.

[73] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, January 2012.

[74] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, December 2012.

[75] Vu Lam, Duy-Dinh Le, Sang Phan, Shin'ichi Satoh and Duc Anh Duong. NII-UIT at MediaEval 2013 violent scenes detection affect task. In *Working Notes Proceedings of the MediaEval Workshop*, October 2013.

[76] Vu Lam, Duy-Dinh Le, Sang Phan, Shin'ichi Satoh and Duc Anh Duong. NII-UIT at MediaEval 2014 violent scenes detection affect task. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[77] Vu Lam, Sang Phan, Duy-Dinh Le, Shin'ichi Satoh and Duc Anh Duong. NII-UIT at MediaEval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[78] Ivan Laptev, Marcin Marszalek, Cordelia Schmid and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[79] Tom LH. Li, Antoni B. Chan and Andy HW. Chun. Automatic musical pattern feature extraction using convolutional neural network. In *International MultiConference of Engineers and Computer Scientists (IMECS)*, March 2010.

[80] J. Lin and W. Wang. Weakly-supervised violence detection in movies with audio and video based co-training. In *Advances in Multimedia Information Processing – PCM*, pages 930–935. Springer, 2009.

[81] Keng-Sheng Lin, Ann Lee, Yi-Hsuan Yang, Cheng-Te Lee and Homer H Chen. Automatic highlights extraction for drama video using music emotion and human face features. *Neurocomputing*, 119:111–117, 2013.

[82] Stuart P Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[83] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia (ACMMM)*, pages 83–92, October 2010.

[84] Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, March 2010.

[85] N. Malandrakis, A. Potamianos, G. Evangelopoulos and A. Zlatintsi. A supervised approach to movie emotion tracking. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2376–2379, May 2011.

[86] Dominique Maniry, Esra Acar, Frank Hopfgartner and Sahin Albayrak. A visualization tool for violent scenes detection. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 522–524, April 2014.

[87] P Marin Vlastelica, Sergey Hayrapetyan, Makarand Tapaswi and Rainer Stiefelhagen. KIT at MediaEval 2015–evaluating visual cues for affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[88] MediaEval Affect Task: Violent Scenes Detection (VSD) Task Proceedings, 2013. `http://ceur-ws.org/Vol-1043/`.

[89] MediaEval Violent Scenes Detection (VSD) Task Proceedings, 2014. `http://ceur-ws.org/Vol-1263/`.

[90] MediaEval Affective Impact of Movies (including Violence) Task Proceedings, 2015. `http://ceur-ws.org/Vol-1436/`.

[91] Ionut Mironica, Bogdan Ionescu, Mats Sjöberg, Markus Schedl and Marcin Skowron. RFA at MediaEval 2015 affective impact of movies task: A multimodal approach. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[92] Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

[93] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein and Anderson Rocha. RECOD at MediaEval 2015: Affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[94] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 331–340, February 2009.

[95] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, pages 332–339, August 2011.

[96] Jianwei Niu, Xiaoke Zhao and Muhammad Ali Abdul Aziz. A novel affect-based model of similarity measure of videos. *Neurocomputing*, 173:339–345, 2016.

[97] Lei Pang and Chong-Wah Ngo. Multimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 619–622, June 2015.

[98] Maja Pantic and Alessandro Vinciarelli. Implicit human-centered tagging [social sciences]. *Signal Processing Magazine, IEEE*, 26(6):173–180, 2009.

[99] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier and Patrick Gros. Multimodal information fusion and temporal integration for violence detection in movies. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2393–2396, March 2012.

[100] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier and Patrick Gros. Technicolor/INRIA team at the MediaEval 2013 violent scenes detection task. In *Working Notes Proceedings of the MediaEval Workshop*, October 2013.

[101] Rosalind W Picard. Affective computing. Technical report 321, MIT Media Laboratory, 1995.

[102] Robert Plutchik and Henry Kellerman. *Emotion: theory, research and experience*, volume 3. Academic press New York, NY, 1986.

[103] Ashfaqur Rahman and Brijesh Verma. Novel layered clustering-based approach for generating ensemble of classifiers. *IEEE Transactions on Neural Networks*, 22(5):781–792, May 2011.

[104] Mark Robertson. 500 Hours of Video Uploaded To YouTube Every Minute [Forecast]. Accessed: 2016-04-04.

[105] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, December 1980.

[106] Bahjat Safadi and Georges Quénot. A factorized model for multiple SVM and multi-label classification for large scale multimedia indexing. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2015.

[107] Markus Schedi, Mats Sjöberg, Ionuţ Mironică, Bogdan Ionescu, Vu Lam Quang, Yu-Gang Jiang and Claire-Hélène Demarty. VSD2014: A dataset for violent scenes detection in hollywood movies and web videos. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2015.

[108] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, December 2005.

[109] E. Schmidt, J. Scott and Y. Kim. Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 325–330, October 2012.

[110] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi and Thierry Dutoit. UMons at MediaEval 2015 affective impact of movies task including violent scenes detection. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

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

[112] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty and Liming Chen. The MediaEval 2015 Affective Impact of Movies Task. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

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

[114] Mats Sjöberg, Ionut Mironica, Markus Schedl and Bogdan Ionescu. FAR at MediaEval 2014 violent scenes detection: A concept-based fusion approach. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[115] Cees GM Snoek, Marcel Worring and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *ACM International Conference on Multimedia (ACMMM)*, pages 399–402, November 2005.

[116] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[117] Mohammad Soleymani, Anna Aljanaki, Frans Wiering and Remco C Veltkamp. Content-based music recommendation using underlying music preference structure. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6, June 2015.

[118] Bob L Sturm and Pardis Noorzad. On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In *International Symposium on Computer Music Modeling and Retrieval*, pages 379–394, June 2012.

[119] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[120] Chun Chet Tan and Chong-Wah Ngo. The Vireo team at MediaEval 2013: Violent scenes detection by mid-level concepts learnt from YouTube. In *Working Notes Proceedings of the MediaEval Workshop*, October 2013.

[121] Lorenzo Torresani, Martin Szummer and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *Proc. European Conference on Computer Vision (ECCV)*, pages 776–789, September 2010.

[122] George Trigeorgis, Eduardo Coutinho, Fabien Ringeval, Erik Marchi, Stefanos Zafeiriou and Björn Schuller. The ICL-TUM-PASSAU approach for the MediaEval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[123] Jian Tu, Zuxuan Wu, Qi Dai, Yu-Gang Jiang and Xiangyang Xue. Challenge huawei challenge: Fusing multimodal features with deep neural networks for mobile video annotation. In *IEEE International Conference on Multimedia and Expo (ICME) Workshops*, pages 1–6, July 2014.

[124] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394–409, 1994.

[125] Andrea Vedaldi and Brian Fulkerson. VLFeat: An open and portable library of computer vision algorithms. In *ACM International Conference on Multimedia (ACMMM)*, pages 1469–1472, October 2010.

[126] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, June 2006.

[127] Heng Wang, A. Klaser, C. Schmid and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176, June 2011.

[128] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558, December 2013.

[129] Shangfei Wang and Qiang Ji. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, 6(4):410–430, October 2015.

[130] Matthias Wimmer, Björn Schuller, Dejan Arsic, Gerhard Rigoll and Bernd Radig. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 145–151, January 2008.

[131] Ting-Fan Wu, Chih-Jen Lin and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, December 2004.

[132] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM International Conference on Multimedia (ACMMM)*, pages 461–470, October 2015.

[133] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *Computing Research Repository (CoRR)*, abs/1511.04798, 2015.

[134] Can Xu, Suleyman Cetintas, Kuang-Chih Lee and Li-Jia Li. Visual sentiment prediction with deep convolutional neural networks. *Computing Research Repository (CoRR)*, abs/1411.5731, 2014.

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

[136] Min Xu, Jesse S Jin, Suhuai Luo and Lingyu Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *ACM International Conference on Multimedia (ACMMM)*, pages 677–680, October 2008.

[137] Min Xu, Namunu C Maddage, Changsheng Xu, Mohan Kankanhalli and Qi Tian. Creating audio keywords for event detection in soccer video. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 2, pages II–281–4 vol.2, July 2003.

[138] Min Xu, Jinqiao Wang, Xiangjian He, Jesse S Jin, Suhuai Luo and Hanqing Lu. A three-level framework for affective content analysis and its case studies. *Multimedia Tools and Applications (MTAP)*, 70(2):757–779, May 2014.

[139] Rong Yan and Milind Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 657–663, June 2005.

[140] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, Oswald Lanz, and Nicu Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(6):1070–1083, 2016.

[141] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu and Nicu Sebe. Multitask linear discriminant analysis for view invariant action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014.

[142] Yan Yan, Yi Yang, Deyu Meng, Gaowen Liu, Wei Tong, Alexander G. Hauptmann and Nicu Sebe. Event oriented dictionary learning for complex event detection. *IEEE Transactions on Image Processing*, 24(6):1867–1878, 2015.

[143] Jianchao Yang, Kai Yu, Yihong Gong and Tingwen Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, June 2009.

[144] Xiaowei Yang, Kuansan Wang and Shihab A Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38(2):824–839, March 1992.

[145] Ashkan Yazdani, Krista Kappeler and Touradj Ebrahimi. Affective content analysis of music video clips. In *ACM International Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM)*, pages 7–12, November 2011.

[146] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 492–497, October 2009.

[147] Yun Yi, Hanli Wang, Bowen Zhang and Jian Yu. MIC-TJU in MediaEval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*, September 2015.

[148] Zeynep Yucel and Albert Ali Salah. Resolution of focus of attention using gaze direction estimation and saliency computation. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–6, September 2009.

[149] Bowen Zhang, Yun Yi, Hanli Wang and Jian Yu. MIC-TJU at MediaEval violent scenes detection (VSD) 2014. In *Working Notes Proceedings of the MediaEval Workshop*, October 2014.

[150] Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao and Qi Tian. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 12(6):510–522, 2010.

[151] Sicheng Zhao, Hongxun Yao, Xiaoshuai Sun, Xiaolei Jiang and Pengfei Xu. Flexible presentation of videos based on affective content analysis. In *International Conference on MultiMedia Modelling (MMM)*, pages 368–379, January 2013.

[152] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC Press, 2012.