

Recognizing Film Aesthetics, Spectators' Affect and Aesthetic Emotions from Multimodal Signals

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par

Michal MUSZYNSKI

de

Wloszczowa (Poland)

Thèse N° 5298

GENÈVE

Repro-Mail - Université de Genève

2018



**UNIVERSITÉ
DE GENÈVE**

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION INFORMATIQUE

Thèse de Monsieur Michal MUSZYNSKI

intitulée :

«Recognizing Film Aesthetics, Spectators' Affect and Aesthetic Emotions from Multimodal Signals»

La Faculté des sciences, sur le préavis de Monsieur T. PUN, professeur ordinaire et directeur de thèse (Département d'informatique), Monsieur G. CHANEL, docteur et codirecteur de thèse (Département d'informatique), Monsieur S. MARCHAND-MAILLET, professeur associé (Département d'informatique), Madame N. BERTHOUBE, professeur (Interaction Center, University College London, United Kingdom), Monsieur L. CHEN, professeur (Département mathématiques - informatique, Ecole Centrale de Lyon, Université de Lyon, Ecully, France), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 20 décembre 2018

Thèse - 5298 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

I would like to dedicate this thesis to my loving brother, parents, friends, and everyone who helped me in my life ...

"The bringing together of theory and practice leads to the most favourable results; not only does practice benefit, but the sciences themselves develop under the influence of practice, which reveals new subjects for investigation and new aspects of familiar subjects."

P.L. Chebyshev

Acknowledgements

These four years of my PhD studies that were an unforgettable life experience I met many outstandingly smart and good people.

My first thanks go to my supervisors Prof. Thierry Pun and Dr. Guillaume Chanel who gave me a lot of valuable advice and freedom while carrying out my doctoral research. Because of that, I become a mature researcher and acquire unique skills. I also thank Dr. Theodoros Kostoulas and Prof. Patrizia Lombardo for their support in writing papers and their guidance.

I am grateful for being able to collaborate with Dr. Leimin Tian. I have certainly learned from you a lot.

I am also grateful to my thesis committee members: Prof. Stéphane Marchand-Maillet, Prof. Nadia Berthouze, and Prof. Liming Chen for evaluating my research and their valuable comments.

I thank all former and current CVML laboratory members for their friendliness and helpfulness all these years, in random order: Dr. Mohammad Soleymani, Dr. Guido Bologna, Dr. Séverine Cloix, Dr. Anna Aljanaki, Dr. Phil Lopes, Sunny Avry, Soheil Rayatdoost, Chen Wang, Viviana Weiss, Lara Broi, Coralie Grossrieder, Prof. Alexandros Kalousis, Dr. Edgar Roman-Rangel (Paco), Dr. Olivier Schwander, Dr. Ke Sun, Dr. Magda Gregorova, Pablo Strasser, Amina Mollaysa, Lionel Blondé, Frantzeska Lavda, Dr. Grigorios Anagnostopoulos, Jason Ramapuram, Prof. Sviatoslav Voloshynovskiy, Dr. Taras Holotyak, Dr. Maurits Diephuis, Dr. Sohrab Ferdowsi, Dr. Dimche Kostadinov, Thomas Charlon, Olga Taran, Shideh Rezaeifar, Denis Ullmann, and Behrooz Razeghi.

I cannot miss thanking my brother, Grzegorz and my parents, Olga and Wojciech for their love, support, and generosity all these years. The moral values that I have been taught make me who I am and who I want to be.

Last but not least, I would like to thank you all my friends from all over the world who I met in Geneva all these years for all discussions, support, and adventures.

Résumé

Les expériences esthétiques sont communes dans nos vies. Malgré cela, les processus impliqués dans la génération de ces expériences ne sont pas entièrement compris. De plus, il n'existe pas de théorie exhaustive capable de définir et d'expliquer l'expérience esthétique dans l'art. Le défi consiste principalement à comprendre les différentes étapes du traitement de l'information esthétique, telles que l'analyse perceptuelle, les processus cognitifs et l'évaluation qui donnent lieu à des jugements esthétiques et émotionnels.

L'objectif de cette thèse est d'analyser l'expérience esthétique de plusieurs spectateurs. Nous visons en particulier à détecter les moments fortement esthétiques des films, ainsi qu'à reconnaître les émotions ressenties par les spectateurs. Les résultats de cette recherche peuvent être utilisés pour la détection et la conception de scènes esthétiques et émotionnelles, pour résumer des vidéos, ou encore pour faire une prédiction affective et esthétiques des contenus visuels.

Cette thèse offre tout d'abord un parcours de l'état de l'art concernant l'expérience esthétique dans le contexte des films cinématographiques. Les émotions "*quotidiennes*" et les émotions esthétiques sont définies, et une distinction claire est établie entre les émotions induites et celles perçues par les spectateurs. Plusieurs représentations émotionnelles et leurs caractéristiques sont discutées. Le concept de synchronisation interpersonnelle pour la visualisation de films en groupe est déterminé. Une revue de la littérature sur l'analyse du contenu esthétique et affectif des vidéos est présentée. Les travaux existants sur la détection d'émotions, ainsi que sur la détection de moments forts à partir des contenus vidéo et des réactions des spectateurs, sont décrits et discutés. Finalement, les principales limites des recherches existantes sont soulignées.

Les bases de données multimédia esthétiques et affectives actuellement disponibles sont ensuite décrites en détail. La base de données LIRIS-ACCEDE créée pour étudier l'expérience émotionnelle d'un film cinématographique est sélectionnée et étendue pour étudier l'expérience esthétique. Différents protocoles qui permettent de collecter des annotations qui décrivent l'expérience esthétique et les émotions perçues par le public durant la visualisation d'un film sont décrits. L'analyse statistique de ces annotations est effectuée.

Ce travail démontre que les moments forts d'un film sont capables d'induire un large spectre d'émotions. L'intensité de ces émotions, décrites par un niveau d'excitation (*arousal*) et d'hédonicité (*valence*), dépend fortement de la catégorie esthétique à laquelle appartient le moment considéré et le genre de film. Les méthodes et les résultats sur la détection des moments fortement esthétiques, basés sur le niveau de synchronisation entre l'activité électrodermale (EDA) et l'accélération (ACC), sont aussi présentés. Les résultats suggèrent que le niveau de synchronisation entre les signaux des spectateurs, que ce soit pour les signaux ACC ou EDA, est discriminant pour la détection des moments fortement esthétiques. En particulier, les mesures de synchronisation par paires sont les plus stables et elles permettent d'obtenir la meilleure performance pour la détection des moments fortement esthétiques indépendamment du genre et de la catégorie de film.

La relation entre les émotions induites et perçues par les spectateurs des films cinématographique est examinée. Une incohérence dans les annotations des émotions induites et perçues est observée. En particulier, on constate que les émotions induites et perçues des spectateurs ne sont pas toujours positivement corrélées. Il est aussi observé que les émotions perçues et induites sont caractérisées par des moments esthétiquement forts. Finalement, les émotions induites sont détectées à partir des mesures EDA et ACC du spectateur, ainsi que du contenu du film. À cette fin, nous constatons que les modèles *Long Short-Term Memory Recurrent Neural Network* (LSTM-RNN) surpassent les modèles de *Support Vector Regression* (SVR) et les *Deep Belief Network* (DBN) à cause de leur capacité à prendre en compte les informations temporelles et à combiner hiérarchiquement les informations multimodales (ACC, EDA, les indices émotionnels et les caractéristiques audiovisuelles).

Ce travail montre que les moments fortement esthétiques induisent des émotions esthétiques, au-delà des émotions "*quotidiennes*". Les émotions esthétiques ressenties par les spectateurs sont différentes suivant la catégorie des moments forts, ainsi que suivant le genre du film. Plus particulièrement les émotions esthétiques d'un film ne peuvent pas être décrites avec précision dans l'espace *valence-arousal* comme peuvent l'être les émotions "*quotidiennes*". Nous avons trouvé quatre dimensions émotionnelles pour représenter avec précision les émotions esthétiques. L'influence de la personnalité sur les émotions esthétiques est évaluée car des différences dans la classification des scènes esthétiques sont observées suivant la personnalité des participants. Finalement, on démontre que les émotions esthétiques peuvent être prédites à partir des réactions des spectateurs (EDA et ACC).

Pour résumer, ces résultats permettent de comprendre les processus impliqués dans l'expérience esthétique d'un film. Toutefois, la compréhension de l'expérience esthétique est une tâche difficile en raison de sa complexité et de sa subjectivité. L'expérience esthétique est influencée par plusieurs facteurs, tels que la personnalité, l'expérience personnelle, l'humeur,

l'intérêt qui sont difficiles à quantifier objectivement. En conclusion, l'expérience esthétique d'un film ne peut pas être étudiée sans prendre en compte les réactions multimodales des spectateurs dans des conditions naturelles, par exemple lorsqu'ils regardent un film en groupe dans une salle de cinéma.

Abstract

Even though aesthetic experiences are common in our lives, processes involved in aesthetic experience are not fully understood. Moreover, there is no comprehensive theory that explains and defines the concept of aesthetic experience in art. The challenge of studies on aesthetic experiences is to understand different stages of aesthetic information processing, such as perceptual analysis, cognitive processes, and evaluation resulting in aesthetic judgments and emotions.

The main goal of this thesis is to analyse film aesthetic experience evoked in spectators. In particular, we aim to detect aesthetic highlights in movies, as well as recognize induced emotions and aesthetic emotions elicited in spectators. The outcomes of the research on induced emotions, aesthetic emotions, and aesthetic highlights can be used for emotional and aesthetic scene detection, emotional and aesthetic scene design, video summarization, and prediction of affective and aesthetic content.

In this thesis, a background review on film aesthetic experience is provided. "*Everyday*" and aesthetic emotions are defined and a clear distinction between induced and perceived emotions of movie audiences is made. Several emotion representations and the characterization of emotion elicitation are discussed. The concept of interpersonal synchronization with regard to watching movies together is determined. An extensive literature review on aesthetic and affective content video analysis is also provided. Existing work on aesthetic and affect recognition as well as highlight detection from video content and spectators' reactions is described and discussed. The main limitations of the existing state of the art research are emphasized.

Currently available aesthetic and affective multimedia databases are described in details. The continuous LIRIS-ACCEDE database that was created to study film emotional experience in a movie theater is selected and extended to study film aesthetic experience. Protocols for collecting annotations of aesthetic highlights in movies, perceived emotions and aesthetic emotions felt by movie audiences are described. The statistical analysis of the annotations is carried out.

It is shown that aesthetic highlights in movies elicit a wide range of emotions. The amount of these emotions (a level of arousal and valence intensity) strongly depends on

the aesthetic highlight category and on the movie genre. Also, methodology and results of aesthetic highlight detection based on the level of synchronization among spectators' electrodermal activity (EDA) and acceleration (ACC) measurements are presented. The results suggest that the level of synchronization among spectators' EDA and ACC signals is discriminative for aesthetic highlight detection in the context of watching movies together. In particular, pairwise synchronization measures are stable measures of synchronization and achieve the best performance of aesthetic highlight detection independently of movie genre and highlight categories.

The relationship between induced and perceived emotions of movie audiences is investigated. An inconsistency in induced and perceived emotion annotations is observed. In particular, it is found that induced and perceived emotions of movie audiences are not always positively correlated. Furthermore, it is observed that both perceived and induced emotions are characterized by aesthetic highlights. Finally, induced emotions are recognized from spectators' EDA and ACC measurements as well as movie content. To this end we find that Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) models outperform Support Vector Regression (SVR) and Deep Belief Network (DBN) models because their ability to take into account temporal information and hierarchically combine knowledge-inspired affective cues with audio-visual movie content and movie audience responses.

It is shown that aesthetic highlights in movies evoke aesthetic emotions in spectators that are beyond "*everyday*" emotions. Aesthetic emotions that are felt by spectators are associated with the category of aesthetic highlights as well as the movie genre. In fact, movie aesthetic emotions cannot be accurately described in the arousal-valence space like "*everyday*" emotions. Four emotional dimensions that can accurately represent aesthetic emotions are found. Furthermore, the influence of personality on aesthetic emotions is assessed by noticing the differences in aesthetic scene ratings with regard to personality traits. Also, it is shown that aesthetic emotions can be predicted based on spectators' reactions (EDA and ACC signals).

To summarize, these promising results allow researchers to better understand processes involved in film aesthetic experience. Nevertheless, understanding of film aesthetic experience is a challenging task due to its complexity and subjective nature. Film aesthetic experience is influenced by several factors, such as personality, life experience, mood, and interest that are difficult to objectively quantify. The conclusion can be made that film aesthetic experience cannot be investigated without taking into account multimodal reactions of spectators in naturalistic conditions, e.g., watching movies together in a movie theater.

Table of contents

List of figures	xvii
List of tables	xix
Nomenclature	xxi
1 Introduction	1
1.1 Film aesthetic experience	1
1.1.1 "Everyday" emotions	5
1.1.2 Perceived vs. induced emotions	5
1.1.3 Aesthetic emotions	6
1.1.4 Emotion representation	7
1.2 Emotion elicitation and aesthetic highlights in films	8
1.3 Interpersonal synchronization	10
1.4 Motivation	11
1.5 Research questions	11
1.6 Contributions	12
1.7 Thesis structure	13
2 State of the art in aesthetics and affect recognition	15
2.1 Aesthetic and affect recognition based on video content	16
2.2 Aesthetics and affect recognition based on spectators' reactions	20
2.3 Highlight detection	25
2.3.1 Highlight detection from multimedia content	27
2.3.2 Highlight detection from spectators' reactions	30
2.4 Limitations of the existing research	30
3 Affective and aesthetic corpus development	33
3.1 Existing aesthetic and affective multimedia databases	33

3.2	Stimulus selection	39
3.3	Aesthetic highlight annotations	40
3.4	Perceived emotion annotations	43
3.4.1	Transcription and affective cue annotation	44
3.4.2	Annotations of disfluency and non-verbal vocalisation in movie dialogues	44
3.4.3	Annotating perceived movie emotions	46
3.4.4	Agreement on perceived and induced emotion annotation	49
3.5	Aesthetic emotion annotations	50
4	Aesthetic highlight detection in movies from synchronization of spectators' reactions	55
4.1	Detection system of aesthetic highlights in movies	56
4.2	Synchronization measures	58
4.2.1	Pairwise synchronization	59
4.2.2	Group synchronization	63
4.2.3	Overall synchronization	66
4.3	Results	67
4.3.1	Emotions and aesthetic highlights	67
4.3.2	Dependencies between synchronization measures	70
4.3.3	Aesthetic highlight detection	71
4.4	Discussion and conclusions	74
5	Studying the relationship between induced and perceived emotions of movie audiences	77
5.1	Multimodal feature extraction	78
5.1.1	Movie audience reaction based features	79
5.1.2	Movie content based features	80
5.2	Recognition models	82
5.2.1	Long short-term memory recurrent neural networks	83
5.2.2	Deep belief networks	85
5.2.3	Support vector regression	86
5.3	Experimental results	86
5.3.1	Perceived and induced emotions	86
5.3.2	Perceived and induced emotions vs. aesthetic highlights	88
5.3.3	Induced emotion recognition	90
5.4	Discussion	95

5.4.1	Limitations of our study	96
5.4.2	Available modalities and sample size	96
5.4.3	Model selection	97
5.5	Conclusion	98
6	Exploring aesthetic emotions of movie audiences	101
6.1	Physiological and behavioural feature extraction	102
6.1.1	Statistical features	103
6.1.2	Wavelet features	103
6.1.3	Synchronization features	103
6.2	Recognition model for aesthetic emotions	104
6.3	Results	105
6.3.1	Dependencies between aesthetic and " <i>everyday</i> " emotions	105
6.3.2	Dependencies between aesthetic highlights and aesthetic emotions	108
6.3.3	Personality and aesthetic preferences	112
6.3.4	Aesthetic emotion recognition	115
6.4	Discussion and conclusions	117
7	Conclusions and perspectives	119
7.1	Outcomes of the research	119
7.2	Responding to research questions	121
7.3	Lessons learned	122
7.4	Conclusions and perspectives	123
	References	127
	Appendix A List of publications	143

List of figures

1.1	An example of an aesthetic object with ambiguous concepts: Picasso P. (1942), <i>Bull's Head</i> , seat and handlebars of a bicycle.	2
1.2	The model of film aesthetic experience that is obtained from Leder's model of aesthetic experience of modern art [121]. Solid lines denote the connections between different components and arrows indicate the direction of influence. The dashed lines denote the relationships investigated by the individual research questions of this thesis.	3
3.1	The 5 categories of aesthetic highlights in movies [141].	40
3.2	A snapshot of aesthetic highlight annotations in the movie <i>After the Rain</i> . . .	41
3.3	Statistical analysis of aesthetic highlights annotated in the C. LIRIS-ACCEDE database, the distribution of : (a) the numbers of the particular highlight category per movie, (b) the duration of the particular highlight category per movie.	41
3.4	Overall Cohen's kappa coefficient between different categories of aesthetic highlights in movies: strong overlaps (yellow) and weak overlaps (purple). . .	42
3.5	The Pearson's Correlation Coefficient (CC) and the Concordance Correlation Coefficient (CCC) between start and end timings of DISfluency and Non-verbal Vocalisation (DIS-NV) annotations (from Leimin Tian [142, 192]). . .	45
3.6	A snapshot of perceived emotion annotations in the movie <i>Spaceman</i> on Amazon Mechanical Turk.	47
3.7	Mean variance of perceived and induced emotion annotations.	49
3.8	A snapshot of aesthetic emotion annotations in the movie <i>Islands</i> on Amazon Mechanical Turk.	50
3.9	Mean variance of aesthetic emotion annotations over the C. LIRIS-ACCEDE database.	51
3.10	Mean value of merged aesthetic emotion annotations over the C. LIRIS-ACCEDE database.	53

3.11	Mean absolute (abs.) derivative of merged aesthetic emotion annotations over the C. LIRIS-ACCEDE database.	53
4.1	The scheme of an unsupervised highlight detection system based on synchronization among spectators' physiological or behavioural signals [141]. . . .	57
4.2	An example of: (a) the filtered Vietoris-Rips complex, (b) a number of connected components for different values of filtration parameter, (c) the persistent diagram [140].	64
4.3	The weighted average CC effect size between the synchronization measures (yellow and purple color indicate strong correlation and anti-correlation, respectively). The synchronization measures are computed over spectators': (a) EDA signals, (b) ACC signals [141].	70
5.1	Structure of HL model using movie based features [142, 192].	83
5.2	Structure of HL model using all features [142, 192].	84
5.3	The weighted average of the CC between perceived (Per) and induced (Ind) emotional dimensions of movie audience (yellow and purple color indicate strong correlation and anti-correlation, respectively) [142, 192].	87
6.1	The weighted average of the CC between the big five personality traits: extraversion, agreeableness, conscientiousness, neuroticism, and openness (yellow and purple color indicate strong correlation and anti-correlation, respectively).	113

List of tables

2.1	The summary of previous work on aesthetic and affective characterization of video content.	17
2.2	The summary of previous work on aesthetics and affect recognition based on spectators' physiological and behavioural reactions.	21
2.3	The summary of previous work on video highlight detection.	26
3.1	The summary of existing affective and aesthetic databases.	35
3.2	Detailed statistics of 8 movies selected from the C. LIRIS-ACCEDE movies (from Leimin Tian [142, 192]).	43
3.3	Movie transcript annotation agreement: the Pearson's Correlation Coefficient (CC) and the Concordance Correlation Coefficient (CCC) between start and end timings of utterances and words (from Leimin Tian [142, 192]).	44
3.4	The amount of DISfluency and Non-verbal Vocalisation (DIS-NV) annotations in movie dialogues (from Leimin Tian [142, 192]).	46
3.5	Mean level of movie audience's perceived emotions per movie [142].	48
4.1	The weighted average effect size (fixed-effect model) of arousal and valence during aesthetic highlights over all the C. LIRIS-ACCEDE database [141].	68
4.2	The weighted average effect size (fixed-effect model) of arousal during aesthetic highlights calculated per movie genre [141].	69
4.3	The weighted average effect size (fixed-effect model) of valence during aesthetic highlights calculated per movie genre [141].	69
4.4	Performance (AUC) of our highlight detection system evaluated per category of aesthetic highlights and movie genre, different synchronization measures applied to EDA signals of spectators [141].	72
4.5	Performance (AUC) of our highlight detection system evaluated per category of aesthetic highlights and movie genre, different synchronization measures applied to ACC signals of spectators [141].	73

5.1	Dependencies between aesthetic highlights and perceived and induced emotions of movie audience (small, medium, and large magnitudes of the overall effect in bold) [142].	89
5.2	Performance of unimodal induced emotion recognition using SVR, DBN, and LSTM-RNN models [142].	92
5.3	Performance of multimodal induced emotion recognition from movie content based features using SVR, DBN, and LSTM-RNN models [142].	94
5.4	Performance of multimodal induced emotion recognition from audience reaction and movie content based features using SVR, DBN, and LSTM-RNN models [142].	95
6.1	The weighted average effect size of the CC (fixed-effect model) between 4 Principal Components (PCs) and induced arousal and valence dimensions calculated over the C. LIRIS-ACCEDE database.	106
6.2	The weighted average effect size of the CC (fixed-effect model) between 4 Principal Components (PCs) and induced arousal dimension calculated per movie genre.	107
6.3	The weighted average effect size of the CC (fixed-effect model) between 4 Principal Components (PCs) and induced valence dimension calculated per movie genre.	107
6.4	The weighted average effect size (fixed-effect model) of the five aesthetic emotions during aesthetic highlights over all the C. LIRIS-ACCEDE database.	109
6.5	The weighted average effect size (fixed-effect model) of awe intensity during aesthetic highlights calculated per movie genre.	109
6.6	The weighted average effect size (fixed-effect model) of boredom intensity during aesthetic highlights calculated per movie genre.	110
6.7	The weighted average effect size (fixed-effect model) of disgust intensity during aesthetic highlights calculated per movie genre.	110
6.8	The weighted average effect size (fixed-effect model) of being touched intensity during aesthetic highlights calculated per movie genre.	110
6.9	The weighted average effect size (fixed-effect model) of wonder intensity during aesthetic highlights calculated per movie genre.	110
6.10	The weighted average of the CC between the Big-Five Personality Traits and average annotation ratings of aesthetic emotions.	114
6.11	Performance of unimodal aesthetic emotion recognition using DBNs.	115

Nomenclature

Acronyms / Abbreviations

ACC ACCeleration

ANOVA ANalysis Of VAriance

AUC Area Under the Curve

BBRBM Bernoulli-Bernoulli Restricted Boltzmann Machine

BN Bayesian Network

PB Blood Pleasure

BVP Blood Volume Pulse

CC Pearson's Correlation Coefficient

CCC Concordance Correlation Coefficient

CCRF Continuous Conditional Random Field

CMMC Combining Multiple Measures and Clustering

CNN Convolutional Neural Network

CSA CrowdSourced Annotation

CSP Common Spatial Pattern

DBN Deep Belief Network

DIS-NV DISfluency and Non-verbal Vocalisation

DL Decision-Level fusion

DTW Dynamic Time Warping

ECG ElectroCardioGram

EDA EletroDermal Activity

EEG ElectroEncephaloGram

EG Eye Gaze

ELM Extreme Learning Machine

EMG ElectroMyoGram

EOG ElectroOculoGram

FDA Fisher Discriminant Analysis

FE Facial Expression

FIS Fuzzy Inference System

FL Feature-Level fusion

FSM Facial and Shoulder Motion

GBRBM Gaussian-Bernoulli Restricted Boltzmann Machine

GP Gaussian Process

HIT Human Intelligent Task

HL Hierarchical fusion

HMM Hidden Markov Model

HR Heart Rate

IAPro Individual Affective Profile

LLD Low-Level Descriptor

LLRM Latent Linear Ranking Model

LP Linear Regression

BLSTM-RNN Bidirectional Long Short-Term Memory Recurrent Neural Network

LSTM-RNN Long Short-Term Memory Recurrent Neural Network

LTDM Latent Topic Driving Model

MAP Mean Average Precision

MAPro Mean Affective Profile

MBH Motion Boundary Histogram

MEG MagnetoEncephaloGram

MEM Maximum Entropy Model

MFCC Mel Frequency Cepstral Coefficients

MLP MultiLinear Regression

MSE Mean Square Error

NB Naive Bayes

NDL Normalized Damerau-Levenshtein

NI Nonlinear Interdependence

PAS Phase-Amplitude Synchronization

PC Principal Component

PDRM Pairwise Deep Ranking Model

PL Physiological Linkage

PLP Perceptual Linear Predictive

PSD Power Spectral Density

PS Periodicity Score

QDA Quadratic Discriminant Analysis

RBF Radial Basis Function

RBM Restricted Boltzmann Machine

Res Respiration

ROC Receiver Operating Characteristic

RSVD Reduced Singular Value Decomposition

RVM Relevance Vector Machine

S-COR S-estimator with CORrelation covariance matrix

S-DM S-estimator with Diffusion Map covariance matrix

S-HK S-estimator with Heat Kernel covariance matrix

S-PLV S-estimator with Phase Locking Value covariance matrix

S-WMI S-estimator with Windowed Mutual Information covariance matrix

SDD Shape Distribution Distance

SG Shoulder Gesture

SRCC Spearman's Rank Correlation Coefficient

STFT Short Time Fourier Transform

SVM Support Vector Machine

SVR Support Vector Regression

Temp Temperature

WMGSRP Weighted Mean Galvanic Skin Response Profile

WMI Windowed Mutual Information

WPT Wavelet Packet Transform

Chapter 1

Introduction

In this chapter, we provide an overview of aesthetic experience, emotions, emotion representations, emotion elicitation, aesthetic highlights, and interpersonal synchronization. Also, we discuss our motivation to study film aesthetic experience. We then formulate research questions on film aesthetic experience. Finally, we summarize the contributions of this thesis and describe its structure.

1.1 Film aesthetic experience

Aesthetic experience is one of the most substantial but also one of the most weakly defined concepts in art. For example, Marković [130] proposed the following definition of aesthetic experience: *"It can be defined as a special kind of relationship between a person and an artistic object in which a particular object absorbs the person's mind and overshadows other surrounding objects and events"*. Aesthetic experience is also considered to be the subjective part of an artistic exposure and corresponds to a feeling of being engaged with a piece of art. An aesthetic experience is different from the everyday experiences [53, 164], since it is supposed to be a special state of mind in which the attention of a person is focused on an artistic object while all other common objects, events, and everyday concerns are overshadowed.

There are several different definitions of aesthetic experience, for example, it can be defined as an effortless mental energy flow induced by the awareness of agreement between incoming information and our goals [48, 49]. Aesthetic experience can be associated with the concept of peak experience that assumes attention is fully focused on a particular object. The object is seen as separated from its everyday purpose [131]. Also, aesthetic experience can be linked with the concept of absorption that refers to having episodes of amplified attention [189]. In [111], aesthetic experience is defined as the perception and understanding

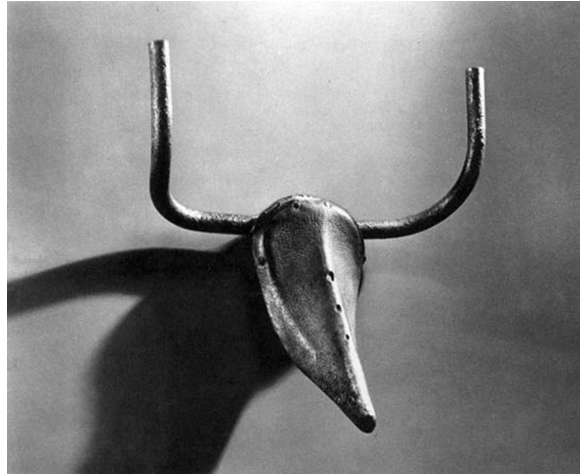


Fig. 1.1 An example of an aesthetic object with ambiguous concepts: Picasso P. (1942), *Bull's Head*, seat and handlebars of a bicycle.

of creative processes occurring in art. The creative action can happen when ambiguous concepts are assembled into a new whole object. For example, an old bicycle seat is mounted next to handlebars to create the *Bull's Head* of Pablo Picasso, as shown in Figure 1.1.

Leder et al. [121] investigated a link between aesthetics and emotions and proposed a model of aesthetic and emotional experience of modern art. Figure 1.2 presents a model of film aesthetic experience derived from Leder's model. The model contains 3 stages of aesthetic information processing: **perceptual analysis**, **cognitive processes**, and **evaluation**.

An art object (film, video, and clip) is the input of the model (arrow 1). To have aesthetic experience, audiences have to pre-classify an object as art taking into account contextual information (arrow 2). For example, the appearance of an object in an art place, such as cinemas, theaters, galleries, and museums is likely to start aesthetic processing. At the beginning, the piece of art is perceptually considered. Basic visual processing is mainly involved to analyze colors, symmetry, order, and complexity independently of exposure time [117]. Aesthetic information processing involves analysis of content and form, e.g., dialogues between main characters in movies and special effects in spectacular scenes.

When the knowledge of an art object is limited, the output of this process is mainly focused on the content part, such as stories, symbolism, and metaphors. With an increase in art knowledge, the initial representation of an artwork is shifted from the narrative part to art-specific compositions, such as physical features and structural regularities (e.g., symmetry, shape, and curvature). Some features (e.g., symmetry and order) of artworks are attractive and essential because they appeal the identification process of visual stimuli [206]. As a result, audiences are able to perceive more details.

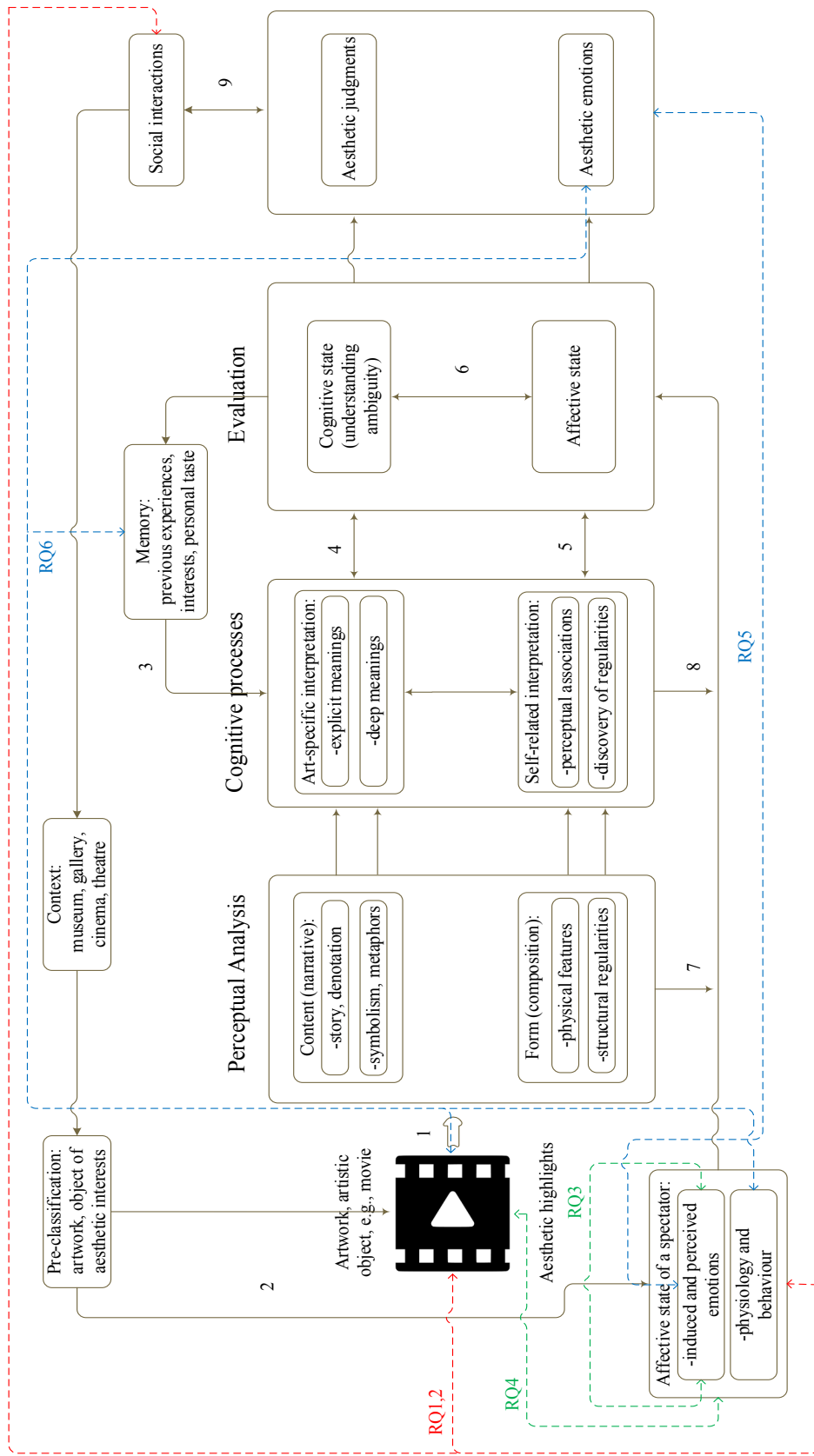


Fig. 1.2 The model of film aesthetic experience that is obtained from Leder's model of aesthetic experience of modern art [121]. Solid lines denote the connections between different components and arrows indicate the direction of influence. The dashed lines denote the relationships investigated by the individual research questions of this thesis.

Aesthetic information processing of artistic objects depends on a person's memory regarding personal experiences, art-specific interests, personal taste, and art knowledge, e.g., prior knowledge of main characters' story (arrow 3). This means that aesthetic experience is affected by familiarity with artworks. The ability to process the style of artistic objects is a cognitive process and strongly depends on a person's knowledge and experience of art.

Cognitive processing of the piece of art relies on recognizing style and visual properties while inexperienced persons often consider only content [51]. Competence in art supports cognitive processing to reveal explicit and deep meanings, and the classification of art style provides art-specific interpretation and ambiguity understanding [82]. Art knowledge could reinforce self-related interpretation, such as perceptual associations with familiar objects and discovery of regularities.

The cognition process is strongly linked with evaluation by two feedback loops (arrows 4 and 5), as shown in Figure 1.2. The outcomes of cognitive processes are continuously evaluated. This results in the increasing level of understanding and the reduction of ambiguity. When understanding is not successful, the information processing is redirected to the previous stage. It is reflected by two feedback loops (arrows 4 and 5) in the model, as shown in Figure 1.2.

Besides, it was reported that aesthetic experience could influence the affective states of a person [19]. The initial affective states of art audiences can be continuously influenced by the outcome of affective evaluation (arrows 6, 7 and 8). Thus, the model includes the results of all previous processing stages that could change affective states. Visual and cognitive judgments with affective states result in aesthetic emotions and aesthetic judgments that are the main outputs of the aesthetic experience model. Aesthetic emotions essentially depends on aesthetic information processing, art knowledge, a subjective sphere of emotions, and preferences. They are derived from affective states and cognitive appraisal at the evaluation stage [58].

By contrast, aesthetic judgments mainly are the object-related cognitive part of aesthetic information processing. Both of them are influenced by social interactions and display places (arrow 9). For example, watching a movie together with other people in a movie theater can influence aesthetic experience of each spectator through emotional contagion. As a result, spectators' physiological and behavioural reactions can be similar. Despite the fact that it does not cover all various aspects of aesthetic experience, Leder's model precisely describes the basic processes involved in aesthetic experience [93]. There are a few modification of Leder's model in the literature, for instance, Marković's model mainly associated aesthetic experience with arousal [130]. This significantly limits a range of emotions related to aesthetic experience.

1.1.1 "Everyday" emotions

"Everyday" emotions are emotions that are mainly evoked by real life events in contrast to aesthetic emotions that are only elicited by artistic objects. Compared to aesthetic emotions, "everyday" emotions contain body poses, gestures, facial expressions, and other actions as responses to emotional situations. Nevertheless, both of them are parts of aesthetic experience according to Leder's model [121], as shown in Figure 1.2. The number of emotion definitions is very large [164] and it is difficult to study and review all of them. In this thesis we consider the most consensual definition of emotions proposed by Kleinginna and Kleinginna [108]: *"Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive."*

Discussion on emotions was started by philosophers in Ancient Greece. Since that time, many different scientists have made contributions to the various models and their assumptions. Decades of research on emotions provide three main approaches, such as the basic emotion model [65], the component process model based on the appraisal theory [7], and the psychological constructionist model [77].

1.1.2 Perceived vs. induced emotions

Perceived emotions correspond to the perception of affective content of artistic objects, e.g., perceiving emotions expressed by main characters in movies while induced emotions are evoked in audiences by this affective content. Both of them are parts of film aesthetic experience, as shown in Figure 1.2. Many studies on movie emotions naively assume that induced and perceived emotions of movie audiences are consistent and they are not distinguished from each other. Research on music emotions was the first to propose the distinction between the perceived emotions of affective content and the induced emotions of listeners [100]. This was supported by finding that induced emotions can have more intensive arousal and less intensive valence ratings with regard to perceived emotions. It was empirically shown by Gabrielsson [76], that perceived emotions from music and induced emotional responses do not have a consistent correlation. The author found that songs perceived as happy could increase the severity of depression. Other studies showed that perceived emotions are more objective than induced emotions [133], and annotations of

perceived emotions have higher agreements than annotations of induced emotions [178] because induced emotions are influenced by personal experience and preferences [151].

There has been a limited number of studies on the relationship between perceived and induced emotions of movie audiences. Hanjalic and Xu [88] assumed positive correlations between perceived and induced emotions of movie audiences to estimate spectators' affective responses. In Tan's work [184] on movie emotions, it was hypothesized that induced emotions are affected by perceived emotions by means of empathy. Besides, Baveye et al. [11] considered that intended emotions of movie directors are not always consistent with induced emotions of movie audiences. Tarvainen et al. [187] defined the main difference between perceived and induced movie affect: the former describing properties of affective movie content, the latter characterizing a spectator's emotional responses to affective movie content.

Films pass information through multiple channels to movie audiences. Spectators interpret the movie content and perceive emotions expressed by actors (perceived emotions). These induce emotions in movie audiences (induced emotions). Movie actors play emotional scenes based on the emotional description of movie scripts (expressed emotions). When movie directors write scripts, they intend to elicit specific emotions in movie audiences (intended emotions), e.g., anger, fear, and joy.

1.1.3 Aesthetic emotions

Aesthetic emotions that a part of aesthetic experience are responses to artworks like films, videos, clips, music, songs, paintings, sculptures, and jewelry, as shown in Figure 1.2. Silvia [46, 168, 169] proposed that aesthetic appraisal could cover a wide range of emotions, such as pleasure, pride, surprise, anger, disgust, contempt, shame, guilt, regret, embarrassment, confusion, etc. The author considered them as aesthetic emotions due to their link with the appraisal of an artistic narrative. Nonetheless, the author did not define the explicit criteria which allow researchers to discriminate aesthetic and non-aesthetic emotions (e.g., What is the difference between aesthetic pleasure and non-aesthetic pleasure?).

Frijda [75] described aesthetic emotions in a more flawless way, establishing two types of aesthetic emotions: complementing and responding emotions. The former are "everyday" emotions (real life emotions) that are evoked by artwork content and the latter are elicited by the form of the artwork. Responding aesthetic emotions are considered to correspond to humans' fascination of art structure and style itself. Also, Cupchik [52] proposed reactive and reflective models of aesthetic emotions based on the cognitive theories of emotions. The reactive model corresponds to pleasure and arousal elicited by artwork content while the reflective model describes emotions evoked by artistic narratives.

According to Kubovy [119], aesthetic emotions are pleasures of the mind that have no distinctive physiological and behavioural expressions unlike basic emotions. The author considered pleasures of the mind as wide collections of emotions which are spread over time, for instance, while watching a film or reading a book. The collections of emotions are transformed into other emotions as a result of narrative changes. Scherer [164] distinguished "*everyday*" emotions and aesthetic emotions: the former have adaptive functions that need the appraisal of goal relevance and coping potential and the latter are not utilitarian but are intrinsic. Aesthetic emotions are not elicited to satisfy basic needs but rather to appreciate a work of art. Also, Scherer provided the following list of aesthetic emotions: being moved/awe, wonder, admiration, bliss, ecstasy, fascination, harmony, rapture, and solemnity. According to Scherer, the non-adaptive nature of aesthetic emotions does not fully exclude actions and bodily responses, for example, moist eyes, goose pimples, and shivers. Similarly, Marković [130] defined aesthetic emotions as special feelings of unity and engagement with artistic objects, arguing that aesthetic emotions which are induced by the appraisal of art form (e.g., composition and structure) are only pleasurable. Nonetheless, the author supposed that aesthetic emotions evoked by the content of artworks may be both pleasurable and unpleasurable.

1.1.4 Emotion representation

Various representations of emotions have been developed. They are mainly obtained from basic emotion models, component process models, and psychological constructionist models. Categorizations of emotions seem to be a natural approach to emotion representations. They are originally derived from human languages in which words and expressions describe many emotional states. Discrete representations of emotions are mainly motivated by Darwin's work that included a number of universal emotions [65, 164]. Darwin argued that utilitarian emotions are essential due to their relevance to humans' survival. Ekman [65] proposed six basic emotions: anger, disgust, fear, happiness, sadness, and surprise based on his research on facial expressions while Plutchik [152] defined eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. Plutchik hypothesized that the primary emotions are biologically inspired and have evolved to reinforce the reproduction of species.

Also, Scherer proposed a list of affect categories including admiration/awe, amusement, anger, anxiety, being touched, boredom, compassion, contempt, contentment, desperation, disappointment, disgust, dissatisfaction, envy, fear, and feeling [164]. When the number of emotional classes increases, some ambiguities emerge because of language and personal differences. In addition, these categorical representations are not able to tackle the granularity

of human emotions since emotional categories cannot cover the wide range of emotions perceived and felt by movie audiences.

Continuous representations of emotions have been developed as well. Wundt [200] proposed the first dimensional representation of emotions in the 1890s. Dimensional theories of emotion propose that emotions are spanned in a 2- or 3-dimensional space. The most well-known such space is the arousal-valence(pleasure)-dominance space originally derived from cognitive theories [159]. The arousal scale varies from bored to excited while the valence scale ranges from negative (unpleasant) to positive (pleasant). In addition, the dominance scale ranges from submissive to dominant. This 3-dimensional representation was extended by attaching the predictability dimension because it was shown that three dimensions are not sufficient to completely represent the wide range of emotions for semantic analysis of some languages [74]. It is worth mentioning that there is no optimal number of dimensions for emotion representations. The number of dimensions is strictly dependent on the objectives of the model. Besides, psychophysiological studies on emotions have shown that emotions elicited by media stimulus can be sufficiently represented in the 2-dimensional arousal-valence space [63]. For example, Baveye et al. [12, 125] successfully used the 2-dimensional arousal-valence space to continuously annotate induced emotions in movie audiences over time.

1.2 Emotion elicitation and aesthetic highlights in films

Watching short videos only evoke specific emotions in movie audiences while watching films can reliably elicit different emotional reactions (film emotion elicitation) [155]. Nevertheless, some emotions like anxiety are difficult to be evoked by movie stimuli. Furthermore, individual differences and prior knowledge of the films could influence spectators' responses to films [85]. In addition, changes in physiology, behaviour, cognition, judgement, and experience are loosely connected with film emotion elicitation [134]. Also, movies are cognitively complex stimuli that could engage movie audiences with fictional reality and different unreal characters. That is why it remains unclear whether film emotions are similar to daily emotions or not. Other challenges of film emotion elicitation is standardization because films are complex and different from one another in many details, e.g., numbers of characters, colour saturation, brightness, shot length, and music in the background.

A movie highlight is a short scene that can be defined as a major or special interest to movie audiences [204]. Formally, a highlightness measure can be defined for every moment in a movie [182]. In this thesis, we define **aesthetic highlights** as scenes that are full of affective and aesthetic content. It is worth pointing out that our definition of aesthetic

highlights is associated with movie form and content rather than personal preferences and feelings. However, aesthetic highlights that we define are supposed to elicit a wide range of emotions (see Chapter 4). Personalized aesthetic highlights are beyond the scope of this thesis.

Wiley [199] hypothesized that emotions evoked by everyday life events (called "*everyday*" emotions) and emotions elicited during watching a film (called "*movie emotions*") are different. The author proposed that movie audience's attention is drawn by the film itself and the surrounding physical environment. The latter provides the atmosphere for watching a movie and separates movie experience from real life. Besides, emotions of movie audiences are clear and well defined because they are stimulated by film narratives with precise emotional descriptions. "*Movie emotions*" are elicited by movie dialogues, character and theme development that are enhanced by different movie forms (e.g., technical choices, special effects, use of the camera, lightening, and music in the background). Also, movie scripts are written according to the certain rules and emotions are transferred to viewers through multimodal channels. According to Wiley [199], humans have a thirst to frequently feel emotions. It can be justified by needs to forget about "*everyday*" problems. Movie experience is desired because it is quite intense and contains many emotions in comparison with daily life.

Marković [130] proposed that aesthetic emotions, such as admiration, delight, rapture, awe, and so on are induced by an exceptional relationship with artistic objects. In particular, aesthetic emotions are elicited by the appraisal of artistic objects' form and content. The former corresponds to the evaluation of symbolic structure and regularities, e.g., the length of movie scenes, first-person shots, and the saturation of colours. The latter assesses the content of artwork, e.g., movie story development, movie dialogues, and tensions among main characters.

An important concern for research on emotions is how humans' emotional states can be validated since we can only suppose that the particular range of emotions is evoked by specific stimuli. To assess emotional experience (evoked emotions), we can measure physiological and behavioural reactions supported by self-reports of feelings. Studies on affect show that recordings of peripheral physiological and behavioural changes, and affective reports are associated with emotional experience [84, 129]. In addition, a factor analysis was used to discover a relationship between physiological changes and emotional responses [54]. Affective ratings, brain activities, corrugator responses, heart rate, and skin conductance responses were analyzed. Two factors that respectively explained 40% and 31% of data variance were found. Electroencephalography signals, skin conductance response signals, and arousal assessments were strongly associated with the first latent variable (emotional

arousal). Corrugator, heart rate, and valence ratings were connected with the second latent variable (emotional valence).

It is important to notice that emotions can result in action tendencies and behavioural adaptation for social interactions. Emotions can influence life goals and plans and effect changes in the motor control of communication channels, e.g., different facial expressions [164]. It was shown that the feelings of happiness, sadness, anger, and anxiety led to changes in physiology, behaviour, cognition, judgement, and experience [122]. This means that information on humans' emotional states is passed through multimodal channels.

1.3 Interpersonal synchronization

Interpersonal synchrony can be considered as individuals' temporal coordination during social interactions. For example, spectators' physiological and behavioural reactions can be synchronized while watching movies together in a movie theater. A large number of terms, such as mimicry, social resonance, co-ordination, synchrony, synchronization, attunement, and chameleon effect have been used in the literature [59]. All of them describe the interdependencies between behaviours of dyads, partners, and couples. Also, turn-taking and mutual attention are strongly connected with synchrony.

Research on synchrony is related to the study of communicative interaction and language. In terms of dialog theories, a conversation is a joint activity that requires coordination at two levels: content and process [40]. The former includes that conversation partners co-ordinate what is said and attempt to reach common understanding while the latter covers prediction of beginnings and endings of conversations based on syntax, morphology, and intonation [40]. We here recall one out of several synchrony definitions that is defined as a dynamic process. Delaherche et al. [59] proposed the following definition: "*Synchrony is the dynamic and reciprocal adaptation of the temporal structure of behaviors between interactive partners. Unlike mirroring or mimicry, synchrony is dynamic in the sense that the important element is the timing, rather than the nature of the behaviors*". This co-regulation of behaviours has been investigated in terms of social, physiological, and developmental components [39, 136]. Besides synchrony among individuals' behaviours, co-regulation of physiological responses has been examined with regard to patient-therapist, infant-parent, couple therapy, and couple interaction [31, 70, 102, 162].

Also, the presence of this physiological linkage regarding the nature of interactions and individual characteristics has been extensively discussed [194]. Many studies have observed physiological synchrony between romantic couples as part of analysis of the interplay of partners' mood and emotions [71, 91, 123, 127]. A lot of importance is attached to physiological

synchrony since it allows researchers to study how individuals' behaviour can influence other individuals' health and well-being [162, 193, 194]. This suggests that we can study different social interactions by means of measures that assess levels of interpersonal synchrony. Thus, the development of physiological and behavioural synchronization measures is critically important for advancing these studies.

1.4 Motivation

Our motivation to study film aesthetic experience is twofold. From the point of view of research on art, an investigation of an inner affective state of a person exposed to an artistic object can provide insight into the understanding of humans' engagement with art: some features of artistic objects that affect aesthetic experience and human emotions can be identified to explain why people are attracted by art. Physiological and behavioural changes of individuals that are evoked by being exposed to art can be investigated.

From an application perspective, emotional and aesthetic scene detection, emotional and aesthetic scene design, and video summarization and movie recommendation systems require the prediction of affective and aesthetic content. There has already been a large number of existing applications including: personalized content recommendation [24], video indexing [207], efficient movie visualization and browsing [208], movie summarization [103], personalized soundtrack recommendation [166], and optimization of advertising content [202] that involve affective content analysis.

1.5 Research questions

In this thesis, we aim at addressing several research questions related to film aesthetic experience. In particular, we investigate definition of aesthetic highlights in movies, the consistency of induced and perceived emotions of movie audiences as well as the dependencies between "everyday" emotions and aesthetic emotions felt by movie audiences. These lead us to the following research questions, also indicated in Figure 1.2:

1. Do aesthetic highlights elicit emotions in movie audiences? (RQ1)
2. Can the level of synchronization among spectators' reactions be used to detect the different categories of aesthetic highlights? (RQ2)
 - If it is possible, which synchronization measures are the most reliable to efficiently detect aesthetic highlights?

3. Are perceived emotions of the movie content and induced emotions in movie audience always consistent? (RQ3)
4. How can we improve recognition performance of induced emotions in movie audience? (RQ4)
 - Are there other features beyond the audio-visual movie content that can contribute to induced emotion recognition?
 - Are perceived emotions discriminative for induced emotion recognition?
 - Do recognition models benefit from including temporal information and multi-modal signals?
5. Is there a direct relationship between "*everyday*" emotions and aesthetic emotions? (RQ5)
 - Is an arousal-valence space sufficient to accurately represent aesthetic emotions?
6. Are there dependencies among aesthetic highlights, aesthetic emotions, personality, and physiological and behavioural reactions? (RQ6)
 - Do aesthetic highlights elicit aesthetic emotions beyond "*everyday*" emotions?
 - Does personality influence aesthetic preferences?
 - Is it possible to predict aesthetic emotions from physiological and behavioural responses of spectators?

1.6 Contributions

The contributions and achievements of this work can be summarized, as follows:

- We show that aesthetic highlights evoke a wide range of emotions in movie audiences, studying the direct link between emotional dimensions (arousal-valence space) and aesthetic highlights (RQ1).
- We investigate several approaches to synchronization estimation to measure the amount of synchronization among multiple spectators' reactions and detect aesthetic highlights in movies (RQ2).
- We create one of the largest annotation database of aesthetic emotions and highlights which will help researchers to study movie audiences' responses to aesthetic content.

This database consists of the annotations of 30 full-length movies derived from 9 movie genres: action, adventure, animation, comedy, documentary, drama, horror, romance, and thriller (RQ1, RQ2, RQ5, RQ6).

- We carry out the first analysis of the relationship between perceived and induced emotions of movie audiences. We then use movie audiences' perceived emotions to predict their induced emotions (RQ3, RQ4).
- We are the first to quantitatively analyze the differences between "*everyday*" and aesthetic emotions of movie audiences (RQ5).
- We are the first to establish the link between aesthetic highlights and aesthetic emotions (RQ6).
- We find the effect of personality on aesthetic preferences (RQ6).

These contributions and achievements have been published in journal and conference papers [112, 113, 115, 139–142, 192]. This research was carried out in collaboration with Theodoros Kostoulas¹, Patrizia Lombardo² and Leimin Tian³.

1.7 Thesis structure

Chapter 2 presents the state of the art in affect and aesthetic content analysis as well as in highlight detection in videos and movies. The limitations of existing research are discussed. **Chapter 3** describes the current existing affective video databases with their limitations. Also, it introduces the new databases created to overcome these limitations and facilitate research on film aesthetic experience.

Chapter 4 describes a novel approach to detection of aesthetic highlights in movies based on synchronization of spectators' physiological and behavioural reactions.

Chapter 5 addresses recognition of induced emotions from multimodal signals, including movie content and movie audience reactions.

Chapter 6 describes a novel approach to aesthetic emotion recognition from spectators' physiological and behavioural signals.

Chapter 7 summarizes the contributions and accomplishments of this thesis and discusses future research directions.

¹TK is now a lecturer at the Faculty of Science and Technology at the Bournemouth University, United Kingdom.

²PL is now an emeritus professor at the Department of French at the University of Geneva, Switzerland.

³LT is now a research fellow at the Caulfield School of Information Technology at the Monash University, Australia.

Chapter 2

State of the art in aesthetics and affect recognition

Aesthetic and affective video content analysis aims at automatically recognizing a wide range of spectators' emotions. These emotions are the main outcomes of aesthetic experience (see Section 1.1) evoked by watching multimedia: movies, videos, clips, and so on. Aesthetic and affective video content analysis is different from video emotion recognition that focuses on the automatic estimation of emotions expressed by actors in video recordings. Research on aesthetic and affective video content can be categorized by modalities that are processed: video content [2, 12, 88, 94, 101, 128, 172–174, 181, 187, 197, 201, 207], and spectators' physiological and behavioural reactions [28, 29, 72, 78, 98, 110, 118, 125, 143, 156, 171, 176, 177].

In addition to aesthetic and affective video content analysis, we can distinguish highlight detection in multimedia based on its content [32, 66, 67, 80, 81, 124, 145, 146, 167, 182, 196, 204, 205] and spectators' physiological and behavioural reactions [35, 73]. The main goal of highlight detection is to determine multimedia scenes that are relevant to spectators. In particular, highlight detection from spectators' reactions tends to discover patterns in their physiological and behavioural signals that can be used to determine the occurrences of highlights in multimedia.

It is observed that there is convergence of the results of research on aesthetic and affective video content analysis as well as highlight detection that process different modalities, such as audio-visual content, spectators' physiological signals and behavioural actions (e.g., facial expressions). It is shown that fusion of multimodal signals is required to improve performance of emotion recognition [99] and highlight detection [66]. However, not all modalities contain the same amount of discriminative information on spectators' emotions [171] and movie highlights [35]. For example, movie audiences' reactions are characterized

by different dynamics in comparison to movie content features. For this reason, there is a need to study different strategies of multimodal fusion that allow different multimodal features to be incorporated at different levels of models.

Most of researchers have focused on recognizing emotions that are acted or are evoked by controlled stimuli (e.g., short videos, and movie excerpts instead of full-length movies) in laboratory conditions [35, 99]. Thus, their proposed methods and recognition models cannot be directly used to predict naturalistic emotions in real conditions. Some studies attempted to assess emotional states in naturalistic and ecological situations, for example, detection of highlights in movies projected in a movie theatre based the analysis of spectators' physiological signals [73] or emotion assessment of movie audiences from their physiological signals [125]. As a result, a performance drop of emotion recognition is observed in comparison with acted emotion recognition. This suggests that natural emotions are subtle and complex. For example, more than one emotion can be felt at the same time while watching movies. Also, some movie scenes have to first be situated in the context of the movie (story development) to be understood by movie audiences. Moreover, many studies, except [187], do not distinguish between induced emotions in movie audiences and perceived emotions from movie content, assuming that they are always consistent. This assumption could be invalid for full-length movies because movie context can cause discrepancies between what spectators feel and perceive. Furthermore, the influence of aesthetic movie features on spectators' reactions is largely overlooked. Besides, physiological and behavioural measurements are contaminated with noise and artifacts. These data often are incomplete due to sensor device failures during data collection.

A drop performance of emotion recognition is also observed when individual-dependent and individual-independent models are trained on physiological signals [4]. Thus, differences between individuals' reactions influence training and decrease the performance of models. Overall, there is a lack of recognition and detection models that are able to overcome these limitations and preserve performance independently of acted/natural emotion recognition as well as processing different individuals' reactions.

2.1 Aesthetic and affect recognition based on video content

Aesthetic experience is evoked in spectators by the form and content of multimedia, such as special effects, usage of cameras, music in the background as well as theme and character development (see Section 1.1). One of possible approaches to aesthetic and affect recognition aims to build computational models that are able to associate audio and visual patterns in video stimuli with various emotions of spectators.

Table 2.1 The summary of previous work on aesthetic and affective characterization of video content.

Authors	Stimuli	Models and modalities	Outputs	Ground truth	Performance
Hanjalic et al. [88]	movie scenes, soccer broadcasts	arousal-valence curves	continuous arousal-valence functions	none	qualitative evaluation
Soleymani et al [174]	21 full-length movies	Bayesian framework with A-V feat.	3 emotional classes	1 annot.	accuracy: 64% F1: 63%
Malandrakis et al. [128]	30 min video clips from 12 movies	2 HMMs with A-V feat.	7 categories interpolated to arousal-valence curves	7 annot.	CC for arousal: 0.54, valence: 0.23
Kang [101]	scenes from 6 movies	HMMs with V feat.	3 emotional classes	10 annot.	accuracy: 79%
Wang et al. [197]	2040 scenes from 36 movies	2 SVMs with A-V feat.	7 emotional classes	3 annot.	accuracy: 75%
Sun et al. [181]	10 movies	4 HMMs with A-V feat.	4 emotional classes	30 annot.	precision: 68% recall: 79%
Xu et al. [201]	videos from 24 movies	5 HMMs with A-V feat.	5 affective classes	unknown	accuracy: 81%
Soleymani et al. [172, 173]	64 movie scenes from 8 movies	RVM with A-V feat. or physiological signals	arousal and valence score per video	8 annot.	MSE for each participant
Zhang et al. [207]	552 music videos	2 SVRs with A-V feat., user profile	arousal and valence score per video	10 annot. 27 annot.	performance of 2 tasks
Irie et al. [94]	206 scenes from 24 movies	LTDM with A-V feat.	9 emotional classes per scene	16 annot.	agreement: 86%
Acar et al. [2]	DEAP database	2 CNNs and SVM with A-V feat.	4 classes of arousal and valence for each video	32 annot.	accuracy: 53%
Baveye et al. [12]	30 full-length movies	CNNs and SVRs with AV feat.	arousal and valence scores at frame level	10 annot.	CC for arousal: 0.33, valence: 0.30
Tarvainen et al. [187]	14 movie clips	LR and ELM with A-V feat.	arousal and valence scores, affective attributes	73 annot.	performance of 3 tasks

(A-V feat. stands for audio and video features and annot. stands for annotators).

Hanjalic et al. [88] were the first to propose affective video content analysis by continuously projecting video features into the arousal-valence space. They selected low-level audio-video features, such as motion intensity, shots lengths, loudness, and speech rate to characterize affective content that could elicit emotions in spectators. After mapping the low-level audio-video features to the arousal-valence space, they obtained arousal-valence curves. Nevertheless, they only provided a qualitative evaluation of the model. In particular, they found the smooth transitions from one arousal and valence level to another for consecutive scenes.

Table 2.1 presents a list of selected previous studies that are relevant to emotion assessment from video content features. Unfortunately, it is difficult to compare these studies because they are different from each other regarding: the type of stimuli, recognition models, outputs of the models, ground truth collection, and performance measured by various metrics. Each of the selected previous work is detailed with respect to these five criteria in Table 2.1.

Over the last years, affective content analysis has mainly focused on emotions evoked by movie scenes, movie excerpts, short videos, and music clips. Each scene was considered separately regardless of movie context. Contextual dependencies between consecutive scenes were not taken into account. Only a few studies [12, 174] analyzed affective content of full-length movies that could elicit more subtle and complex emotions over time than video excerpts. Moreover, almost all these studies assumed that felt emotions and perceived emotions of spectators were the same or highly positively correlated, except for Tarvainen's work [187] in which there was a clear distinction between them.

A wide range of machine learning models have been used to predict emotions evoked by affective content. Classifiers, such as Hidden Markov Model (HMM) [101, 128, 181, 201], Support Vector Machine (SVM) [2], Naive Bayes (NB) [174], Latent Topic Driving Model (LTDM) [94] were applied to different audio and visual features of video content for emotion classification. To predict continuous ratings of evoked emotions, e.g., in the arousal-valence space, regression models, such as Relevance Vector Machine (RVM) [172], [173], Support Vector Regression (SVR) [12, 207], Linear Regression (LP) [187], and Extreme Learning Machine (ELM) [187] were used. In addition to model selection, it is important to find audio-visual features that are always correlated with affective states of spectators. Various sets of audio and visual features were extracted to describe affective content. For example, Kang [101] extracted low-level visual features: color, motion, and shot cut rate to characterize movie scenes. Also, Convolutional Neural Networks (CNNs) were used to learn a mid-level representation of low-level audio-visual features [2]. At the level of video segments, one CNN was fed by Mel Frequency Cepstral Coefficients (MFCC) features of audio while the other CNN used color channels as low-level visual features. It is worth mentioning that

Baveye et al. [12] used deep transfer learning techniques to extract high-level features. Deep learning models that were pre-trained on images were applied to movie frames to generate abstract features of affective movie content. The transfer learning techniques could improve the performance of emotion recognition when large datasets with emotional annotations were not available.

Several emotion representations that were discussed in Section 1.2 have been used for affective content analysis. On the one hand, Kang [101] classified movie scenes into 3 emotional categories: fear, sadness, and joy, but on the other hand, Baveye et al. [12] recognized continuous arousal and valence ratings of movie audiences that were annotated at the frame level.

To evaluate a computational model that is able to associate video content with spectators' affective states, it is necessary to collect emotional ground truth, e.g., emotions that are felt by movie audiences. Generally, finding the emotional ground truth is a difficult task and requires many annotations. As shown in Table 2.1, a number of annotators varies from one study to another. The largest amount of annotators (more than 70 annotators) were involved in studies on movie affect and aesthetics [187].

Performance metrics of emotion recognition models are selected based on a representation of emotions that evoked by affective content. Generally, average values of metrics are reported when cross-validation is run. When discrete categories are assigned to emotional states, emotion recognition is considered as a classification problem. This is why, a wide range of performance measures can be used. It is important to mention that some emotions are less frequently elicited than others. Thus, distributions of emotional classes are skewed. This class imbalance should be taken into account by selecting appropriate performance metrics, such as average recall, precision, and F1 score calculated for each class [181]. When emotions are rated in a continuous dimension space, the Pearson's Correlation Coefficient (CC) and the Mean Square Error (MSE) are most common performance measures. For example, Baveye et al. [12] reported CC and MSE values to evaluate the performance of movie emotion recognition.

As we can see in Table 2.1, the results of discrete and continuous emotion recognition varies from one study to another. These differences in performance can be explained by the fact that diverse machine learning models with different audio-visual features were applied to emotion recognition. Moreover, each study had their own cross validation settings and different performance metrics. Thus, fairly comparisons of recognition models could not be made.

2.2 Aesthetics and affect recognition based on spectators' reactions

Aesthetic experience influences emotions, and physiology and behaviour of individuals. That is why another approach to aesthetic and affect recognition attempts to find computational models that are able to connect changes in physiology and behaviour of spectators with their emotional states. This can be achieved by means of several models fed by multimodal signals. A list of relevant work on emotion recognition from physiological and behavioural signals is summarized in Table 2.2.

It is difficult to make comparisons between these studies since they vary regarding several criteria: the type of stimuli, recognition models and multimodal signals, outputs of the models, ground truth collection, performance metrics, and results. Soleymani et al. [172, 173] demonstrated that individuals' physiological reactions were as discriminative as video content to predict individuals' emotions evoked by video stimuli.

A high number of participants in studies make the results more significant. For example, Kroupi et al. [118] involve 32 participants to study the emotional experience of multimedia content. The authors were interested in phase-amplitude coupling between signals generated by the peripheral nervous system and the central nervous system. In particular, electrodermal activity (EDA) and electroencephalogram (EEG) were recorded when subjects were watching music clips. The authors measured coupling between these signals by means of phase-amplitude synchronization algorithm (PAS alg.). An analysis of the results suggested that synchronization between EEG and EDA signals increased with high arousal music clips as well as with high and low valence music clips in comparison with neutral ones.

Different types of stimuli, such as images, video clips, music videos, movie excerpts, and movies can be selected to elicit emotions. For example, Chanel et al. [29] used 100 images to elicit emotions while Nicolaou et al. [143] displayed 134 video segments to evoke emotional experience. Also, Li et al. [125] projected 30 movies a movie theatre to better understand movie emotional experience where spectators could share their emotions with each other. In addition, Koelstra et al. [110] used 20 music videos to investigate music emotions and music preferences.

Table 2.2 The summary of previous work on aesthetics and affect recognition based on spectators' physiological and behavioural reactions.

Authors	Stimuli	Models and modalities	Outputs	Ground truth	Performance
Chanel et al. [29]	100 images	NB, FDA with EDA, BP, HR, Res, Temp, EEG feat.	2 or 3 arousal classes	self-assess. of 4 subj.	accuracy per subj.
Soleymani et al. [171]	20 videos	LSTM-RNN, CCRF, SVR, MLR with EEG, FE feat.	arousal and valence score per time window	5 annot.	performance of multiple models
Li et al. [125]	30 movies	WMGSRP (13 subj.)	arousal score	10 annot.	CC: 0.26 SRCC: 0.34
Soleymani et al. [177]	20 video clips	SVM with EEG, EG (24 subj.)	3 arousal and valence classes	10 annot.	accuracy for arousal: 69%, valence: 76%
Kroupi et al. [118]	40 one-min music clips	PAS alg. (EEG and EDA of 32 subj.)	3 arousal and valence classes	self-assess. of 32 subj.	statistical analysis
Ghaemmaghami et al. [78]	36 movie clips	NB with MEG feat.(30 subj.)	4 movie genres	3 annot.	mean accuracy of 36%
Koelstra et al. [110]	20 music videos	SVM with EEG, EDA, Res, Temp, ECG, BVP, EMG, EOG feat. (6 subj.)	2 arousal, 2 valence, like-dislike classes	self-assess. of 6 subj.	performance of multiple models
Nicolaou et al. [143]	134 video segments	(BLSTM-RNNs), SVR with MFCC and FSM feat. (4 subj.)	arousal and valence score	4 annot.	performance of multiple models
Soleymani et al. [176]	20 short videos	SVM with PSD of EEG (30 subj.)	3 arousal and valence classes	30 annot., 9 annot.	performance of multiple tasks
Fleureau et al. [72]	15 video clips	GP with EDA, HR, EMG feat. (10 subj.)	negative and positive valence class	self-assess. of 10 subj.	multiple metrics
Ringeval et al. [156]	9.5h of recordings	LSTM-RNN with LLD of A-V, ECG, EDA (27 subj.)	arousal and valence score	6 annot.	CCC for arousal: 0.80 and valence: 0.53

(*subj.* stands for subjects, *A-V* stands for audio-video, *feat.* stands for features, *annot.* stands for annotators, and *self-assess.* stands for self-assessment).

Various machine learning classifiers, such as NB [29, 78], Fisher Discriminant Analysis (FDA) [29], SVM [110, 171, 176, 177], and Gaussian Process (GP) [72] have been used to determine classes of emotions. For example, NB and FDA classifiers were fed by wavelet based features extracted from peripheral signals and 6 EEG frequency bands to assess induced emotions [29]. Multiple physiological signals of individuals, including EEG, EDA, blood pressure (BP), heart rate (HR), respiration (Res), and temperature (Temp) were recorded to detect classes of self-reported arousal. Emotion classification performance was measured by accuracy and it was strongly participant-dependent. This suggests that building emotion recognition models that are robust to inter/intra-participant variability of physiological and behavioural responses is a very challenging task.

To predict continuous ratings of emotions, a wide range of regression models, such as Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [156, 171], Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN) [143], Continuous Conditional Random Field (CCRF) [171], SVR [143, 171], and Multilinear Regression (MLR) [171] were applied to individuals' multimodal responses. Only CCRF, LSTM-RNN, and BLSTM-RNN models are able to include temporal information on physiological and behavioural reactions to recognize emotional states. Both of the RNN models can learn long-term dependencies between signals, e.g., a relationship between emotional reactions and patterns in EDA while CCRF models can only capture consecutive emotional states due to their sequential structure.

Soleymani et al. [171] used LSTM-RNN, CCRF, SVR and, MLR models for continuous emotion detection. These models were fed by power spectral densities (PSDs) of EEG signals and landmark point based features of facial expressions (FEs). The authors found that LSTM-RNN and CCRF models performed emotion recognition at the same level and FEs had more discriminative power for emotion recognition than EEG signals.

Ringeval et al.[156] investigated the relevance of integrating contextual information in machine learning models. To predict arousal and valence annotations of several raters from audio, video, electrocardiogram (ECG), and EDA recorded during spontaneous interactions, LSTM-RNN models were used. Low-level descriptors (LLD) were extracted from each modality with various window sizes to test different model architectures. The authors showed that LSTM-RNN models could capture the dependencies among emotional ratings, audio features, and video features. Also, the authors discovered that decision-level fusion led to higher performance than feature level fusion. This suggests that all modalities do not have the same discriminative power for emotion recognition. Nevertheless, the authors did not consider hierarchical fusion of multiple modalities since audio-video and physiological signals have different dynamics of changes and noise levels.

Nicolaou et al. [143] proposed to predict spontaneous affect from FEs, shoulder gestures (SGs), and audio cues. Emotions in the arousal-valence space were recognized by BLSTM-RNN, and SVR models fed by Mel-frequency Cepstrum Coefficients (MFCC), prosody and pitch features of audio as well facial and shoulder motion (FSM) features. BLSTM-RNN models outperformed SVR models due to their ability to take into account temporal information.

Several modalities can be measured to assess emotional states of individuals. Generally, physiological and behavioural responses of individuals can be distinguished. Furthermore, physiological signals can be split into signals generated by the peripheral nervous system, e.g., EDA and the central nervous system, e.g., EEG signals.

Soleymani et al. [177] proposed emotional tagging for video clips using EEG signals, and eye gaze data (EGD) of viewers. PSD based features were extracted from EEG signals and other features were obtained from EGD. SVM models classified videos into 3 classes of arousal and valence applying modality fusion at the feature and decision level. The multi-modal system achieved the best classification accuracy of 69% and 76% for 3 valence and arousal classes, respectively. The promising results suggest that the fusion of physiological and behavioural signals can be beneficial for emotion recognition.

Koelstra et al. [110] recognized emotions induced by music videos from electroencephalography (EEG) signals and peripheral physiological signals. SVM models were fed by PSD or common spatial pattern (CSP) features of EEG signals as well as statistical features of EDA, Res, Temp, ECG, blood volume pulse (BVP), electromyogram (EMG), and electrooculogram (EOG) to classify music videos into 2 arousal, 2 valence and like-dislike classes. The authors showed that peripheral physiological signals were more discriminative than EEG signals for music emotion recognition.

Ghaemmaghami et al. [78] proposed an approach for classification of movie clips into four movie genres, such as comedy, romantic, drama, and horror using magnetoencephalography (MEG). MEG features were extracted by means of spectral power analysis on theta, alpha, beta, and gamma frequency bands to feed NB classifiers. The authors showed that there was a correlation between brain activity in the visual and temporal lobes and audio-visual movie features. Unfortunately, the proposed system requires recording brain activity in a magnetically shielded room with controlled illumination. That is why it cannot be used in real conditions.

Soleymani et al. [176] showed how a reduced set of EEG channels could be used for multimedia implicit tagging with a slight drop of performance. SVM classifiers with a Radial Basis Function (RBF) kernel were fed by PSD based features for arousal and valence classification. Then, the authors proposed to aggregate multiple subjects' responses

to improve tagging accuracy by means of averaging features over all participants. These promising results encourage to study another aggregation approach that could take into account multiple individuals' reactions to multimedia.

Other studies attempted to assess emotions by means of peripheral physiological signals only. For example, Li et al. [125] introduced a weighted mean galvanic skin response profile (WMGSRP) of spectators to study temporal dependencies between induced arousal and peaks of EDA. This was an initial attempt to aggregate physiological responses of multiple individuals. The authors found an overall CC of 0.26 and a Spearman's Rank Correlation Coefficient (SRCC) of 0.34 between these two signals. Nevertheless, the sensitivity of the WMGSRP to the threshold selection could bias the results. Values of the threshold varied from one movie to another. Moreover, the WMGSRP is only applicable to EDA due to assumption on the existence of specific signal components.

Fleureau et al. [72] proposed a real time affect detection system from peripheral signals of individuals. The detector classified each video events into positive and negative valence classes by means of Gaussian Process (GP) models fed by statistical descriptors of EDA, HR, and EMG. The authors showed that affective states of single and multiple individuals could be detected in realistic conditions.

The studies that are detailed in Table 2.2 assigned continuous emotional ratings in the arousal-valence space or arousal and valence classes to emotions evoked in individuals. On the one hand, Ringeval et al. [156] and Nicolaou et al. [143] used continuous arousal and valence scores in their work. On the other hand, Soleymani et al. [177], [176] described elicited emotions by means of 3 classes of arousal and valence. In addition to emotional classes, Kolestra et al. [110] considered annotations of like-dislike classes to investigate music preferences.

Generally, there are two approaches for the collection of emotional ground truth: self-assessment and external annotations. The self-assessment relies on self-reporting of participants' emotions. This means that each participant directly assesses its affective state while/after performing a task. The advantage of this approach is the usage of only questionnaires while its pitfalls are that some participants are unreliable at reporting and they can attempt to hide their real emotions. For example, Fleureau et al. [72] asked participants to self-report emotions while watching video clips.

External annotation is an implicit assessment that attempts to describe individuals' affective states based on annotations of other individuals that are not involved in performing a task. For example, Li et al. [125] carried out research on movie emotional experience in which one group of 13 participants were watching movies in a movie theatre while the other group of 10 participants were annotators of emotions evoked by movie content. It is not clear which

amount of annotators is sufficient to obtain the reliable ground truth. A number of annotators should be increased when a low annotation agreement is observed. For instance, when very sublime aesthetic emotions related to art are annotated, a large number of annotators are required due to the subjective nature and complexity of these emotions.

Several performance measures can be used to evaluate an emotion recognition model. In general, average values of performance measures are reported when cross-validation is performed. Chanel et al. [29] calculated classification accuracy to evaluate NB and FDA classifiers when they identified that the distributions of emotional classes were balanced. In other cases, classifiers could be not correctly trained and all instances are assigned to the majority classes while classification accuracy remains very high. When emotional classes are imbalanced, performance metrics: F1 score, precision, and recall should be calculated per each class to measure unbiased classification performance. For example, Fleureau et al. [72] computed multiple performance metrics: accuracy, specificity, and sensitivity to evaluate valence classification using GP models. For continuous emotional scores, Ringeval et al. [156] calculated the Concordance Correlation Coefficient (CCC) instead of the MSE and CC to measure the performance of arousal and valence predictions.

As we can observe in Table 2.2, the results of discrete and continuous emotion recognition from physiological and behavioural reactions of individuals depend on: recorded modalities, extracted features, machine learning models, ground truth collection, evaluation metrics, and so on. Nevertheless, we observe that peripheral physiological signals and EEG signals can be often used interchangeably for emotion recognition. Also, it is worth mentioning that the fusion of physiological and behavioural reactions of individuals can be beneficial to emotion recognition since information on affective states are passed through multiple channels.

2.3 Highlight detection

A highlight can be considered as a short scene or a sequence of scenes that are relevant to spectators: killing scenes in action movies, spectacular views in video clips, scoring points in basketball game broadcasts, and so on. In particular, highlights can be associated with the form and content of multimedia. The former includes audio-visual features of multimedia, such as colour saturation, brightness, shot length, and music in the background. The latter corresponds to the developments of main characters, dialogues and tensions among main characters, main characters' emotions, story development, and event occurrences.

Table 2.3 The summary of previous work on video highlight detection.

Authors	Video types	Models and modalities	Outputs	Ground truth	Performance
Gong et al. [81]	10 baseball videos	MEM with A-V feat.	7 categories of baseball highlights	3 annot.	recall: 70% precision: 63%
Li et al. [124]	8 soccer videos	BN with A-V feat.	6 categories of soccer highlights	unknown	recall: 90% precision: 90%
Sigari et al. [167]	10 hours of soccer videos	FIS with A-V feat.	subjective evaluation	6 annot.	recall: 91%, precision: 95% (goal detection)
Yao et al. [205]	100 hours of first person videos	PDRM with V feat.	objective goal events Highlight curve	12 annot.	2 tasks 3 metrics
Yang et al. [204]	6500 short form videos	BLSTM-RNN with video spatial-temporal feat.	7 sport categories of search terms	6 annot.	the performance of multiple tasks
Eyben et al. [67]	4 full length movies	LSTM-RNN with PLP feat.	voicing or silence class	unknown	performance of multiple tasks
Sun et al. [182]	600 youtube videos	LLRM with dense motion feat.	6 categories	5 annot.	mean average precision: 54 %
Penet et al. [146]	15 movies	BNs with A feat.	gunshot, explosion scores	9 annot.	recall 70% (gunshots) 50% (explosions)
Gninkoun et al. [80]	15 movies	NB, LDA, QDA with A-V and text feat.	violence and non-violence class	7 annot.	multiple metrics
Chen et al. [32]	4 movies	SVM and clustering with V feat.	violence and non-violence class	unknown	avg. recall: 85 %, precision: 100 %
Fleureau et al [73]	5 movies	IAPro and MAPro with EDA signals of 128 subj.	probability of highlights	3 annot.	lack of metrics
Chênes et al. [35]	64 scenes from 8 movies	SVM with PL feat. of EMG, BVP, EDA, Temp (8 subj.)	highlight and non-highlight class	18 annot.	accuracy: 79%
Jaho et al. [98]	8 video clips	BN with motion vectors (10 subj.)	ranking of video frames	self-assess. of 10 subj.	MAP per subj.

(A-V feat. stands for audio-video features, subj. stands for subjects, and annot. stands for annotators).

In the area of highlight detection, most of the research has focused on audio and video features of videos, clips, movies, and sport broadcasts. There are some studies that attempted to use spectators' reactions to identify important moments in multimedia content. Table 2.3 summarizes a list of selected studies on highlight detection in sport broadcasts, videos, and movies. Unfortunately, a comparison of these studies is extremely difficult due to different definitions of highlights.

Various definitions of highlights were defined and studied. Sigari et al. [167] considered scoring goals as one out of sport highlights while Penet et al. [146] defined gunshots and explosions as movie highlights. Jaho et al. [98] investigated personal highlights that evoke emotional reactions. Also, a number of annotators is study-dependent, for example, Chênes et al. [35] involved 18 annotators to find movie highlights that are major interest to movie audiences.

Moreover, studies on highlight detection vary from one to another with respect to the following criteria: video types, detection models fed by multimodal features, collection and definition of ground truth, performance metrics, and results. Most of previous work on highlight detection relied on machine learning models, such as LSTM-RNN [67], BLSTM-RNN [204], Bayesian Network (BN) [124, 146], and SVM [32], Maximum Entropy Model (MEM) [81], Pairwise Deep Ranking Model (PDRM) [205], Latent Linear Ranking Model (LLRM) [182] that were fed by audio-video content features. Thus, the analysis of spectators' physiological and behavioural reactions to the content and form of multimedia was largely overlooked. Almost all studies were limited to highlight detection in short movie excerpts, videos and clips without taking into account the fact that consecutive scenes in movies are strongly dependent on each other. Furthermore, these scenes are made by movie directors on purpose to elicit specific reactions and emotions in movie audiences.

In general, highlight detection is considered as a classification problem (highlight and non-highlight classes). The results of highlight detection were reported by various performance measures: recall, precision, and accuracy. Jaho et al. [98] computed the mean average precision (MAP) per participants to measure personal highlight detection based on facial motion measurements while Penet et al. [146] calculated recall to evaluate the performance of gunshot and explosion highlight detection in movies from audio features.

2.3.1 Highlight detection from multimedia content

As shown in Table 2.3, the research on highlight detection from multimedia content can be split into two categories: detection of events in sport broadcasts interesting for the majority of sport audiences, and detection of relevant video and movie scenes, e.g., speech, sounds, and violent moments.

A wide range of sport highlight systems were developed. These systems rely on specific characteristics of these sport broadcasts, e.g., the fixed number of cameras at fixed locations around sport fields. Sport highlights have similar transitional patterns of unique views, text lags behind their corresponding events, and audio features, e.g., audience cheering.

Several approaches to highlight detection attempted to take advantage of these specific characteristics and structures of sport broadcasts. For example, Gong et al. [81] proposed the integration of multimedia features: image, audio and speech clues as well as contextual information to automatically detect and classify baseball highlights using a framework based on a maximum entropy model (MEM). The proposed framework outperformed HMMs and the BN-based approaches for detecting highlights in basketball games. This work showed how to fuse multimodal features with taking into account temporal information for short highlight detection.

Li et al. [124] used a BN based framework that was adapted to fuse multimodal clues, i.e. audio, visual, and textual information to improve the detection of highlights in soccer videos. The authors proposed to merge multimodal features with learning dependencies among them.

Sigari et al. [167] implemented a fuzzy inference system (FIS) to summarize the content of broadcast soccer videos using on-demand feature extraction. The proposed approach partitioned videos into some highlight segments and then gradually extracted low-level and mid-level features, e.g., logos, shot views, goal mouths, and scoreboards. The importance of each highlight segments was estimated using a FIS. The system performed well on soccer broadcasts and achieved a precision of 95% and a recall of 91%.

Yao et al. [205] used a pairwise deep ranking model (PDRM) to learn the relative relationship between pairs of highlight and non-highlight segments for sport first-person video summarization. In particular, the authors proposed the two-stream structure of the PDRM that is able to capture the appearance of video frames and temporal dynamics across frames for highlight detection in videos. The PDRM attached a highlight score to each segment. Higher highlight scores were assigned to highlight segments than non-highlight segments. The proposed deep model outperformed the state of the art Rank SVM around 10%.

Sun et al. [182] proposed an automatic system for ranking domain specific highlights in personal unconstrained videos by applying a novel latent linear ranking model (LLRM). The authors assume that edited videos are more likely to contain highlights than the trimmed parts of the raw videos to obtain pair-wise rankings for the training of the LLRM. The LLRM outperformed classification and motion analysis. The results showed that the automatic system could retrieve skating, surfing, skiing, gymnastics, parkour, and dog activity without additional human supervision for domains.

Nevertheless, all of the mentioned frameworks were dedicated to highlights in specific sport broadcasts. The characteristic attributes of sport broadcasts were taken into account to detect these sport highlights. It is therefore difficult to apply these frameworks to more generic videos like movies.

By contrast, Yang et al. [204] proposed a more generic approach to highlight detection in movies, using unsupervised learning of spatial-temporal features. Generic deep learning features are computationally efficient and accurate in characterizing both appearance and motion of objects in videos. In particular, the authors designed a recurrent autoencoder with shrinking exponential loss that reduces its sensitivity to noisy data. Then, the autoencoder was combined with BLSTM-RNNs to capture the temporal structure of highlight segments.

Also, research on highlight detection from multimedia content focused on retrieving video and movie highlights: voicing and silence moments, gunshots, explosions, violent scenes, and so on. Eyben et al. [67] detected voice activity in movies using LSTM-RNNs that are able to learn long range dependencies between two time series. The LSTM-RNNs were fed by perceptual linear predictive (PLP) features to indicate voicing or silence with noise. The proposed model outperformed the state of the art algorithms. The promising results suggest that this approach could be used for dialogue detection in films.

Penet et al. [146] also detected audio events in movies. The variability between the soundtracks of the movies was modelled and balanced by means of a factor analysis. The factor analysis compensation of soundtrack variability was validated by audio event detection using a BN based system with audio features. The system could detect gunshots, explosions, and screams. It can be generalized to detect other types of sound events in movies.

Violence detection can be considered as highlight detection. The goal is to detect whether or not a video consists of violent scenes. This is associated with revealing physical actions that can cause human injuries. Chen et al. [32] successfully integrated face, blood and motion information to determine whether an action scene is violent or not. Violent scene detection was decomposed into action scene detection and bloody frame detection under the specific definition of violence. Firstly, the input video was segmented into scenes. Then, features such as motion intensity, camera motion ratio, average shot length, and shot cut frequency were extracted to feed a SVM classifier. Secondly, the face, blood, and motion information were integrated to define clusters of blood color pixels. Finally, violent content was detected based on blood-color decision boundaries that were empirically selected. Even though the system reached an average precision of 100% and an average recall of 85%, it is limited to the detection of violent scene in action movies. Therefore, this approach cannot be extended to other movie genres that contain more subtle violent scenes without blood and a lot of motion.

2.3.2 Highlight detection from spectators' reactions

A few studies attempted to detect highlight in videos and movies based on spectators' physiological and behavioural reactions to multimedia content. Fleureau et al. [73] created an individual affective profile (IAPro) of a spectator and a mean affective profile (MAPro) of a movie audience using spectators' EDAs recorded during a movie projection. The peaks of the MAPro were observed while highlights identified by the audience occurred. Nevertheless, there were no quantitative evaluation of the system performance. Besides, these models cannot be applied to any other physiological signals or behavioural signals because they rely on tonic and phasic components of the EDAs.

Chênes et al. [35] used physiological linkage (PL) of spectators' physiological signals for determining highlights in movie scenes. The authors considered the PL as a physiological index of social interactions. Peripheral physiological signals of each participant, such as EMG, BVP, EDA, and Temp were recorded to compute the PL between every pairs of participants as features. Then, the vectors of PL values were averaged to feed a SVM classifier. The proposed system could detect highlights that were relevant to the majority of spectators and reached a classification accuracy of 79%. However, the spectators did not interact among themselves because they were separately watching videos without any social context. Thus, the PL could only occur, assuming the same spectators' perception and interpretation of the stimuli.

Joho et al. [98] detected personal highlights in videos based on the analysis of viewers' facial activities. Motion vectors of certain face regions were extracted using a real time facial expression recognition system. Then, a BN was fed by these vectors to detect self-reported highlights. The effectiveness of facial motion units and the whole system was measured by the MAP of video frame rankings. Although, the proposed system only required a web camera, it cannot be used to detect movie highlights based on spectators' reactions recorded in a dark movie theater. Moreover, we expect that only watching strongly emotional scenes can evoke facial activities and facial expressions.

2.4 Limitations of the existing research

Many aesthetic and affect recognition systems rely on machine learning models, such as HMMs, NBs, BNs, SVMs, SVRs, CNNs, LSTM-RNNs, and CCRFs that were trained on small datasets. Most of the models are fed by predefined handcrafted features of videos (see Tables 2.1 and 2.3). The bottleneck of these approaches is that strong domain knowledge is required because features are strongly task-dependent. In contrast, there are deep learning models that can be trained on raw signals. In particular, features are extracted in the first layers

of these models. However, the disadvantage of deep learning models is that a large number of training instances are required for training in comparison with handcrafted approaches.

Besides, aesthetics and affect recognition from viewers' physiological and behavioural reactions is even more limited with regard to fitting complex models to the data. There is a small amount of available datasets (see Table 2.2). In general, these datasets vary from one study to another in terms of stimuli, tasks, as well as physiological and behavioural signals that were collected. This significantly reduces the range of machine learning models that can be applied.

The lack of training instances limits the comparisons of models. Consequently, the variance of the results is large and their reproducibility is sometimes impossible. Furthermore, models are designed for a specific task and thus they cannot be generalized easily. This causes that these models are trained and tested on a small specific dataset with a high risk of overfitting.

Another problem is that many existing models do not take into account the fact that feeling emotions and aesthetic experiences are evoked over time and are sequential processes, as well as videos and movies are constructed in a specific manner in which consecutive scenes are strongly dependent on each other. Only LSTM-RNNs, HMMs, and CCRFs out of the mentioned models in this Chapter are capable of including this temporal information. LSTM-RNNs can learn long-term dependencies between signals. Also, HMMs and CCRFs are machine learning models with sequence of hidden states that are able to capture consecutive emotional states. In the other cases, temporal information can be only included during feature extraction.

The last but not least issue is that measurements of physiological and behavioural signals are often corrupted due to electrode contact noise and sensor device failures during data collection. Furthermore, audio-video features are contaminated with video background noise. Moreover, the quality of ground truth is questionable due to low annotation agreement and small numbers of annotators. As a result, there is a need to study and develop machine learning models that are able to deal with noisy features and a lack of reliable labels.

This thesis will attempt to overcome these limitations. Research on film aesthetics and affect will be carried out on massive datasets of full-length movies with reliable annotations. Movie stimuli will represent a wide range of movie genres. A large number of spectators will participate in experiments to obtain significant results. Recognition models will be designed to include temporal information and learn long term-dependencies among elicited emotions, spectators' physiological and behavioural reactions as well as movie content features. Film aesthetic experience will be entirely studied based on dependencies among emotions, physiology and behaviour of spectators as well as form and content of films.

Chapter 3

Affective and aesthetic corpus development

In this Chapter, we review the existing databases that allow researchers to work on various aspects of aesthetic and affective video content analysis: emotion elicitation, emotional characterizations of videos, violence and affect detection as well as spectators' moods, emotions, and personalities. In particular, we point out their limitations and drawbacks. Regarding the type of stimuli, these databases can be split into two categories: video excerpts and full-length movies. In this thesis, we are only interested in the latter.

To study aesthetic highlights in movies, induced emotions, perceived emotions, and aesthetic emotions felt by movie audiences, we decided to annotate the existing Continuous LIRIS-ACCEDE database [12, 125]. This database was created to carry out research on film emotional experiences in a movie theater (realistic conditions). The large amount of movies that come from several movie genres guarantees fairly comparable studies. Our annotations were collected to provide a gold standard for aesthetic highlight detection in movies, induced emotion recognition, and aesthetic emotion recognition to the affective computing and multimedia community. In this Chapter we describe the protocols for collecting emotional and aesthetic annotations of full-length movies. Also, we provide the statistical analysis of annotations that we collected. In Chapters 4, 5, and 6 we use these annotations to evaluate the performance of the models that we propose.

3.1 Existing aesthetic and affective multimedia databases

Creating an affective and aesthetic database is required since many existing databases only contain movie excerpts instead of full-length movies. Also, the existing databases do not

contain annotations that characterize film aesthetic experience. Most of available databases include a small set of movie excerpts for emotion elicitation in laboratory conditions. The selected movie excerpts are supposed to evoke strong emotional reactions in individuals. Movie excerpts can only represent a small part of the whole movie content. Complex and subtle emotions that could be potentially evoked during a full-length movie projection are omitted.

Table 3.1 summarizes a list of selected databases that allow researchers to carry out aesthetic and affective video content analysis. These databases were developed with different goals, which is why it is difficult to compare them. They differ on several criteria: the type of stimuli, recorded modalities, the number of subjects participating in experiments, the collection of categorical and dimensional ground truth as well as initial research goals. Most of the existing databases contain various short videos: music videos [1, 109], different video and movie clips [10, 13, 26, 64, 137, 163, 170, 177, 180, 187, 188], and movie scenes [185, 186] as stimuli, except for a few databases [12, 62, 125] that consist of full-length movies. Regarding recorded modalities, these selected databases can be divided into two groups: the databases with audio and video content of stimuli and the databases with physiological and behavioural reactions of individuals.

Schaefer et al. [163] released the FilmStim database that consists of 70 movie excerpts that are supposed to evoke emotional states in experimental psychology experiments. For each emotion, 10 most frequently mentioned scenes were selected. Then, 70 movie excerpts were rated by 364 participants with regard to multiple emotional dimensions. Eventually, ranking scores were computed for 24 classification criteria, e.g., subjective arousal, positive and negative affect, and positive and negative affect scores derived from the differential emotional scales. Also, discrete scores for anger, disgust, sadness, fear, amusement and tenderness as well as 15 mixed feelings were annotated. It is important to mention that all the movie excerpts were labelled at a global level. In particular, one label was assigned to each movie excerpt. This is insufficient for the characterization of the dynamic emotion elicitation while watching movies.

Table 3.1 The summary of existing affective and aesthetic databases.

Database	Stimuli	Modality	Size	Ground truth	Description
HUMAINE [64]	50 clips (5-180 sec)	A-V record., ECG, EDA, Res, Temp	unknown	categorical and dimensional	everyday action and interaction
FilmStim [163]	70 movie excerpts (1-7 min)	movie A-V	364 annot.	categorical and dimensional	emotional stimuli
DEAP [109]	120 music videos (1 min)	EEG, EDA, Res, Temp, ECG, BVP, EMG, EOG, FE	32 subj. with self-assess	multiple dimensions	music video stimuli
MAHNOB -HCI [175]	20 short clips (35-117 sec)	audio, FE, EG, EEG, ECG, EDA, Res, Temp	30 subj. with self-assess.	categorical and dimensional	video, movie excerpt stimuli
EMDB [26]	52 non-auditory movie clips (40 sec)	movie video, EDA, HR	113 annot. 32 subj.	arousal, valence and dominance	emotional stimuli
VSDS [62]	15 full -length movies	video, sound, subtitles	6 annot.	physical valence, high-level concepts	violence movie scenes
D. LIRIS- ACCEDE [10, 13]	9800 video clips (8-12 sec)	movie A-V	1517 annot.	arousal and valence	affective tagging
C. LIRIS- ACCEDE [12, 125]	30 full -length movies	movie A-V, EDA, ACC	5 annot. 13 subj.	arousal and valence	affective tagging
MSAADS [188, 187]	14 movie clips (1-2.5 min)	movie A-V	73 annot.	83 stylistic, aesthetic, affective attributes	affective analysis
FMDS [185, 186]	50 movie scenes (0.5-3 min)	movie A-V	42 annot.	movie style and mood	multimedia analysis
MediaEval 2015 [170]	10900 short clips (8-12 sec)	movie A-V	1517 annot. and 17 annot.	2 violence, 3 arousal, 3 valence classes	violence and affect detection
DECAF [1]	40 music videos (1 min)	FE, MEG, ECG, EOG and EMG	30 subj. with self -assess., 7 experts	categorical and dimensional	multimedia analysis
ASCERTAIN [180]	36 affective movie clips (51-128 sec)	EEG, ECG, EDA and FE video record.	58 subj. with self-assess.	affective ratings	emotion, personality
AMIGOS [137]	16 short, 4 long videos	EEG, ECG, EDA	40 subj. 3 annot.	personality scales affect, mood, personality	affect, mood, personality

(A-V stands for audio-video, subj. stands for subjects, annot. stands for annotators, self-assess. stands for self-assessment, and record. stands for recording).

Baveye et al. [10, 13] created the Discrete LIRIS-ACCEDE (D. LIRIS-ACCEDE) database. This database is one of the most significant emotional databases with respect to numbers of movies and annotators. It is composed of 9800 video clips extracted from 160 movies. The ratings of each excerpt regarding arousal and valence were done by 1517 annotators. Thus, this dataset can be considered as a benchmark for affective tagging of movie scenes. Nevertheless, the authors did not take into account the fact that some scenes extracted from one movie could be dependent on each other. The consecutive scenes might share similar information on the form and content of the movie. This can result in overestimating the performance of affective tagging systems.

Tarvainen et al. [187, 188] introduced a Movie Style, Aesthetics, and Affect Data Set (MSAADS) to carry out affective movie content analysis including movie style and aesthetics. There had been a lack of data on perceptual stylistic and aesthetic attributes of film before this data set came out. The authors emphasized that there could be a difference between perceived and felt affect in their preliminary research. Nevertheless, the analysis can only be carried out on movie scenes instead of whole movies. The disadvantage of this data set may be that consecutive scenes are separately considered although many aesthetic attributes in terms of form and content are intentionally shown over time by moviemakers.

Also, Tarvainen et al. [185, 186] created a Film Mood Data Set (FMDS) to investigate film mood of different scenes. The film scenes were distinguished between each other based on location, time of the day, dialogues, and music. The dataset allows researchers to analyze dependencies between film mood, perceptual stylistic, and audio-video features for various scene type. Nevertheless, there is only some information on selected scenes without taking into account a movie scenario context. Therefore, studies on aesthetics and affect of full-length movies cannot be carried out.

Sjöberg et al. [170] proposed the Affective Impact of Movies Task that is a part of the MediaEval 2015 Benchmarking Initiative. Actually, this database is an extension of the D. LIRIS-ACCEDE database. The database was created to address induced affect and violence detection in short videos. Thus, each short video could be classified into violence and non-violence scenes, and 3 classes of arousal (calm, neutral, and active) and valence (negative, neutral, and positive). The overall goal of the initiative is to conduct studies that can lead to the design of an automatic video search system that is able to fit users' particular mood, age, and preferences.

Demarty et al. [62] built the Violent Scene Data Set (VSIDS) by annotating 15 full-length movies in terms of physical violence scenes with 10 high level concepts. The goal of the database was to develop a violent scene detection system based on multimodal information, such as video, sound, and subtitles tracks. Also, the detection of violent scenes in movies is

strongly associated with the analysis of induced emotions since we expect that violent scenes evoke strong negative emotions.

The other group of the selected databases provides physiological and behavioural signals of individuals participating in experiments. Douglas-Cowie et al. [64] built the HUMAINE database that is composed of three naturalistic and six induced reaction databases. The number of participants in each of these nine databases varies from 8 to 125 and these nine databases contain different modalities, e.g., peripheral physiological signals and audio-video recordings of individuals. The main goal of the HUMAINE project was to provide the affective computing community a wide range of data containing several emotional labels at a global and frame level. The former includes labels of emotion-related states, mixed emotions, context, key events, emotion words, and appraisal categories while the latter consists of labels of emotion intensity, acts, arousal, valence, power, and so on. The authors collected individuals' physiological signals, such as ECG, EDA, Res, and Temp. It is worth mentioning that behavioural cues of individuals can be extracted from audio and video recordings that are incorporated in this database.

Koelstra et al. [109] created the DEAP database for the multimodal analysis of human affective states. EEG and peripheral physiological signals: EDA, Res, Temp, ECG, BVP, EMG, and EOG of 32 participants were collected during watching a subset of 40 music videos obtained from 120 one-minute long music video excerpts. Moreover, frontal face videos were also recorded for 22 out of the 32 participants. Also, these 40 one-minute long excerpts of music videos were rated by 14 annotators regarding the levels of arousal, valence, like/dislike, dominance, and familiarity.

Soleymani et al. [175] created the MAHNOB-HCI database for emotion recognition and implicit tagging. This database includes multimodal recordings of emotional responses to 20 short excerpts extracted from movies and videos. The authors collected 30 participants' audio, FE, EG, EEG, ECG, EDA, Res, and Temp during watching these 20 excerpts. Then these participants self-reported their felt emotions by means of arousal, valence, dominance, and predictability dimensions as well as emotional categories: disgust, amusement, joy, fear, and sadness.

Abadi et al. [1] introduced the DECAF database with physiological responses to affective multimedia content. Compared to the DEAP and MAHNOB-HCI database, this database includes brain activity recorded by a MEG sensor. This only requires little physical contact with the subject's scalp; it therefore facilitates naturalistic affective responses. Also, the DECAF contains synchronously recorded FE, ECG, EOG, and EMG. This database allows researcher to compare video and music stimuli as well as the peripheral physiological signals and MEG signals for emotion elicitation and recognition, respectively. In addition, the

analysis of dependencies between subjects' self-assessments of affective states (arousal, valence, and dominance) and their physiological reactions to music and video clips could be carried out. Also, this database comprises continuous arousal and valence annotations over time done by seven experts for dynamic emotion prediction.

Carvalho et al. [26] released the EMBD database that contains 52 non-auditory movie clips. The stimuli were selected to elicit a wide range of emotions that were annotated by 113 annotators with regard to valence, arousal, and dominance dimensions. Also, the EDA and the HR of 32 subjects were collected while watching these movie clips.

Baveye et al. [12, 125] who created the D. LIRIS-ACCEDE database also released the Continuous LIRIS-ACCEDE (C. LIRIS-ACCEDE) database that consists of continuous arousal and valence annotations of felt emotions while watching 30 full-length movies. These 30 movies varies from each other regarding their genre, content, duration, and language to represent a wide range of movies. This database also provides physiological and behavioural signals of spectators watching these movies in a cinema theater. The main contribution of this database is the creation of a benchmark for affective analysis of full-length movies at the scene level based on both audio and video as well as spectators' EDA and movement acceleration (ACC) signals.

Subramanian et al. [180] built the ASCERTAIN database for emotion and personality recognition. This database contain physiological signals recorded by means of commercial physiological sensors. The ASCERTAIN database is the first database that covers personality traits, emotional states, and physiological responses. It consists of big-five personality scales, self-assessment of arousal, valence, engagement, liking, and familiarity. Also, EEG, ECG, EDA, and FE were recorded during watching affective movie clips. The great advantage of this database is that it allows researchers to study the relationships between viewers' affective state ratings, personality scales, and physiological responses.

Miranda-Correa et al. [137] proposed the AMIGOS database for affect, personality, and mood research on individuals and groups. The authors intended to evoke emotions using short and long videos in two different social contexts: individual settings and group settings. The former corresponded to watching videos separately while the latter corresponded to watching videos together with other subjects. This database allows researchers to study multimodal affective responses based on physiological responses of subjects with respect to their personality and mood, social context, and video duration. The data was collected in two experiments. In the first experiment, 40 subjects were watching 16 short emotional videos. In the second experiment, the subjects were watching 4 long videos, some of them separately and the others in groups. Measurements of EEG, ECG, and EDA were collected using wearable sensors. In addition, frontal HD video and both RGB and depth full body

videos were recorded. Then, the subjects were asked to self-assess their felt emotions while watching videos regarding: valence, arousal, control, familiarity, liking, and basic emotions. Also, the external-assessment of valence and arousal level were collected from 3 annotators.

Much research focused on the analysis of specific emotions evoked by movie excerpts. Most of annotations are at a global level (one emotional annotation per movie excerpt) regardless of dynamic emotion changes from one scenes to another. Researchers investigated aesthetic attributes at the global level instead of analyzing the temporal context and time dependencies in full-length movies. Furthermore, the influence of movie content and form on emotions is neglected. That is why there is a need to study together aesthetics and affect of full-length movies including dependencies between consecutive scenes. To do so, it is required to create an aesthetics and affect database.

3.2 Stimulus selection

The C. LIRIS-ACCEDE database was selected to be annotated with respect to aesthetic highlights in movies, perceived emotions, and aesthetic emotions felt by movies audiences [12, 125], because no other database has the following characteristics:

- The C. LIRIS-ACCEDE database contains 30 full-length movies with a large amount of emotional and aesthetic scenes that can influence spectators' affective states.
- These movies represent several movie genres: action, adventure, animation, comedy, documentary, drama, horror, romance, and thriller that can elicit various aesthetic experiences in spectators.
- The physiological and behavioural reactions of 13 spectators watching 30 movies (the total duration of the movies is 7 hours, 22 minutes, and 5 seconds) in a darkened air-conditioned amphitheater were collected. The Bodymedia armband sensors attached to spectators' hands measured EDA and movement ACC signals.
- Continuous annotations of induced emotions in the arousal-valence space were collected from another 10 participants. During annotation collection, these movies were grouped into four sets according to their duration. Each of 10 participants watched selected movies from two sets once and then annotated continuous arousal and valence scores (value range [-1,1]) of the emotions they felt during watching (induced emotions). Then, the means of scores provided by the participants over each second of the movie were used as the gold-standard.

3.3 Aesthetic highlight annotations

Section 3.3 is based on our joint work with Patrizia Lombardo [141]. Our experiment is an extension of the work [113, 112, 139, 140] that investigated aesthetic highlights in the movie *Taxi driver*. The proposed structure of aesthetic highlights is chosen based on various film theories and experts' feedback on the annotation process [14, 20, 27, 57, 60, 61], as is shown in Figure 3.1. We can distinguish two general categories of aesthetic highlights: Form and

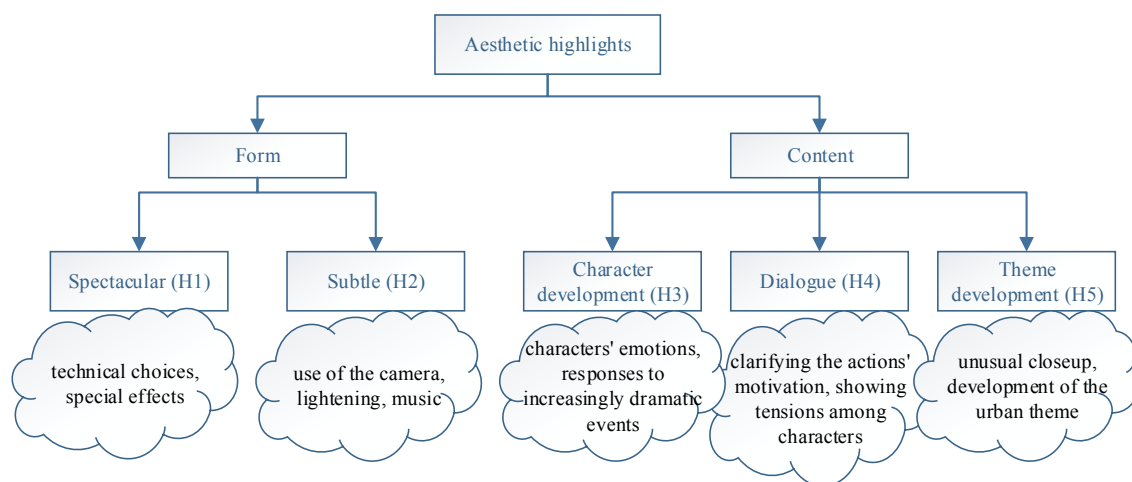


Fig. 3.1 The 5 categories of aesthetic highlights in movies [141].

Content. Form (highlights H1, H2) corresponds to manners in which subjects are presented in films, e.g., adding special effects and playing music in the background. Content (highlights H3, H4, and H5) covers the subjects presented in the films, such as developments of main characters' emotions, dialogues that cause actions and tensions among main characters as well as a specific theme development in a movie, e.g., occurrence of events or conversations that result in mental or emotional strains of characters.

As shown in Figure 3.2, aesthetic highlights in the movies from the C. LIRIS-ACCEDE database were annotated by an expert with technical support of one person using an open-source annotation software [107], similarly to the previous work [113]. The annotations are the objective assessment of the movies including 5 categories of aesthetic highlights, as presented in Figure 3.1. We selected movie scenes with high levels of aesthetic values and emotions. These scenes are constructed by moviemakers in a manner to establish engagement between spectators and movies. The structure of these scenes is designed to increase the enjoyment of watching the whole movie because it provides the context of the full story to movie audiences. A strong aesthetic experience can evoke specific affective states in spectators.

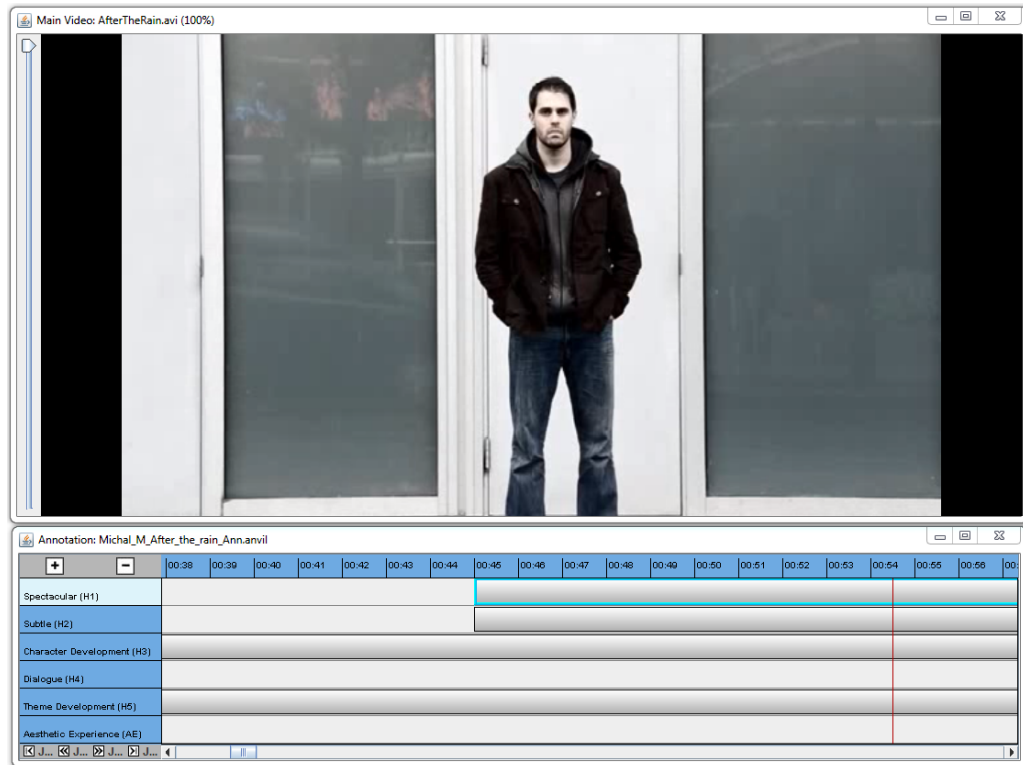


Fig. 3.2 A snapshot of aesthetic highlight annotations in the movie *After the Rain*.

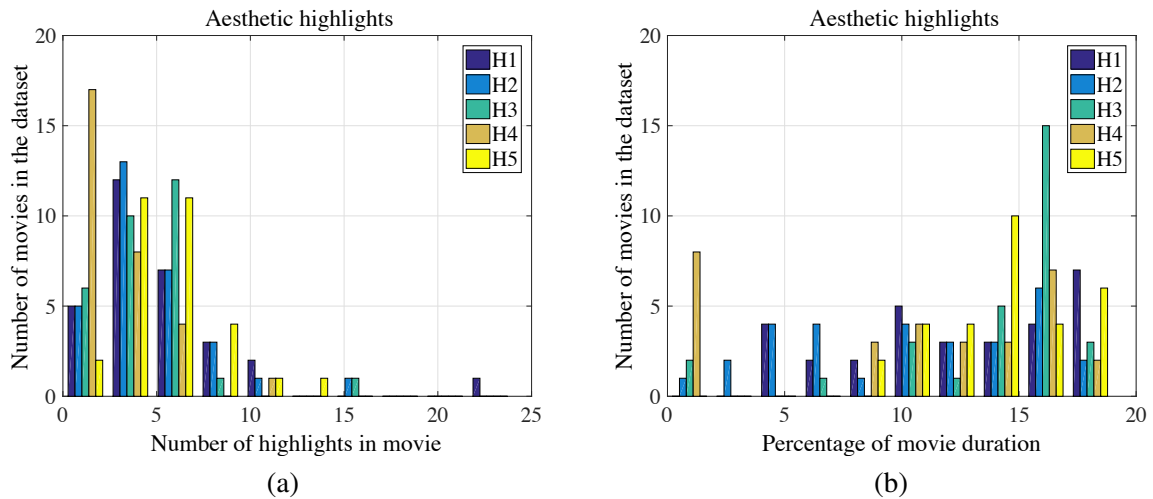


Fig. 3.3 Statistical analysis of aesthetic highlights annotated in the C. LIRIS-ACCEDE database, the distribution of : (a) the numbers of the particular highlight category per movie, (b) the duration of the particular highlight category per movie.

Figure 3.3 plots the distributions of the number of the aesthetic highlights per movies and the percentage of movie duration for the whole C. LIRIS-ACCEDE database. We observe

that there are no more than 25 highlights of a given type in a movie. The duration of these highlights is not longer than 20% of a movie duration. That means that only particular scenes are considered as aesthetic highlights.

In addition, we used Cohen's kappa coefficient κ to measure some overlaps between pairs of aesthetic highlight categories, as shown in Figure 3.4 [41]. To take into account the fact that one movie varies from another regarding the amount of aesthetic highlights and their duration, we applied a fixed-effect model to κ values to find the overall κ value [21]. Thus, we computed the weighted average of κ over all 30 movies. We then interpreted the practical significance of these overlaps between different aesthetic highlight categories, assuming that an overall value of κ at around 0.4, 0.6, and 0.8 correspond to the weak, moderate, and strong overlap, respectively [135].

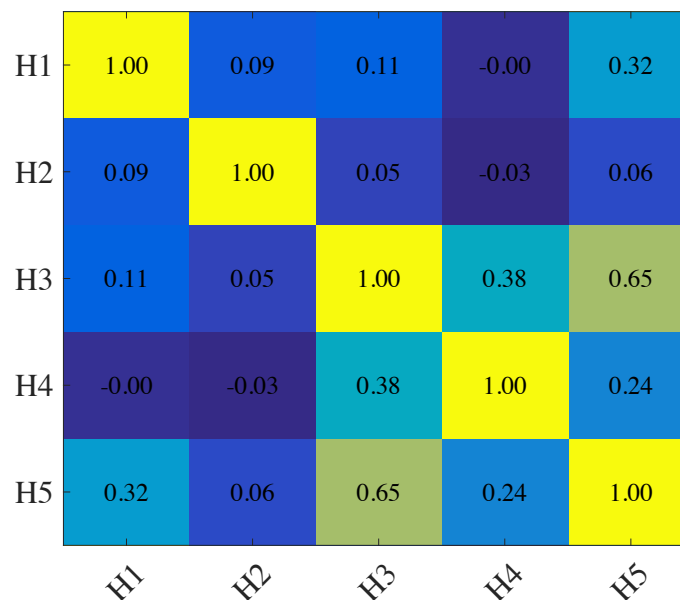


Fig. 3.4 Overall Cohen's kappa coefficient between different categories of aesthetic highlights in movies: strong overlaps (yellow) and weak overlaps (purple).

As a result, we observe that there is a moderate overlap between aesthetic highlights H3 and H5. This can be justified by the fact that main characters' emotions and their responses to dramatic events are often a part of movie theme development. Furthermore, it is worth mentioning that aesthetic highlights H3 and H4 weakly cover each other. It means that a certain amount of emotions are expressed during dialogues between main characters. This is investigated more in Chapter 5 in which we use emotional annotations of dialogues to predict emotions induced in movie audiences. We should also consider the dependence between highlights H1 and H5 because they tie movie form and content regardless of a non-significant

overlap between these two highlight categories (an overall κ of 0.32). This describes how the important events of a movie story are presented.

3.4 Perceived emotion annotations

Section 3.4 is based on our joint work with Leimin Tian [142, 192]. We here describe how the extended annotations of the C. LIRIS-ACCEDE database were collected. Also, their detailed statistics are provided. These include transcripts of movie dialogues and affective cue labels in Section 3.4.1, perceived emotion annotations in Section 3.4.3 as well as the analysis of agreement between perceived and induced emotion annotations in Section 3.4.4.

Table 3.2 Detailed statistics of 8 movies selected from the C. LIRIS-ACCEDE movies (from Leimin Tian [142, 192]).

Movie	Genre	Utterance count	Mean sent. duration (sec)
After the Rain	Drama	77	3.000
First Bite	Romance	54	2.056
Nuclear Family	Comedy	147	2.694
Payload	Adventure	121	2.488
Spaceman	Adventure	133	2.489
Superhero	Drama	161	2.832
Tears of Steel	Adventure	79	2.165
The Secret Number	Drama	98	2.724

We chose 8 English movies out of 30 movies from the C. LIRIS-ACCEDE database, as presented in Table 3.2. These movies, such as *Spaceman*, *Nuclear Family*, and *The Secret Number* contain significantly more dialogues than the other movies from this database [12, 13]. Moreover, these movies come from different movie genres and are in the double-reality art form. Double reality is an abstract concept in which the main characters exist between two different worlds. This might be also referred to spectators' association with movies when their real world and the imaginary movie world are mixed. Thus, the movie audiences could become more engaged with the main characters. This is particularly interesting for understanding perceived and induced emotions of movie audiences due to the strong engagement with movies. These 8 movies last 118 minutes, and contains of 870 utterances.

3.4.1 Transcription and affective cue annotation

The movie transcription and affective cue annotations were done by two expert annotators. To make the annotation process efficient, audio recordings of the movies were firstly processed by the IBM Watson Speech to Text service¹. The service provides automatic speech transcription with word timings. The auto-generated transcripts were then manually corrected and annotated by two annotators working simultaneously. Each of them annotated 5 movies out of 8 selected movies.

Table 3.3 Movie transcript annotation agreement: the Pearson's Correlation Coefficient (CC) and the Concordance Correlation Coefficient (CCC) between start and end timings of utterances and words (from Leimin Tian [142, 192]).

Labels	Start (CC)	End (CC)	Start (CCC)	End (CCC)
Utterance	0.998	0.998	0.997	0.998
Word	0.999	0.999	0.999	0.999

To evaluate the annotation agreement, including word timing alignment, the movies *First Bite* and *Spaceman* were annotated by two annotators. Then, the Normalized Damerau-Levenshtein (NDL) distances of the transcripts, as well as the CC and the CCC of the word timings were calculated to measure the alignment of two sets of annotations [9]. The NDL distance measures the distance between two strings. It is defined as the minimum number of string operations that is needed to transform one string to the other normalized by the length of the longer string out of the pair. The NDL distance of 0 indicates that the two strings are identical. Thus values around 0 corresponds the high annotation agreement. We found that 94.8% of the word transcriptions are identical for both annotators and the average NDL distance of word transcriptions is 0.049. In addition, the CC and CCC for the word and utterance timings of the transcript are presented in Table 3.3. We observe that the utterance and word timings annotated by the two annotators are strongly correlated. This suggests that these two annotators strongly agreed on movie transcriptions.

3.4.2 Annotations of disfluency and non-verbal vocalisation in movie dialogues

Also, the same two annotators annotated the following categories of DISfluency and Non-verbal Vocalisation (DIS-NV) in movie dialogues: filled pauses (e.g., "eh" or "hmm"), fillers (verbal filled pauses), stutters, laughter, and audible breath (remaining words are labelled as

¹<https://www.ibm.com/watson/developercloud/speech-to-text.html>, retrived on 2016.12.20

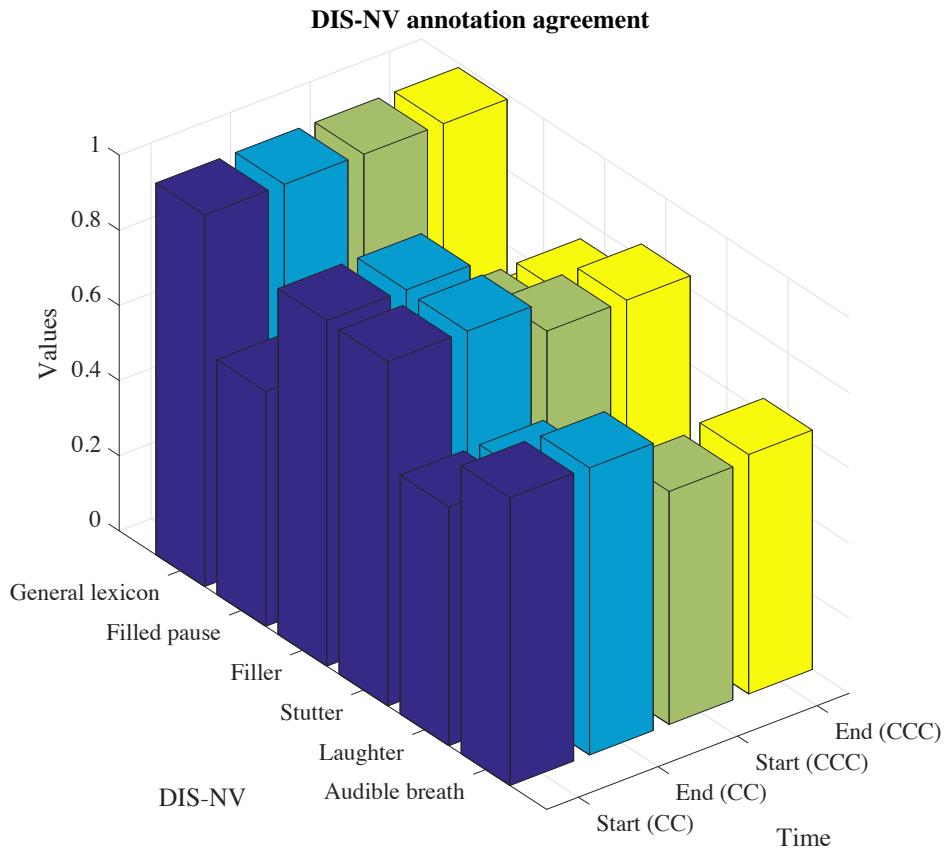


Fig. 3.5 The Pearson's Correlation Coefficient (CC) and the Concordance Correlation Coefficient (CCC) between start and end timings of DISfluency and Non-verbal Vocalisation (DIS-NV) annotations (from Leimin Tian [142, 192]).

general lexicons). It has been shown that the DIS-NVs could indicate speaker emotions in spontaneous dialogues [191]. To evaluate annotation agreement, we split the annotations into six subsets regarding the DIS-NV categories. We then calculated the CC and CCC of start and end timings of the words for each subset. Despite of the fact the annotation agreement on DIS-NV labels is lower compared to the movie transcription agreement, the annotations are strongly correlated, as presented in Figure 3.5. This means that annotating DIS-NV labels, e.g., laughter and audible breath is a subjective task because there could be some ambiguity in interpretation made by environmental noise and playing music in the background.

Table 3.4 reports the statistics of DIS-NV categories in each movie. Generally speaking, there are more disfluencies than non-verbal vocalisations in these movies. Also, filler is the most common category of the DIS-NV. As we can see, romance movie *First Bite* contains the least DIS-NVs of all movies, while drama *Superhero*, comedy *Nuclear Family* and

Table 3.4 The amount of DISfluency and Non-verbal Vocalisation (DIS-NV) annotations in movie dialogues (from Leimin Tian [142, 192]).

Movie	General lexicon	Filled pause	Filler	Stutter	Laughter	Audible breath
After the Rain	532	0	9	0	0	0
First Bite	185	6	2	0	2	0
Nuclear Family	748	18	0	18	0	5
Payload	712	1	15	2	1	3
Spaceman	686	4	5	13	5	5
Superhero	910	9	16	0	2	8
Tears of Steel	273	0	6	18	0	7
The Secret Number	549	1	12	0	0	3

adventure movie *Payload*, *Spaceman*, and *Tears of Steel* contain the most DIS-NVs. It is worth pointing out that amounts of DIS-NVs are indicators of speaker uncertainty [126] and a level of conflict [195]. Our observation is that dramas, comedies, and adventure movies have more DIS-NVs than the other movie genres. This may indicate that there is a high level of uncertainty in these movie dialogues and thus the related story development, as well.

3.4.3 Annotating perceived movie emotions

The emotion annotation process is supposed to be more subjective in comparison to movie transcription. Thus, many annotators are required to do this task. Previous work on emotions has recommended having more than 6 annotators to achieve reliable emotion annotations [23]. Recent developments in the area of crowdsourcing allows us to have easy access to larger numbers of annotators. To collect a large amount of annotations in a time and cost efficient manner, perceived emotions of movie audience were annotated by means of Amazon Mechanical Turk². It is a crowd-sourced annotation platform provided by Amazon to perform Human Intelligence Tasks (HITs).

Firstly, we segmented these 8 movies into utterance excerpts based on manual transcription of utterance timings. Then, we collected at least 10 annotations from different annotators for each movie clip. In our studies, we assume that annotators can correctly understand and perceive affective movie content. As shown in Figure 3.6, the annotators were instructed to annotate the emotions expressed by main movie characters along arousal, power, and valence dimensions with 1 to 9 integer scores. The explanation of each emotion dimension and meaning of the different scores were provided to the annotators. Each HIT consisted of

²<https://requester.mturk.com/>


Please rate the emotions expressed by the movie characters in the following videos.

Please rate the intensity of each **emotion that the movie characters show** on a 9 point scale: (1) = "lowest", (5) = "medium", (9) = "highest".

Arousal: Low Arousal: bored or inactive; Medium Arousal: calm or neutral; High Arousal: excited or activated.

Valence: Low Valence: negative (unhappy, sad, angry); Medium Valence: neutral; High Valence: positive (elated, happy).

Power: Low Power: being dominated and submissive; Medium Power: neutral; High Power: dominating and in control.



Play video




























Bored										Excited
Negative										Positive
Dominated										Dominating

Fig. 3.6 A snapshot of perceived emotion annotations in the movie *Spaceman* on Amazon Mechanical Turk.

5 continuous utterance excerpts from the same movie, displaying in their original order to provide movie context. Each movie utterance was in different HITs to reduce annotation bias. Although the HITs were posted in random order, we tracked all the previous annotators of each movie to avoid that an utterance could be annotated by the same annotator more than one time. Annotators could only annotate a clip when it finished playing. Also, annotators were only able to submit their work after annotating all movie excerpts. To sum up, we published 1809 HITs. As a results, we collected annotations from 129 annotators with various cultural and educational backgrounds.

The scores of the crowd-sourced annotations were normalized to $[-1,1]$ to be consistent with the induced emotion annotation range from the C. LIRIS-ACCEDE database. We then calculated the means of these annotations collected at the level of each utterance of the movie dialogues. In particular, this provided us utterance-level arousal, power, and valence annotations of perceived emotions of movie audiences. Table 3.5 presents statistics of the

Table 3.5 Mean level of movie audience’s perceived emotions per movie [142].

Movie	Arousal	Valence	Power
After the Rain	<i>-0.149±0.142</i>	<i>-0.238±0.155</i>	<i>-0.118±0.138</i>
First Bite	0.003±0.212	-0.043±0.204	<i>0.055±0.175</i>
Nuclear Family	<i>0.106±0.301</i>	-0.037±0.385	<i>0.117±0.251</i>
Payload	<i>0.073±0.213</i>	-0.045±0.257	<i>0.121±0.210</i>
Spaceman	<i>0.127±0.198</i>	<i>0.115±0.265</i>	<i>0.122±0.148</i>
Superhero	<i>0.127±0.212</i>	0.032±0.254	<i>0.088±0.229</i>
Tears of Steel	<i>0.238±0.232</i>	-0.063±0.228	<i>0.202±0.183</i>
The Secret Number	<i>0.067±0.199</i>	-0.054±0.160	<i>0.127±0.137</i>

Numbers in italics indicate means that are significantly different from 0 (p -value < 0.05) while numbers in bold italics indicate means that are significantly different from 0 and the largest for all the movies regarding a given emotional dimension.

perceived emotion annotations for each movie. Even though the averages of perceived emotion dimensions vary from one movie to another, the variances are in the same order of magnitude. Our observations are supported by the one sample t -test at the significance level of 0.05. We used the t -test because we can assume that the distributions of average annotations are Gaussian due to the Central Limit Theorem. As we can see, there are some movies that are close to the neutral state (value 0) with regard to average perceived emotions, such as the movie *First Bite*. This means that the movies contain a balanced number of scenes with various emotional tones.

Besides, adventure movies have higher arousal, valence, and power than the other movie genres. This means that movie events with specific emotional tones dominate the content of this movie genre. Moreover, the observation is that the romance movie is the closest to the neutral state in terms of arousal, as shown in Table 3.5. This suggests that there is a balance between the amount of exciting and relaxing scenes in these movies. Furthermore, the comedy includes movie scenes with the highest emotional discrepancies between one another.

3.4.4 Agreement on perceived and induced emotion annotation

In this section we investigated differences between induced and perceived emotion annotations. Figure 3.7 shows the distributions of the average variances of the annotations on each movie. Please note that "Per-A", "Per-V", and "Per-P" correspond to perceived arousal, valence, and power, respectively. Consequently, "Ind-A" and "Ind-V" represent induced arousal and valence, respectively. We used the original annotations per second from the C. LIRIS-ACCEDE database for induced emotions while we processed the perceived emotion annotations at the level of utterances. Then, we calculated the variance over all annotators at each emotion annotation step (a second or an utterance of a movie, respectively) for a given movie.

As shown in Figure 3.7, we can observe that the average variance for perceived emotions is larger than that for induced emotions for all movies. The observations are supported by

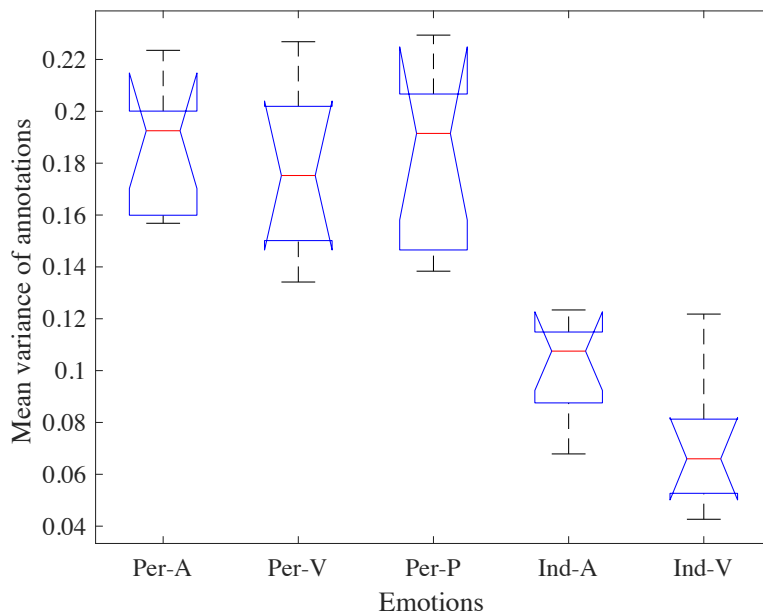


Fig. 3.7 Mean variance of perceived and induced emotion annotations.

an one-way ANalysis of VAriance (ANOVA). We used the the one-way ANOVA because we are allowed to assume that the distributions of average variance are Gaussian due to the Central Limit Theorem. It shows that all means of mean variance distributions are not equal ($p\text{-value} \ll 0.0001$). This could be the result of the crowd-sourced annotation collection for perceived emotions of movie audiences. The perceived emotion annotations were collected from 129 untrained annotators with various cultural and educational backgrounds, while the induced emotion annotations from the C. LIRIS-ACCEDE were provided by only 10 trained

annotators who are postgraduate students living in France. Moreover, we can suppose that these annotations are more coherent than the crowd-sourced annotations because these 10 trained annotators have less various cultural and educational backgrounds as well as similar age. This could influence the emotion induction process and bias the annotations.

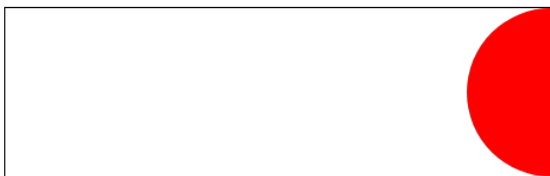
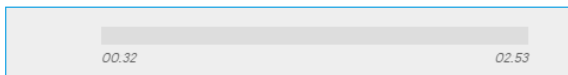
3.5 Aesthetic emotion annotations

Annotating

Wonder

Wonder is a reaction to unknown, surprising, perhaps fantastical phenomena or events. For example, you might feel wonder when a character behaves in an unexpected way, or the lighting and décor in some scene is striking or strange.

- **Low (0)**: you feel wonder a little bit (weak feeling)
 - Moving the ball to the left (**green**)
- **High (1)**: you feel wonder a lot (strong feeling)
 - Moving the ball to the right (**red**)



1.000000

Instructions

- Move cursor inside the box and press SPACE to play movie. Press SPACE again to pause. DONOT use the play/pause button of the video player
- Move cursor left or right while watching the movie to annotate emotions
- Moving cursor outside the box will pause annotation

Show instruction

Fig. 3.8 A snapshot of aesthetic emotion annotations in the movie *Islands* on Amazon Mechanical Turk.

The C. LIRIS-ACCEDE database was collected to provide researchers resources to work on affective content analysis [13]. However, aesthetic emotions evoked in movie audiences have been largely overlooked in previous movie emotion research since movies are a form of art and spectators have emotional responses specific to their artistic aspects. In this work, we

select five aesthetic emotions related to spectators' feelings, namely *awe*, *boredom*, *disgust*, *being touched*, and *wonder* based on various studies on aesthetic emotions to be continuously annotated over time in our experiment [52, 75, 130].

Much of previous work assumes that the emotional state of a human can be identified by a concrete category or a unique value. This assumption can become valid when we collect multiple individuals' emotional responses. Then, we can find emotions that are felt by the majority of them. Previous work on emotion recognition has suggested that having more than 6 annotators provides reliable emotion annotations [23].

Our protocol for collecting aesthetic annotations of the C. LIRIS-ACCEDE database is described below. To collect a large amount of aesthetic annotations in a time and cost efficient manner, aesthetic emotions were annotated online by using Amazon Mechanical Turk. We segmented the selected 30 movies into 84 excerpts with a mean duration of 316 seconds, preserving the context of movie scenes. Then, we collected at least 9 annotations from different annotators for each of 84 movie excerpts and the five aesthetic emotions. Firstly, a basic questionnaire about age, gender, English fluency, and favorite genre was provided. Then, annotators were asked to complete a detailed personality test that measures an individual personality on the Big Five Factors (dimensions) [97]. Finally, annotators were

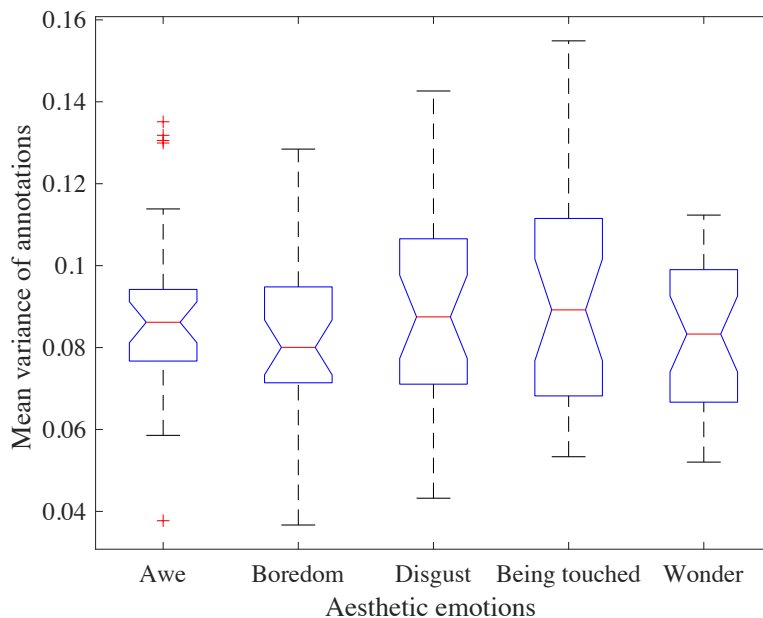


Fig. 3.9 Mean variance of aesthetic emotion annotations over the C. LIRIS-ACCEDE database.

instructed to continuously annotate their feelings of one out of the five aesthetic emotions over time by moving their cursor in an annotation box, as shown in Figure 3.8. The left end

of the box represents not feeling the emotion at all and right end of the box represents an extremely strong feeling of the emotion. Each movement of the cursor was saved with a time stamp. The intensity of all aesthetic scores varies between 0 and 1. We obtained more than 4000 annotations from around 1000 annotators with various cultural and educational backgrounds. To merge multiple annotations, we applied resampling to raw signals and computed means of the annotations collected for each second. Finally, we used a moving average filter with the sliding window of length 30 signal points to remove the noise from the signals due to unintended movements of the cursor.

We studied agreement on annotations of each aesthetic emotion. We used the resampled annotations at the second level to compute the variance over all annotations at the given time stamp. We assume that a variance value can indicate an agreement level among annotators. Figure 3.9 shows the distributions of the average variance of the annotations on each movie from the C. LIRIS-ACCEDE database. To compare the agreement on each aesthetic emotion annotations, we performed an one-way ANOVA. As a result, we report that the means of all mean variance distributions are not significantly different (p -value = 0.56). This means that the agreement on annotations for these five aesthetic emotions is at the same level. However, the ranges of mean variances are bigger for boredom, disgust, and being touched annotations than awe and wonder annotations. We can also see the outliers for awe annotations. Thus, the annotators had problems with agreement on awe intensity since it is a complex aesthetic emotion. This suggests that some emotions evoked by movies are very difficult to annotate and require multiple annotators to find emotional ground truth. Using the crowd-sourced annotation platform allowed us to collect aesthetic emotion annotations from many annotators with various cultural and educational backgrounds. Moreover, this suggests that we obtained the sufficient representation of the annotator population to provide the gold standard of movie aesthetic emotions.

To verify the quality of the annotations that we collected, we investigated the differences between average values and trend changes of merged annotations for each emotion. In particular, we calculated the average values of these five aesthetic emotion intensity over each movie. Even though Figure 3.10 shows that the average values vary from one movie to another for each aesthetic emotion, an one-way ANOVA proves that there is no significant difference between the means of all mean value distributions (p -value = 0.22). Thus, we do not observe a significant bias of aesthetic gold standard annotations towards a specific emotion. It means that the movies from the C. LIRIS-ACCEDE database contain a balanced number of scenes that elicit various aesthetic emotions. Moreover, the mean values of disgust vary the least over the movies, unlike the rest of the aesthetic emotions. This implies that almost all movies evoke feelings of disgust at a similar level of intensity.

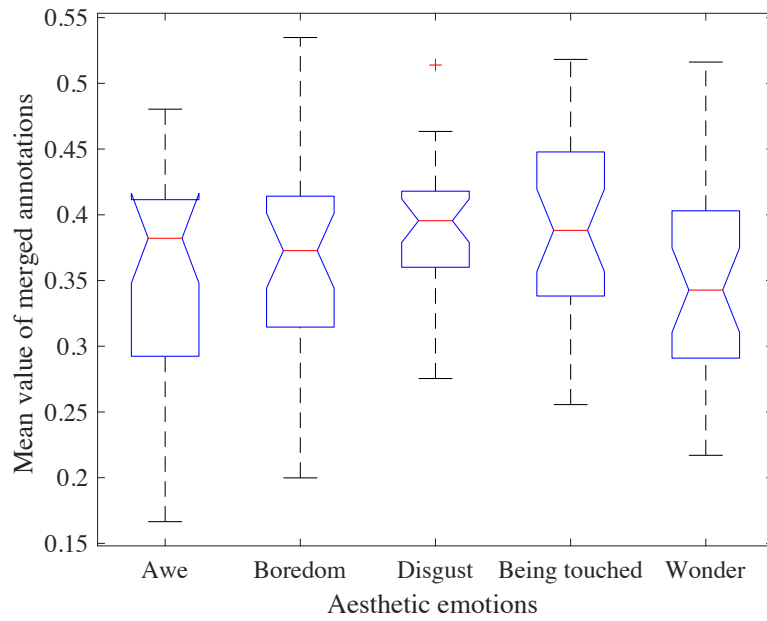


Fig. 3.10 Mean value of merged aesthetic emotion annotations over the C. LIRIS-ACCEDE database.

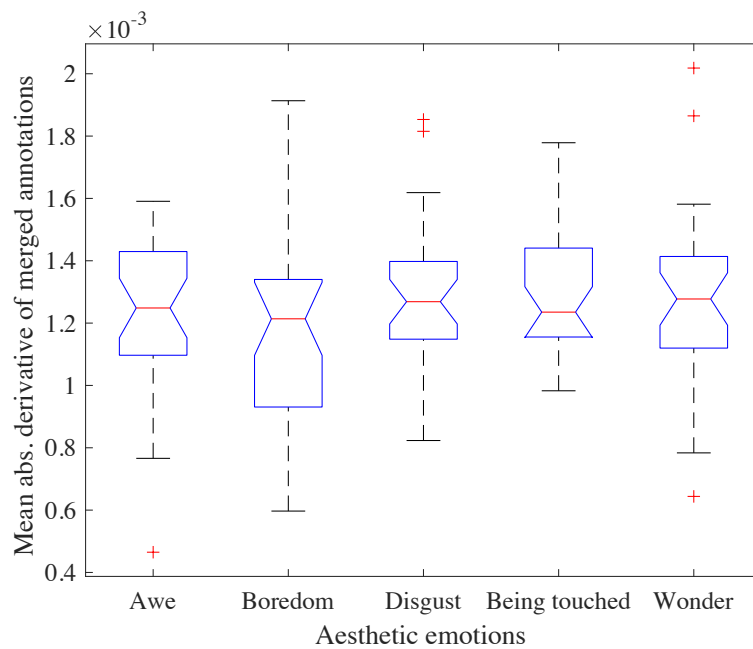


Fig. 3.11 Mean absolute (abs.) derivative of merged aesthetic emotion annotations over the C. LIRIS-ACCEDE database.

Besides, we investigated dynamics of changes in annotation trends, using the absolute derivative of merged annotations for each emotion. In order to do this, we calculated the absolute values of the first derivatives over each movie and then we averaged them, as presented in Figure 3.11. As a result, we verified if the dynamics of trend changes could be influenced by aesthetic emotion categories. To address this question and compare the distributions of mean absolute derivatives, we ran an one-way ANOVA. The analysis do not show significant differences between means of these distributions (p-value = 0.44). We can see that all the aesthetic emotions are gradually elicited over time. This is made on purpose by moviemakers. Nevertheless, the box plots for awe, disgust, and wonder have a few outliers that might indicate the rapid changes of aesthetic emotion intensities for some movies. This can partially explain difficulties in the annotation process and the variance of aesthetic annotations.

Chapter 4

Aesthetic highlight detection in movies from synchronization of spectators' reactions

In this Chapter, we mainly focus on understanding spectators' responses to aesthetic highlights in full-length movies [141]. This corresponds to scenes with high aesthetic attributes in terms of form and content. These scenes are constructed on purpose by the moviemakers in order to establish a connection between the spectators and the movie, and allow spectators to be engaged with the movie. We are interested in analyzing and interpreting movie audiences' aesthetic experiences and reactions to aesthetic movie content. It can make contributions to several applications: aesthetic scene detection, aesthetic scene design, video summarization and movie recommendation systems with aesthetic ratings.

In this Chapter, we also investigate spectators' responses to aesthetic highlights in a social context when spectators watch movies together in a movie theater. In particular, we are interested in the understanding of their physiological and behavioural reactions. We assume that spectators can display similar behaviours and have similar physiological reactions when they watch movies together because: (i) aesthetic choices of filmmakers are made to elicit specific emotional reactions (e.g., special effects, empathy, and compassion toward a character, etc.) and (ii) watching a movie together causes spectators' affective reactions to be synchronized through emotional contagion [90]. For these reasons, we measure synchronization among multiple spectators' physiological and behavioural signals. Then, we use the level of synchronization among spectators' reactions to detect aesthetic highlights in movies.

In order to uncover a relationship between the occurrences of aesthetic highlights in films

and multiple spectators' affective states, we address the following research questions (see a full list of research questions in Section 1.5):

1. Do aesthetic highlights elicit emotions in movie audiences? (RQ1)
2. Can the level of synchronization among spectators' reactions be used to detect the different categories of aesthetic highlights? (RQ2)
 - If it is possible, which synchronization measures are the most reliable to efficiently detect aesthetic highlights?

Below we emphasize what are, to the best of our knowledge, the main contributions of this chapter, highlighting the novelty compared to the state of the art research:

- We are first to quantitatively analyze the direct relationship between emotions induced in movie audiences (arousal-valence space) and the occurrences of aesthetic highlights in movies.
- We investigate different approaches to synchronization estimation, such as pairwise, group, and overall synchronization measures to analyze multiple spectators' reactions. There have been no comprehensive and comparative studies on synchronization measures including multiple spectators' physiological and behavioural responses.
- We use the level of synchronization of movie audiences' EDA and ACC measurements to detect aesthetic highlights. Then, we find that the pairwise approach to synchronization performs aesthetic highlight detection efficiently compared to other measures for several movie genres.
- We create one of the largest databases of aesthetic highlight annotations that will help to study movie audiences' responses to aesthetic content. This database contains aesthetic highlight annotations of 30 full-length movies derived from 9 movie genres: action, adventure, animation, comedy, documentary, drama, horror, romance, and thriller.

4.1 Detection system of aesthetic highlights in movies

In this thesis we assume that physiological and behavioural responses of spectators in the context of watching movies together are discriminative for aesthetic highlight detection in movies. That is why we process spectators' EDA and ACC measurements. The utility and suitability of these signals for emotion and behaviour assessments have been confirmed

by several previous studies [73, 125]. In order to detect aesthetic highlights and analyze spectators' responses to aesthetic stimuli, we propose an unsupervised highlight detection system based on physiological and behavioural reactions of spectators watching movies together, as shown in Figure 4.1. It is composed of three parts: signal preprocessing, synchronization estimation, and detection based on the synchronization level. Filtering and time windowing are included in the signal preprocessing while several synchronization measures are used for the synchronization estimation and highlight detection.

We formulate highlight detection as a binary classification problem (highlight and non-highlight class) to respond to our research question RQ2. There is a possible overlap between different aesthetic highlights, e.g., highlights H3 and highlights H5 (see Section 3.3). A movie scene can contain more than one highlight, for example, spectacular moments and character development [113]. In this thesis we focus our work on detecting a particular category of aesthetic highlights independently of those overlaps. We started with the preprocessing of

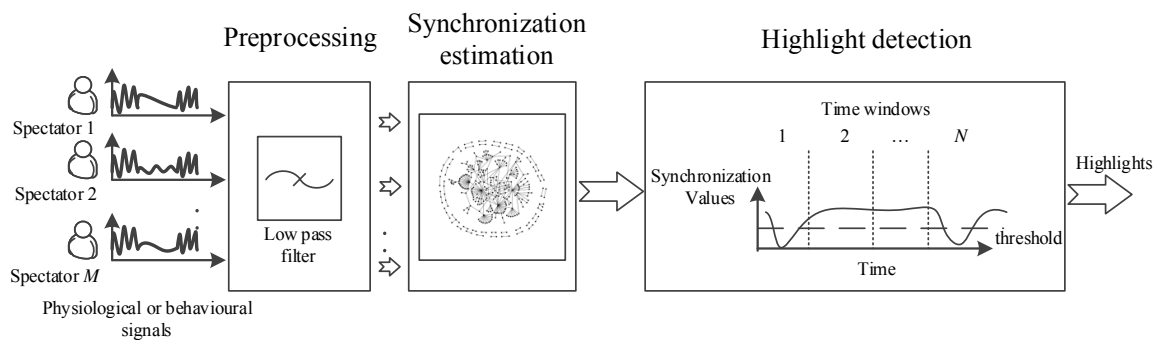


Fig. 4.1 The scheme of an unsupervised highlight detection system based on synchronization among spectators' physiological or behavioural signals [141].

spectators' physiological and behavioural signals that were filtered by a low-pass filter to remove noise and distortions. To measure synchronization among spectators from the C. LIRIS-ACCEDE database, we used the following experimental settings. EDA and ACC measurements of spectators were filtered by a third order lowpass Butterworth filter with cutoff frequency 0.3 Hz, as shown in Figure 4.1. The filtered signals were then segmented into overlapping time windows with a time step and a window length equal 1 second and 5 seconds, respectively (some measurements were discarded due to the amount of artifacts).

The main component of our detection system is an estimator of the synchronization level among spectators that employs synchronization measures. To compute the amount of synchronization for each time window, we can use different synchronization measures. We expect that the value of synchronization increases when spectators jointly react to aesthetic scenes. The choice of a synchronization estimator is related to the type of analysis of

synchronization which we attempt to carry out. We can analyze synchronization at pairwise (local descriptors), overall (global descriptors), and group (trade-off between local and global descriptors) levels to capture different patterns in multiple spectators' responses [55]. In particular, the synchronization analysis at the group level aims at finding subgroups of synchronized responses. Also, the properties of aesthetic highlights, e.g., duration, frequency of occurrence, and sampling frequency of recorded EDA and ACC measurements limit the choice of synchronization measures that can be used.

We evaluate the discriminative power of synchronization measures for aesthetic highlight detection by means of carrying out a receiver operating characteristic (ROC) curve analysis with calculating the area under the ROC curve (AUC) as a performance metric. Aesthetic highlights are determined for each time window based on the value of the estimated synchronization among spectators. If the value of a synchronization measure is higher (lower) than a threshold, we assign the time window to highlight (non-highlight) scenes [112, 113, 139]. The crucial issue of our unsupervised highlight detection system is the choice of a threshold for a given synchronization measure. To overcome this, we evaluate the performance of the detection system for varying thresholds. We calculate the true positive rate and the false positive rate for each value of the threshold. Then, we generate ROC curves and calculate their AUCs.

4.2 Synchronization measures

In this thesis we attempt to investigate different synchronization measures with particular emphasis on the estimation of physiological and behavioural synchronization. This includes all constraints related to highlight detection e.g., sampling frequency of physiological and behavioural signals, the number of signals, the duration of time window, the duration of highlights, etc. To understand different approaches to synchronization estimation for analysis of multiple spectators' signals, we divide synchronization measures into 3 classes: pairwise, group, and overall measures. Pairwise measures establish synchronization between pairs of signals. Group measures can analyze clusters of synchronized signals. Overall measures can simultaneously process an arbitrary number of signals. The obvious disadvantage of group and overall measures is that they are not able to provide information on synchronous activities of single signal pairs due to their global properties. By contrast, pairwise measures can successfully capture synchronization when there are synchronous changes in pairs of signals.

We underline that although we study a large variety of synchronization measures, it is not possible to include all existing synchronization measures in this thesis. Novel measures

are constantly being developed. Our choice of synchronization measures is limited by the fact that we are not able to estimate a covariance matrix by means of some of the pairwise synchronization measures (namely, dynamic time warping, shape distribution distance and nonlinear interdependence) because of their properties. Dynamic time warping is not a normalized measure while shape distribution distance and nonlinear interdependence would require distance measurements to neighbouring time windows.

In this Section we briefly review those three approaches to synchronization (pairwise, group, and overall measures) and we propose to apply them to spectators' physiological and behavioural signals in order to detect aesthetic highlights in movies. A level of a synchronization measure should reveal the synchronized reactions of spectators during watching movies.

For spectators' EDA and ACC signals $\{x_{i,t}\}$, we consider time windows $\{x_i(l)\}$ for $i = 1, 2, 3, \dots, M$, $t = 1, 2, 3, \dots, \mathcal{N}$, and $l = 1, 2, 3, \dots, N$, where M is the number of spectators' signals, \mathcal{N} is the number of samples, and N is the number of time windows.

4.2.1 Pairwise synchronization

The key point of pairwise measures is to measure the amount of synchronization only at the local level, i.e. between two time series. When the number of signals is more than 2, the synchronization value is calculated as a mean synchronization value over all possible non-overlapping pairs of signals at a given time. We begin our review with mentioning about the Pearson's Correlation Coefficient (CC) that is perhaps the most common measure for linear interdependence between two signals and the coherence function quantifies that linear correlations in the frequency domain (find the details [144]). There are some attempts to propose an extension of CC, such as correntropy coefficient [86], modifications of the partial coherence [55, 149]. Although the amplitudes of signals are statistically independent, their instantaneous phases can be strongly synchronized. This refers to phase synchronization [120]. Granger causality is considered as a synchronization measure that is derived from linear stochastic models of time series which measures linear dependencies between signals [18, 83]. Non-linear extensions of Granger causality have been proposed in [6, 34]. Several synchronization measures come from information theory [47]. Mutual information is perhaps the most well-known synchronization measure among them. To study nonlinear dependencies between time series, mutual information is calculated in the time and time-frequency domain [8, 116]. Stochastic event synchrony represents a family of synchronization measures that quantifies the similarity between point process extracted from time-frequency representations of signals [56].

We introduce below the following pairwise synchronization measure: Dynamic Time Warping, Shape Distribution Distance, and Nonlinear Interdependence. In this thesis, we used the mean value of a pairwise synchronization measure over all possible pairs of signals at a given time stamp l as the value of the synchronization measure [112, 115].

Dynamic time warping

Let us suppose there are two time windows $x_i(l)$ and $x_j(l)$ of m samples for $i, j = 1, 2, 3, \dots, M$, $l = 1, 2, 3, \dots, N$. In order to align these two signals, we create a m -by- m matrix D_W which contains the Euclidean distances between pairs of samples from time window $x_i(l)$ and $x_j(l)$ [138]. A warping path W between two time windows is a set of matrix elements which creates a mapping between them. A warping path W of length L is defined as follows

$$W = w_1, w_2, \dots, w_L, \quad (4.1)$$

where w_1, w_2, \dots, w_L are the elements of the matrix D_W and $m \leq L < 2m - 1$.

The total cost $c_W(x_i(l), x_j(l))$ of the warping path W is expressed by

$$c_W(x_i(l), x_j(l)) = \sum_{p=1}^L w_p. \quad (4.2)$$

The optimal warping path between two time windows $x_i(l)$ and $x_j(l)$ is a warping path W^* that has a minimal total cost among all possible warping paths.

Dynamic Time Warping (DTW) distance between two time windows $x_i(l)$ and $x_j(l)$ is the total cost of the warping path W^* , as follows [16]

$$d_{DTW}(x_i(l), x_j(l)) = c_{W^*}(x_i(l), x_j(l)). \quad (4.3)$$

The distance $d_{DTW}(x_i(l), x_j(l))$ is computed for each pair of time windows $x_i(l)$ and $x_j(l)$ for $i, j = 1, 2, 3, \dots, M$, $l = 1, 2, 3, \dots, N$. The computational cost of the dynamic time warping is $O(Nm^2M^2)$ and is bounded by the number M of signals, the number N and the size m of time windows.

Shape distribution distance

Time-delay coordinate embedding is used in analysis of dynamical systems [183]. This method embeds a scalar time series into a m dimensional space to reconstruct the state space trajectory of the underlying dynamical system. For each sample $x_{i,t}$, $i = 1, 2, 3, \dots, M$, $t = 1, 2, 3, \dots, \mathcal{N}$ of time series $\{x_{i,t}\}$, a representation of the delay-coordinate embedding

can be expressed as the following vector $X_{i,t}$ which consists of m components

$$X_{i,t} = [x_{i,t}, x_{i,t+\tau}, x_{i,t+2\tau}, \dots, x_{i,t+(m-1)\tau}], \quad (4.4)$$

where τ is the index delay and m is the embedding dimension. A theoretical discussion on the choice of these parameters is out of the scope of this thesis. The index delay and the embedding dimension are selected based on the duration of aesthetic highlights. Diffusion maps of time-delay coordinate embedding provides a new low dimensional parametrization that is able to capture the changes in physiological and behavioural signals. When diffusion maps are applied [44], an affinity metric $K(x_{i,g}, x_{i,h})$ is defined between pairs of the samples $x_{i,g}$ and $x_{i,h}$ based on their representation in time-delay coordinate $X_{i,g}$ and $X_{i,h}$, respectively. Then, we only take into account a collection \mathcal{M} of samples to define the following kernel

$$K(x_{i,g}, x_{i,h}) = e^{-\frac{\|x_{i,g} - x_{i,h}\|}{\varepsilon}}, \quad (4.5)$$

where ε is a scale parameter of the affinity metric (the parameter is selected based on the mean distance between points in the m -dimensional space) and $g, h = 1, 2, 3, \dots, \mathcal{M}, \mathcal{M} < \mathcal{N}$.

We can consider the collection \mathcal{M} as nodes of an undirected symmetric graph, where two nodes $x_{i,g}$ and $x_{i,h}$ are connected by an edge with the affinity weight $K(x_{i,g}, x_{i,h})$. We pursue the construction of a Markov chain (random walk) on the graph nodes by normalizing the kernel $K(\cdot, \cdot)$. Let K be the kernel matrix, and let $P = D^{-1}K$ be the corresponding transition matrix, where D is a diagonal matrix with elements $D_{gg} = \sum_{h=1}^{\mathcal{M}} K(x_{i,g}, x_{i,h})$. In sequence, we calculate P_ζ analogues to P , where $P(x_{i,g}, x_{i,h})$ is the probability of transition in a single step from node $x_{i,g}$ to node $x_{i,h}$. In addition, we define $P_\zeta(x_{i,g}, x_{i,h})$ as the transition probability in ζ steps from node $x_{i,g}$ to node $x_{i,h}$. A definition of the diffusion distance $D_\zeta(x_{i,g}, x_{i,h})$ between pairs of samples is the following [44]

$$D_\zeta(x_{i,g}, x_{i,h}) = \sqrt{\sum_{z=1}^{\mathcal{M}} (P_\zeta(x_{i,g}, x_{i,z}) - P_\zeta(x_{i,h}, x_{i,z}))^2 w(x_{i,z})}, \quad (4.6)$$

where $w(x_{i,z})$ is a normalization weight. Intuitively, two points are similar when many short paths with large weights connect them. When applying spectral graph theory to the transition matrix P_ζ , the diffusion distance $D_\zeta(x_{i,g}, x_{i,h})$ can be computed by means of its eigenvalues $\{\lambda_T\}$ that tend to 0 and have moduli strictly less than 1 and the corresponding eigenvectors $\{\phi_T\}$ [44]. Let $\Phi_\zeta(x_{i,t})$ for some $\zeta \geq 0$ be the diffusion maps of time series samples $x_{i,t}$ for

$i = 1, 2, 3, \dots, M, t = 1, 2, 3, \dots, \mathcal{M}$ into the Euclidean space \mathbb{R}^s that is expressed by

$$\Phi_{\zeta}(x_{i,t}) = [\lambda_1^{2\zeta} \varphi_1(x_{i,t}), \dots, \lambda_s^{2\zeta} \varphi_s(x_{i,t})], \quad (4.7)$$

where $s \in \{1, 2, 3, \dots, \mathcal{M} - 1\}$ is the new space dimensionality. It is shown that the diffusion distance between samples $x_{i,g}$ and $x_{i,h}$ is equal to the Euclidean distance in the diffusion map space, as follows [44]

$$D_{\zeta}(x_{i,g}, x_{i,h}) = \|\Phi_{\zeta}(x_{i,g}) - \Phi_{\zeta}(x_{i,h})\|. \quad (4.8)$$

The idea is to measure the similarity between local shapes of reconstructed signal manifolds by means of time-delay embedding and diffusion maps [139]. We propose a geometric framework which computes the amount of synchronization between a pair of spectators' physiological or behavioural signals. To capture the unique local geometric properties of a signal manifold, we introduce the local shape cumulative distribution function $F_{x_{i,t}}^{\sigma}(\delta)$ of pairwise diffusion distances for each sample $x_{i,t}$ and its delay samples $x_{i,t}, x_{i,t+1}, \dots, x_{i,t+\sigma}$ defined by

$$F_{x_{i,t}}^{\sigma}(\delta) = \int 1_{\tilde{D}_{\zeta}(x_{i,t}, x_{i,t+q}) \leq \delta} d\mu, \quad (4.9)$$

where $q \in \{1, \sigma\}$, μ is a counting measure and $1_{\tilde{D}_{\zeta}(x_{i,t}, x_{i,t+q})}$ is an indicator function with respect to a delay sample on manifolds. Moreover, σ should be chosen to obtain enough a number of samples required for density estimation ($\sigma = 50$). Besides, $\tilde{D}_{\zeta}(\cdot, \cdot)$ is the cosine distance in the diffusion maps space that can be derived from the Euclidean dot product. Normalization is advantageous to the local shape distribution, as follows

$$\mathcal{F}_{x_{i,t}}^{\sigma}(\delta) = \frac{F_{x_{i,t}}^{\sigma}(\delta)}{F_{x_{i,t}}^{\sigma}(\infty)}. \quad (4.10)$$

For two time series $\{x_{i,t}\}$ and $\{x_{j,t}\}$, the synchronization measure that is named Shape Distribution Distance (SDD) is derived from calculating the Kolmogorov-Smirnov distance between two local shape distributions of their manifold representations for each time step t . SDD is defined, as follows

$$S_{\sigma}(x_{i,t}, x_{j,t}) = \max_{\delta} |\mathcal{F}_{x_{i,t}}^{\sigma}(\delta) - \mathcal{F}_{x_{j,t}}^{\sigma}(\delta)|. \quad (4.11)$$

If two time series are synchronized, $S_{\sigma}(x_{i,t}, x_{j,t})$ is equal to 0 for all time steps. The complexity of the shape distribution distance is $O(M^2 N^3)$ and is bounded by the number M of signals and the number N of time windows.

Nonlinear interdependence

The concept of nonlinear interdependence comes from studies on generalized synchronization that evaluate the interdependence between signals in a reconstructed state space domain [158]. Nonlinear Interdependence (NI) measures the geometrical similarity between the state space trajectories of two dynamical systems. Time-delay embedding is applied to two time series $\{x_{i,t}\}$ and $\{x_{j,t}\}$ for $i, j = 1, 2, 3, \dots, M$, $t = 1, 2, 3, \dots, \mathcal{N}$ to reconstruct the trajectories analogous to shape distribution distance [183]. The mean square Euclidean distance of each sample $x_{i,t}$ to its K nearest neighbours $x_{i,r}$ for $r = 1, 2, 3, \dots, K$ in the delay-coordinate embedding is

$$R^K(x_{i,t}) = \frac{1}{K} \sum_{r=1}^K (X_{i,t} - X_{i,r})^2, \quad (4.12)$$

and the mean squared Euclidean distance conditioned by the equal time partners of the K nearest neighbours of $x_{j,t}$ is

$$R^K(x_{i,t}|x_{j,t}) = \frac{1}{K} \sum_{r=1}^K (X_{i,t} - X_{j,r})^2. \quad (4.13)$$

NI measure is defined as [154]

$$S^K(x_{i,t}|x_{j,t}) = \frac{R^K(x_{i,t})}{R^K(x_{i,t}|x_{j,t})}. \quad (4.14)$$

The number of nearest neighbours should be selected to accurately estimate an average distance a point to its nearest neighbours, e.g., $K = 50$. To make the nonlinear interdependence symmetric, we consider $S^K(x_{j,t}|x_{i,t})$ and we then average these two parameters. When two time series are synchronized (resp. desynchronized), the value of the nonlinear interdependence is close to 1 (resp. 0) for all samples. Searching K nearest neighbours of N time windows for calculating the nonlinear interdependence among all possible pairs of signals M can be found in $O(M^2KN \log N)$ time.

4.2.2 Group synchronization

Group synchronization measures intend to be a trade-off between pair and overall approaches to synchronization. It can capture synchronization among groups of signals based on the connectivity of signal clusters. This approach to synchronization contains a multivariate measure which ascribes a single value to groups of signals in comparison with pairwise or overall measures [140]. We detail below how we can adapt Periodicity Score (PS) [148] to measure synchronization among groups of spectators.

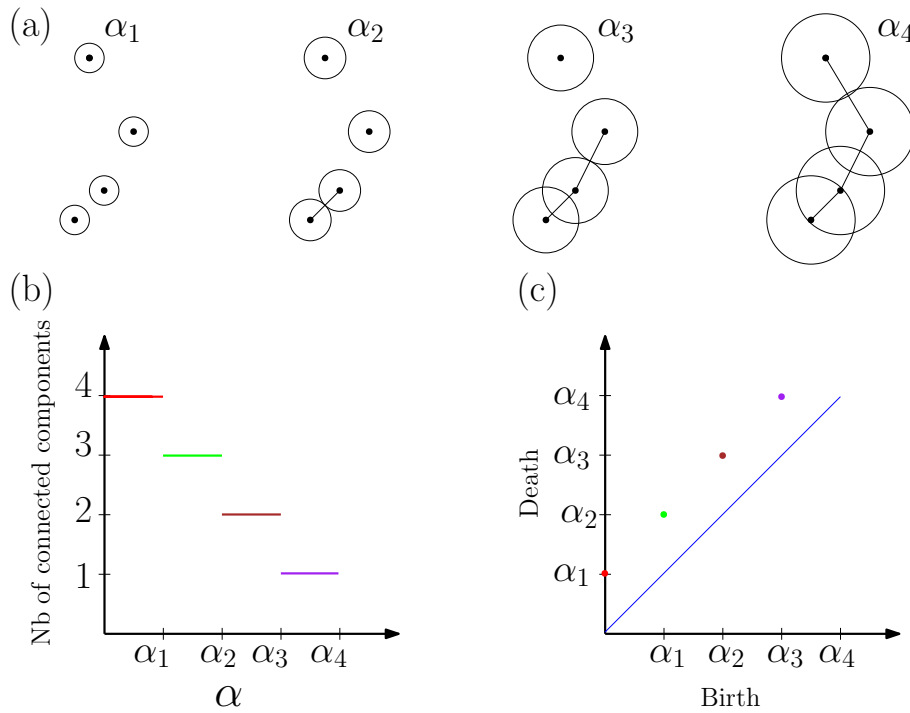


Fig. 4.2 An example of: (a) the filtered Vietoris-Rips complex, (b) a number of connected components for different values of filtration parameter, (c) the persistent diagram [140].

Periodicity score

In this section we detail the usage of PS to measure synchronization among spectators [147, 148]. We analyze time windows $\{x_i(l)\}$ for $i = 1, 2, 3, \dots, M, l = 1, 2, 3, \dots, N$ of spectators' signals as a sequence of points encoded on the Grassmann manifolds preserving their intrinsic dependencies. The Grassmann manifold $G(k, n)$ is defined as a set of k dimensional linear subspaces of the n dimensional vector space. To map the time windows of spectators' physiological or behavioural signals to Grassmann manifolds, we apply Reduced (Thin) Singular Value Decomposition (RSVD) to their Short Time Fourier Transform (STFT). Finally, we associate PS with the synchronized spectators' physiological and behavioural signals.

STFT. We apply STFT to given time windows $\{x_i(l)\}$ for $i = 1, 2, 3, \dots, M$, and we yield a sequence of $x_{\hat{t}, \hat{f}}^{i,l}$ in the time and frequency domain, where \hat{t} is a time frame index and \hat{f} is a frequency band index. Each time window $x_i(l)$ is split into segments with an overlap of 50% to apply STFT. Let $S_x^{i,l}(\hat{t}, \hat{f})$ be the squared magnitude of the STFT, as follows

$$S_x^{i,l}(\hat{t}, \hat{f}) = \|x_{\hat{t}, \hat{f}}^{i,l}\|^2. \quad (4.15)$$

RSVD. Then, we map time windows $\{x_i(l)\}$ of all signals on real Grassmann manifolds to recover the intrinsic dependencies among them [36]. The real Grassmann manifold $G(k, n)$ parametrizes all k dimensional subspaces of the vector space \mathbb{R}^n . A sequence of corresponding matrices $S_x^{i,l}(\hat{t}, \hat{f})$ for $i = 1, 2, 3, \dots, M$ can be mapped to the points on the manifold $G(k, n)$ using RSVD. If we compute RSVD of the matrix $S_x^{i,l}(\hat{t}, \hat{f})$, as follows

$$S_x^{i,l}(\hat{t}, \hat{f}) = U^i \Sigma^i V^{iTr}, \quad (4.16)$$

then the columns of the $n \times k$ orthogonal matrix U^i are a non-unique basis for the column space of $S_x^{i,l}(\hat{t}, \hat{f})$. Thus, U^i can be used to represent the matrix $S_x^{i,l}(\hat{t}, \hat{f})$, and can be identified with a point on the Grassmann manifold $G(k, n)$ [36]. Once the time windows are mapped to a sequence of points on $G(k, n)$, the pairwise distances between these points can be found using a function of the angles between subspaces.

Let U^i and U^j be two k dimensional subspaces representing $x_i(l)$ and $x_j(l)$, we measure the similarity $d_{min}(U^i, U^j)$ of two points on the Grassmann manifold $G(k, n)$ by applying the minimum correlation distance [87]

$$d_{min}(U^i, U^j) = \sin \theta_k, \quad (4.17)$$

where $0 \leq \theta_1 \leq \theta_2 \leq \dots \leq \theta_k \leq \frac{\pi}{2}$ are principal angles between two subspaces.

PS. Finally, we introduce the basics of persistent homology: filtrations and persistence diagrams [79, 147, 148]. Once the sequence of $S_x^{i,l}(\hat{t}, \hat{f})$ for $i = 1, \dots, M$ matrices is mapped to $G(k, n)$ and defines a metric space $(U = \{U^1, \dots, U^M\}, d_{min}(\cdot, \cdot))$, we recall the definition of the Vietoris-Rips complex $Rips_\alpha(U)$ as the set of the simplices $[U^1, \dots, U^P]$ such that $d_{min}(U^a, U^c) \leq \alpha$ for $a, c = 1, \dots, P$. There is an inclusion of $Rips_\alpha(U)$ in $Rips_\beta(U)$ for any $\alpha \leq \beta$. The sequences of inclusions are called filtrations $Filt_\alpha(U)$. An example of the filtered Vietoris-Rips complex is presented in Figure 4.2(a). Persistence diagrams study the evolution of the topology of a filtration, and to capture properties of the metric which is used to generate the filtration. Existing connected components are merged for 0–th homology when α increases, as shown in Figure 4.2(b). Persistent homology tracks the birth (appearance) b and death (disappearance) d of all connected components that are illustrated in Figure 4.2(c). The maximum persistence $mp(dgm(\{x_i(l)\}))$ of a persistence diagram $dgm(\{x_i(l)\})$ for $i = 1, 2, 3, \dots, M$ is defined as follows [148]

$$mp(dgm(\{x_i(l)\})) = \max_{(b,d) \in dgm(\{x_i(l)\})} pers(b, d), \quad (4.18)$$

where $\text{pers}(b, d) = d - b$ for $(b, d) \in \text{dgm}(\{x_i(l)\})$, and as ∞ otherwise. Finally, we can provide PS $S(\{x_i(l)\})$ [148]

$$S(\{x_i(l)\}) = \frac{\text{mp}(\text{dgm}(\{x_i(l)\}))}{\sqrt{3}}. \quad (4.19)$$

PS that is the normalized maximum persistence $\text{mp}(\text{dgm}(\{x_i(l)\}))$ can help us to quantify synchronization among signals because it is capable of measuring their intrinsic geometric dependencies. PS can measure synchronization among groups of signals based on the connectivity of signal clusters. When $S(\{x_i(l)\})$ equals 0, it means that we cannot explore any structure in the data. If a value of $S(\{x_i(l)\})$ rises close to 1, we find some strong connectivity structure in the data. The computational complexity of the periodicity score is $O(mNM^2)$ and is bounded by the number M of signals, the number N and the size m of time windows.

4.2.3 Overall synchronization

The overall approach to synchronization simultaneously processes all the time series and considers them as components of a single interdependent system. Omega complexity is a synchronization measure which is derived from applying a principal component analysis to a covariance matrix [160]. Given M signals, the multivariate time series is viewed as series of temporary maps whose sequence over time defines a trajectory of a dynamical system in a M dimensional space. Omega complexity evaluates in particular the complexity of a trajectory by means of examining its shape along the principal dimensions. S-estimators are an extension of omega complexity based on Shannon entropy [95]. We detail below the concept of the S-estimators with different estimators of the covariance matrix.

S-estimators

All signals can be viewed as a representation of a trajectory that can be modeled in a high-dimensional state-space. The dimensionality of the trajectory in the state-space can be assessed based on a principal component analysis of an estimated covariance matrix. Minimum entropy characterizes the situation when a few normalized eigenvalues only are nonzero, showing a high level of synchronization. Let $C^l = (C_{ij}^l)$ be a matrix in which C_{ij}^l reflects cross-dependence between time windows $x_i(l)$ and $x_j(l)$ for $i, j = 1, 2, 3, \dots, M$, $l = 1, 2, 3, \dots, N$, such as correlation, phase synchronization (phase locking value), synchronization likelihood, windowed mutual information, event synchronization, heat kernel, and diffusion map [50, 95].

The eigenvalue decomposition of C^l is

$$C^l v_u^l = \lambda_u^l v_u^l, \quad (4.20)$$

where eigenvalues $\lambda_1^l \leq \lambda_2^l \leq \dots \leq \lambda_M^l$ are in increasing order and v_u^l , $u = 1, 2, 3, \dots, M$ are corresponding eigenvectors. As the matrix C^l is a real and symmetric, all eigenvalues are real numbers, and the trace of C^l is equal to the number of signals M .

S-estimator is proposed to measure synchronization among signals by means of the distributions of the eigenvalues of the covariance matrix C^l , as proposed in [25]

$$S_l = 1 + \frac{\sum_{u=1}^M \frac{\lambda_u^l}{M} \log\left(\frac{\lambda_u^l}{M}\right)}{\log(M)}. \quad (4.21)$$

When all the signals are synchronized (resp. desynchronized), the value of the S-estimator is close to 1 (resp. 0) for all time windows. The computational complexity of the S-estimator is $O(NM^3)$ and is bounded by the number M of signals and the number N of time windows.

4.3 Results

4.3.1 Emotions and aesthetic highlights

Previous studies have confirmed that physiological reactions of spectators are linked with their emotional states [109, 209]. In our studies we attempt to prove that aesthetic scenes are able to influence the affective states of spectators. In order to evaluate whether aesthetic highlights in various movies evoke emotions, we used a meta analysis [21]. We related the occurrences of highlights in movies to felt emotions (level of arousal and valence) by the spectators from the C. LIRIS-ACCEDE database. To apply a meta-analysis, we considered this database as a set of empirical experiments about a given topic: the level of emotions (arousal/valence) while watching aesthetic highlights in movies [21]. Each category of aesthetic highlights was analyzed independently of the others. Effect-sizes were calculated over individual movies. The effect size is the standardized mean difference that is defined as the difference between mean values of continuous emotion annotations of highlight and non-highlight intervals divided by their pooled standard deviation. Positive values indicate a higher level of arousal/valence of highlight scenes in comparison with non-highlight scenes, whereas negative values of the effect size indicate a lower level.

To integrate the results of the experiments, we used a fixed-effect model [21] that weights each effect size estimate as a function of its precision to calculate an overall estimate. Weights

are inversely proportional to the variances of effect size estimates. In our studies we followed Cohen's benchmarks [43] for the interpretation of the practical significance of a weighted average effect size. We assume that the values around 0.2, 0.5, and 0.8 can be interpreted as small, medium and, large effect sizes, respectively. The weighted average effect size of arousal/valence on the C. LIRIS-ACCEDE database is reported in Table 4.1. In the case of arousal, medium positive effect sizes (>0.5) are found for spectacular highlights H1, character development highlights H3, and theme development highlights H5. Also, medium negative effect sizes (<-0.5) are observed for spectacular highlights H1 and character development highlights H3 regarding valence.

Table 4.1 The weighted average effect size (fixed-effect model) of arousal and valence during aesthetic highlights over all the C. LIRIS-ACCEDE database [141].

Emotions \ Highlights	H1	H2	H3	H4	H5
Arousal	0.76	-0.23	0.70	0.04	0.53
Valence	-0.64	-0.11	-0.55	0.11	-0.22

To investigate the relationship between movie genres and emotions induced by aesthetic highlights, we carried out the same meta analysis for each of the 9 movie genres in the C. LIRIS-ACCEDE (see a list of movie genres in Section 3.2). We infer that the direction and the power of the average effect size strongly depends on the movie genre for both arousal and valence, as shown in Tables 4.2 and 4.3. Strong emotional reactions are expected to be associated with spectacular highlights H1, such as using special effects, increasing saturation of colors, playing with lightening, and camera location. The results from Tables 4.2 and 4.3 confirm our hypothesis. We observe a medium positive effect size (a high level) of arousal for drama, action, romance, and adventure and a large positive effect for horror. Moreover, we identify large positive and negative effects of valence for documentary and horror movies, respectively.

Slow movements of cameras, lightening, shadowing, and playing music in the background during subtle highlights H2 do not evoke strong emotional reactions among spectators. A medium negative effect size of arousal for action movies is observed. Furthermore, a medium positive effect size of valence is only reported for horrors, unlike action and romance movies.

Following the main characters' development and tensions among them (character development highlights H3) can affect the emotional and physiological states of spectators. We find a medium positive effect size of arousal for comedy, adventure, and documentary movies and a large positive effect size for horror and animation movies. Also, we observe a high level of negative valence for animations (medium negative effect), and thriller, romance, and horror movies (large negative effects), as illustrated in Tables 4.2 and 4.3.

Table 4.2 The weighted average effect size (fixed-effect model) of arousal during aesthetic highlights calculated per movie genre [141].

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	0.73	0.49	0.22	0.57	0.34	0.53	0.57	-0.30	1.06
H2	-0.02	-0.18	-0.13	-0.58	-0.46	-0.38	0.11	-0.49	-0.42
H3	0.01	0.92	0.14	0.11	0.65	-0.04	0.56	0.72	1.03
H4	-0.74	0.10	0.25	-0.13	0.92	-0.26	0.15	-	-0.27
H5	0.32	0.78	0.62	-0.01	0.76	-0.16	0.19	-0.19	0.65

Table 4.3 The weighted average effect size (fixed-effect model) of valence during aesthetic highlights calculated per movie genre [141].

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	-0.10	0.14	-1.04	0.19	0.13	-0.13	-0.48	1.20	-1.13
H2	-0.15	-0.10	-0.38	-0.75	-0.28	-0.71	0.08	0.23	0.60
H3	-0.16	-0.75	-0.80	-0.20	0.27	-0.92	0.11	0.45	-1.01
H4	0.07	-0.55	0.41	-0.02	-0.01	0.35	-0.43	-	0.39
H5	0.07	-0.17	-1.02	-0.40	0.56	-0.01	-0.11	-0.69	-0.48

Dialogues among main characters (highlights H4) only for some specific movie genres are able to elicit emotions in spectators. We find a low level of arousal in dramas (medium negative effect size) and a high level of arousal in comedies (large positive effect size) as well as a medium negative effect of valence for animation movies in Tables 4.2 and 4.3. We infer that dialogues carry the emotional tone of the genre. There are a low level of arousal (sad) for dramas and a high level of arousal (joy) for comedies. The long duration of dialogues could cause that spectators' emotions fade over time. Also, the main character are frequently ambiguous movie characters who could stimulate different reactions across the audience. That is why, we can observe in Tables 4.2 and 4.3 that directions of effects vary from one movie genre to another.

Theme development highlights H5 incompletely overlap with the other types of aesthetic highlights, such as spectacular highlights H1 and character development highlights H3. The development of a specific theme is often conjugated with the emotion development of main characters as their responses to dramatic events are presented in a sublime manner. Also, we observe a medium positive effect size of arousal for the following movies genres: animation, thriller, comedy, and horror, as presented in Table 4.2. In terms of valence, we remark a medium positive effect size for comedies and a medium and large negative effect for thriller and documentary movies, as shown in Table 4.3.

4.3.2 Dependencies between synchronization measures

In order to find dependencies between different approaches to synchronization and evaluate their detection performance, we selected the following synchronization measures: Nonlinear Interdependence (NI), Dynamic Time Warping (DTW), Periodicity Score (PS), Shape Distribution Distance (SDD), S-estimators with different covariance matrices, such as CORrelation (S-COR), Phase Locking Value (S-PLV), Windowed Mutual Information (S-WMI), Heat Kernel (S-HK), and Diffusion Map (S-DM). These synchronization measures represents pairwise, group, and overall approaches to synchronization estimation. The values of all the mentioned synchronization measures were computed for each time window. We calculated the CC between those measures to gain insight into the dependencies between them [55, 96].

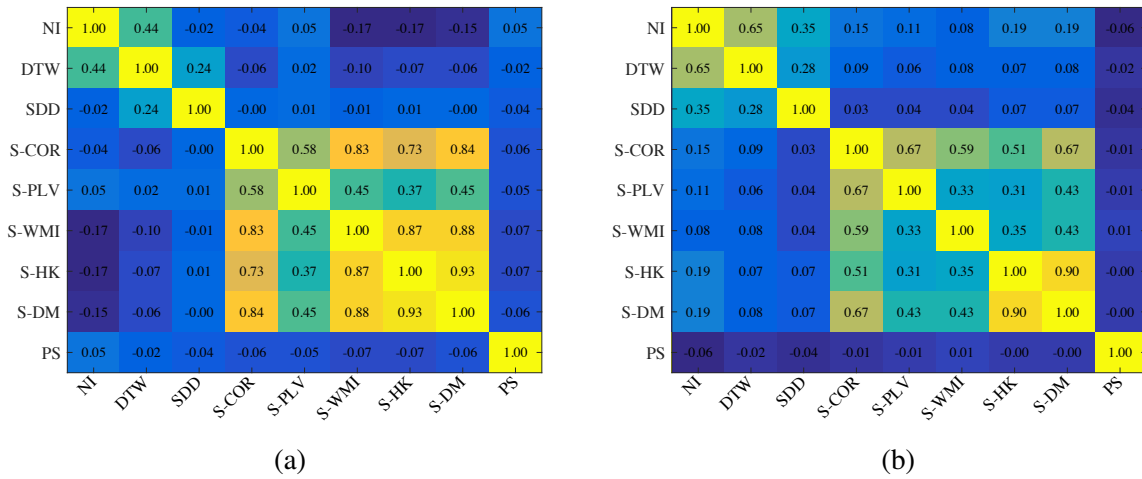


Fig. 4.3 The weighted average CC effect size between the synchronization measures (yellow and purple color indicate strong correlation and anti-correlation, respectively). The synchronization measures are computed over spectators': (a) EDA signals, (b) ACC signals [141].

A statistical analysis requires to weight all CCs between all pairs of the synchronization measures over the different movies from the C. LIRIS-ACCEDE database. To integrate the results and obtain the weighted average effect size of CC, we utilized a fixed-effect model [21]. To interpret the practical significance of a weighted average effect size for CC, we assume that values around 0.1, 0.3, and 0.5 can be interpreted as small, medium, and large effect sizes, respectively [43]. The values of thresholds are different in comparison to the standardized mean difference effect size (see Section 4.3.1).

As seen in Figure 4.3, we find that some synchronization measures are strongly correlated with each other independently of the processed modality (EDA and ACC signals). Taking

into account the values of the a weighted average effect size for CC, we can distinguish three families of synchronization measures: pairwise, group, and overall measures. It becomes clear that all the S-estimators with the different estimators of the covariance matrix are dependent on one another (medium and large positive effect). Furthermore, we can emphasize the strong relations (small, medium, and large positive effect) between all three pairwise synchronization measures: the NI, the DTW, and the SDD. Interestingly, the PS measure seems to be mutually uncorrelated with the other measures. These results are in line with the other studies on synchronization applied to electroencephalograph signals for early diagnosis of Alzheimer's disease [55]. We find as well that some measures (pairwise measures or S-estimators) are strongly correlated or anti-correlated.

4.3.3 Aesthetic highlight detection

In this section we provide the results of aesthetic highlight detection per movie genre using the different approaches to synchronization estimation which are described in Section 4.2. We detected 5 categories of aesthetic highlights (H1, H2, H3, H4, and H5 are defined in Section 3.3) based on the level of the estimated synchronization. If the value of a synchronization measure for a given sliding window is higher (lower) than a changing threshold, we assign the time window to a highlight class.

Since the presented synchronization measures capture different dependencies among signals, we decided to take the advantage of it. We followed a feature fusion approach and combined all the synchronization measures at a given time into one vector. Then, we used clustering based Gaussian mixture models to compute a probability of belonging to the highlight class (resp. non-highlight) for each window. If the probability for a given sliding window is higher (resp. lower) than a changing threshold, we assign the time window to the highlight class. We name it Combining Multiple Measures and Clustering (CMMC) approach. Identified labels are compared to the collected annotations (ground truth) and the true positive and false positive ratio are calculated to obtain ROC curves.

In order to investigate the statistical significance of the results, we used the following validation. Firstly, we computed AUC to evaluate the overall performance of our system as well as the discriminative power of synchronization measures. Furthermore, we referred the performance of our system to the performance of a random classifier ($AUC = 0.5$). Secondly, the synchronization measures that did not perform randomly were placed in rank order for each movie genre. Thirdly, multiple comparisons were made to find groups of measures that perform significantly better than others, such as 1st (the highest performance), 2nd and 3rd group of measures.

Table 4.4 Performance (AUC) of our highlight detection system evaluated per category of aesthetic highlights and movie genre, different synchronization measures applied to EDA signals of spectators [141].

Gen. \H	H1	H2	H3	H4	H5
Drama	1. DTW (0.57) [†] CMMC (0.56) [†]	1. CMMC (0.56) [*] NI (0.55) [*]	1. SDD (0.58) [‡] NI (0.57) [‡]	1. CMMC (0.61) [‡] SDD (0.59) [‡] DTW (0.58) [‡] S-WMI (0.56) [†] S-HK (0.55) [*] NI (0.55) [*] S-DM (0.55) [*]	1. NI (0.58) [‡] DTW (0.57) [‡] CMMC (0.56) [†]
Animat.	1.S-DM (0.59) [‡] S-HK (0.59) [‡] S-WMI (0.58) [‡] S-COR (0.55) [*]	1.* DTW (0.70) [‡]	1.‡SDD (0.72) [‡]	1.‡ CMMC (0.84) [‡]	1. NI (0.59) [‡] DTW (0.59) [‡]
Thrill.	1. SDD (0.62) [‡]	1. SDD (0.62) [‡] CMMC (0.62) [‡] DTW (0.61) [†] S-HK (0.60) [†] S-WMI (0.57) [*] S-DM (0.57) [*]	1. S-DM (0.70) [‡] S-HK (0.68) [‡] S-WMI (0.66) [‡] CMMC (0.65) [‡] DTW (0.64) [‡] S-COR (0.63) [†] S-PLV (0.59) [*]	SDD (0.74) [‡]	1. S-WMI (0.60) [†] S-HK (0.59) [†] S-COR (0.58) [*] S-DM (0.58) [*] S-PLV (0.58) [*]
Action	1.S-WMI (0.59) [‡] NI (0.58) [†] S-HK (0.57) [†] S-DM (0.56) [*] DTW (0.56) [*] CMMC (0.56) [*]	1. S-HK (0.58) [*] NI (0.58) [*]	1. DTW (0.57) [†] CMMC (0.57) [†] SDD (0.55) [*]	1. SDD (0.61) [‡]	1. DTW (0.58) [†] CMMC (0.58) [†] NI (0.56) [*]
Comed.	1. DTW (0.63) [‡] CMMC (0.63) [‡]	1. DTW (0.58) [‡] CMMC (0.56) [†] S-DM (0.54) [*]	1.*SDD (0.62) [‡]	1. S-HK (0.56) [‡] S-WMI (0.56) [‡] SDD (0.56) [‡] S-DM (0.54) [†] S-PLV (0.54) [*]	1.†SDD (0.63) [‡]
Roman.	1.‡SDD (0.82) [‡]	1. DTW (0.60) [‡] CMMC (0.59) [†] NI (0.57) [*]	1. DTW (0.65) [‡] CMMC (0.62) [‡] NI (0.60) [‡] SDD (0.57) [*]	1.‡DTW (0.77) [‡] CMMC (0.77) [‡]	1. NI (0.60) [‡] SDD (0.56) [*] S-COR (0.56) [*]
Advent.	1.*DTW (0.70) [‡] CMMC (0.70) [‡]	1. DTW (0.58) [†] CMMC (0.58) [†] SDD (0.57) [†] NI (0.57) [*]	1. SDD (0.64) [‡] CMMC (0.62) [‡] DTW (0.61) [‡]	1. SDD (0.60) [‡] NI (0.56) [†]	1.*SDD (0.67) [‡]
Docum.	1. CMMC (0.83) [†] DTW (0.75) [*] S-HK (0.72) [*] S-WMI (0.71) [*]	-	1. CMMC (0.97) [‡] SDD (0.83) [†] DTW (0.81) [†] NI (0.78) [*]	-	Any Measures
Horror	1.*DTW (0.69) [‡] CMMC (0.69) [‡]	1. SDD (0.67) [‡] CMMC (0.62) [†] DTW (0.62) [*]	1.‡DTW (0.75) [‡] CMMC (0.74) [‡]	1. SDD (0.61) [‡] NI (0.59) [‡] S-HK (0.57) [†] S-DM (0.57) [†] S-WMI (0.56) [†]	1. DTW (0.60) [‡] CMMC (0.60) [‡] NI (0.55) [†] S-HK (0.55) [*] S-WMI (0.54) [*] S-DM (0.54) [*]

* stand for p -value < 0.05 , † for p -value < 0.01 and ‡ for p -value < 0.001 . We report p -value when we refer the performance of a measure to a random classifier (upper index of AUC performance) and when we find the groups of synchronization measures in the ranking significantly different in terms of performance (the upper index of an ordinal number of measure groups).

Table 4.5 Performance (AUC) of our highlight detection system evaluated per category of aesthetic highlights and movie genre, different synchronization measures applied to ACC signals of spectators [141].

Gen. \H	H1	H2	H3	H4	H5
Drama	Any Measures	1.*SDD (0.63) [‡]	1. CMMC (0.57) [‡] DTW (0.56) [†] NI (0.56) [†]	1.*DTW (0.62) [‡] CMMC (0.59) [‡]	1.*NI (0.61) [‡]
Animat.	1. SDD (0.58) [‡] NI (0.55) [*] CMMC (0.55) [*]	1.*SDD (0.74) [‡]	1. SDD (0.60) [‡] NI (0.57) [†]	1. DTW (0.68) [‡] CMMC (0.67) [‡] SDD (0.64) [‡]	1. NI (0.64) [‡] DTW (0.62) [‡] SDD (0.61) [‡] CMMC (0.59) [‡]
Thrill.	1. MCCM (0.61) [†] DTW (0.58) [*]	Any Measures	1. SDD (0.59) [*] S-WMI (0.59) [*]	1.*CMMC (0.70) [‡] DTW (0.68) [‡]	1. SDD (0.66) [‡] DTW (0.65) [‡] CMMC (0.64) [‡] NI (0.59) [†]
Action	1. SDD (0.63) [‡] S-DM (0.58) [†] S-HK (0.57) [†] S-COR (0.56) [*] S-WMI (0.56) [*]	1. S-WMI (0.59) [*] CMMC (0.58) [*]	1. DTW (0.55) [*]	1. DTW (0.55) [*]	1. SDD (0.60) [‡]
Comed.	1. SDD (0.61) [‡] CMMC (0.58) [‡] NI (0.57) [†]	Any Measures	1. CMMC (0.56) [‡] NI (0.55) [†] SDD (0.53) [*]	1. CMMC (0.56) [‡]	1. [‡] SDD (0.63) [‡]
Roman.	1.*SDD (0.70) [‡]	1. DTW (0.63) [‡] CMMC (0.62) [‡] SDD (0.59) [†] NI (0.57) [*]	1. NI (0.63) [‡] DTW (0.59) [†]	1. S-COR (0.59) [*]	1. DTW (0.60) [‡] CMMC (0.58) [†] NI (0.56) [*]
Advent.	1. SDD (0.58) [†] CMMC (0.56) [*]	Any Measures	1. DTW (0.63) [‡] NI (0.61) [‡] SDD (0.58) [†] CMMC (0.58) [†]	1. DTW (0.63) [‡] CMMC (0.63) [‡] SDD (0.60) [‡]	1. S-WMI (0.55) [*]
Docum.	1. SDD (0.81) [†]	-	1. SDD (0.78) [*] NI (0.77) [*] S-PLV(0.74) [*]	-	1. SDD (0.95) [‡]
Horror	1. [‡] SDD (0.62) [‡]	1. DTW (0.64) [‡] SDD (0.64) [†] CMMC (0.64) [†]	1. [†] SDD (0.63) [‡]	1. DTW (0.57) [‡] NI (0.57) [†] CMMC (0.55) [*]	1. SDD (0.58) [‡]

* stand for p -value < 0.05 , † for p -value < 0.01 and ‡ for p -value < 0.001 . We report p -value when we refer the performance of a measure to a random classifier (upper index of AUC performance) and when we find the groups of synchronization measures in the ranking significantly different in terms of performance (the upper index of an ordinal number of the measure groups).

All the statistical comparisons were made using the two sided Bradley test at 0.05 significance level [22]. Low p-values indicate that there are large differences in the performance of the synchronization measures.

Tables 4.4 and 4.5 present the detection performance (AUC) of all the synchronization measures applied to spectators' physiological and behavioural signals. We only report the first group of synchronization measures that reached significantly the best performance for the given category of aesthetic highlights and movie genre.

In Table 4.4, the results illustrate the discriminative power of the synchronization measures to detect aesthetic highlights in movies based on spectators' EDA signals. In general, we observe that the pairwise synchronization measures obtain the best performance in comparison with the group or overall approaches. The DTW measure achieves the highest of performance for the following movie genre: animation, action, romance, documentary, and horror. The SDD measure reaches the best results for thriller, comedy, romance, and adventure movies. Also, the NI measure could indicate these highlights with the highest performance in drama, action, and romance movies. Moreover, the pairwise synchronization measures appear to have also the most discriminative power for detecting the different categories of aesthetic highlights. The DTW measure can be used to detect highlights H1, H2, and H3 unlike the DDS measure that indicates highlights H3 and H4 with the best performance. Besides, the NI measure can be applied to predict highlights H5.

Table 4.5 presents the detection performance, when synchronization measures are applied to ACC signals. We infer that the pairwise synchronization measures, especially the SDD, reach the best performance per movie genre and aesthetic highlight category. Moreover, the SDD measure obtains the best results for the movie genres: animation, action, comedy, adventure, documentary, and horror, as well as, in terms of highlight type: highlights H1, H2, H3, and H5. The DTW measure performs the best for drama, thriller, action, and romance movies as well as for highlights H4. The NI can be used interchangeably with the DTW to detect these highlights in specific movie genres, such as drama or romance. Also, it can replace the SDD to identify highlights H3. Detection of highlights H4 in animations and comedies only benefits from the basic aggregation of multiple synchronization measures applied to spectators' EDA and ACC signals (CMMC approach). Generally, combining synchronization measures does not improve performance of aesthetic highlight detection.

4.4 Discussion and conclusions

We extended our primary experiment [113] that aimed to understand spectators' physiological and behavioural reactions to movie aesthetic highlights. In this thesis, aesthetic highlight

detection was performed on the C. LIRIS-ACCEDE database that contains 30 movies from 9 movie genres. Regarding our **first research question** (RQ1) we discovered that these highlights evoke some amounts of emotions (arousal and valence intensity) that is strictly related to movie genres. Moreover, we proposed the architecture of an supervised detection system that is able to detect aesthetic highlights in movies based on spectators' physiological and behavioural reactions. We investigated which approach to synchronization estimation, such as pairwise, group, and overall measures obtains the best performance. In response to our **second research question** (RQ2), the results that we obtained prove that the level of synchronization among spectators' EDA and ACC signals in a social setting has a discriminative power to detect the different categories of aesthetic highlights independently of movie genre and recorded modalities. Nevertheless, a general statement cannot be made for different movie genres because of the small number of movies per genre available. Furthermore, we infer from our analysis that all the pairwise synchronization measures are correlated with each other. Also, that is the case for all the overall synchronization measures. To study synchronization, we find that it is enough to evaluate a few measures derived from the different families of synchronization measures instead of using all of them.

Overall, we observe that the pairwise synchronization measures, such as the SDD (Shape Distribution Distance) and the DTW (Dynamic Time Warping) best perform aesthetic highlight detection in movies independently of movie genre and highlight type, responding to the second part of RQ2. The group and overall estimation of synchronization perform at the lowest level. Also, the choice of a covariance matrix estimator, such as the min correlation distance, correlation, phase locking value, windowed mutual information, heat kernel, and diffusion map does not influence performance. When rapid physiological and behavioural reactions are evoked, all the pairwise synchronization measures (the SDD and the NI) seem to take the advantage of including information on neighboring time windows unlike the estimation of the covariance matrix. Moreover, the DTW is able to average the temporal reactions of spectators which vary in dynamics. These features of the estimation methods allow them to suppress the oscillations of the values from one time window to another. In addition, we suppose that considering all spectators signals like one dynamic system suffers from rapid changes of social interactions among spectators through the whole movie, and may result in unstable behaviours of the dynamical system. Analysis of synchronization at the level of pairs could benefit from uncovering stable pairs of spectators through the majority part of a movie.

EDA measurements appear to be more indicative of aesthetic highlights in a social context compared to ACC measurements. The main reason can be that aesthetic experience is associated with a high level of arousal that is depicted in spectators' physiological reactions.

That is coherent with our findings on the annotated scenes contain a large amount of emotions (high level of arousal and valence) for the whole C. LIRIS-ACCEDE database. Spontaneous rapid behavioural reactions could be expected to be evoked when spectators are exposed to very intensive stimuli, e.g., spectacular killing people in a horror movie.

Generally, we observe that pairwise, group, and overall synchronization measures are able to estimate synchronization among spectators' physiological signals when they are exposed to different aesthetic highlights that elicit a high level of arousal and valence, e.g., romance, action, adventure, horror movies, etc. This is not the case for the estimation of synchronization based on ACC measurements, only pairwise synchronization measures are capable of estimating the level of synchronization among behavioural responses (ACC signals) of spectators.

Combining multiple synchronization measures into one vector (CMMC approach) does not significantly improve the performance of aesthetic highlight detection. There is a need to study new strategies of multiple synchronization measure fusion since they are defined in different manners and measure different nonlinear dependencies between signals. This can be considered as one of the future directions of research on synchronization measures.

The main limitation of our work is the amount of available annotated data, the feasibility of running a large scale experiment in a cinema theater, and using unobstructive and reliable sensors. In our studies we uncover that the estimation of synchronization among spectators from their physiological signals results in better performance of highlight detection than from their behavioural (ACC) signals. However, this conclusion can be biased by the placement of sensors. The sensors were attached to spectators' hands when the experiment was conducted. Generally, spectators do not often make limb movements when they watch a movie.

Future work includes collecting more multimodal data to propose general architecture of a detection system. That allows us to apply more complex synchronization measures between different modalities. In the future, we will possibly have access to cost-effective sensors that are capable of capturing currently unavailable modalities, such as audio-video recording of movie audiences in a darkened cinema theater and more spectators' physiological and behavioural signals. A comprehensive approach to understanding aesthetic experience also requires to explore movie content combined with spectators' reactions. The integration of audio-visual movie attributes with spectators' physiological and behavioural signals can be beneficial for aesthetic highlight detection and the understanding of film aesthetic experience.

Chapter 5

Studying the relationship between induced and perceived emotions of movie audiences

In this Chapter, we mainly study the relationship between perceived and induced emotions of movie audiences. The former correspond to perceiving properties of affective movie content while the latter are associated with spectators' emotional reactions to affective movie content. In addition, we investigate different machine learning models to predict movie induced emotions from movie content based features as well as physiological and behavioural (ACC signals) responses of movie audiences. Understanding how affective movie content evokes emotions in movies can make significant contribution to emotion-based content delivery, video summarization and indexing. Also, our work on perceived and induced emotions may help the movie industry to design movie sets and increase spectators' engagement with stories of movie main characters.

It has been argued that the perceived emotions of movie content can influence the induced emotional responses of audiences, for example, by evoking empathy [184]. That suggests a positive correlation between perceived and induced emotions. Nevertheless, Baveye et al. [11] wondered whether intended emotions of moviemakers are always consistent with emotions that evoked in movie audiences or not. To the best of our knowledge, there has been no previous research quantitatively investigating the relationship between perceived and induced emotions of movie audiences. In this thesis, we suppose that there is a relationship between movie audiences' induced and perceived emotions. However, we assume that this relationship is more complex than a positive correlation due to the effect of watching full-length movies. We have to take into account the fact that consecutive scenes are dependent on each other and movie form and content are presented over time with respect to intentions

of moviemakers. This can be one cause of the difference between what spectators perceive and feel.

To investigate the relationship between perceived and induced emotions of movie audiences, we address the following research questions (see a full list of research questions in Section 1.5):

1. Are perceived emotions of the movie content and induced emotions in movie audience always consistent? (RQ3)
2. How can we improve recognition performance of induced emotions in movie audience? (RQ4)
 - Are there other features beyond the audio-visual movie content than can contribute to induced emotion recognition?
 - Are perceived emotions discriminative for induced emotion recognition?
 - Do recognition models benefit from including temporal information and multi-modal signals?

We emphasize the contributions of our work below, highlighting the novelty compared to the state of the art research:

- We carry out the first quantitative analysis of the relationship between perceived emotions and induced emotions of movie audiences as well as aesthetic highlights.
- To the best of our knowledge, we are the first to use movie perceived emotion annotations as affective cues to predict induced emotions.
- In our experiments, we show how performance of induced emotion recognition can be improved by including temporal context, and multimodal fusion that incorporates audio-video features, lexical features, perceived emotion and aesthetic highlight annotations as well as spectators' physiological and behavioural reactions to movie content.

5.1 Multimodal feature extraction

To answer our fourth research question (RQ4), we attempt to use multimodal information that describe affective movie content and spectators' physiological and behavioural reactions to improve induced emotion recognition. Besides audio-visual features of movie content, we extracted high level affective features, such as lexical features in movie dialogues, aesthetic movie highlight annotations, and perceived emotion annotations to describe affective movie

content. The lexical features characterize emotions in dialogues expressed by movie main characters while the aesthetic highlight and perceived emotion annotations describe the aesthetic and affective movie content. In addition, we included statistical descriptors of spectators' physiological and behavioural signals to take into account the fact that induced emotions are encoded in movie audiences' reactions.

The induced emotion annotations in the arousal-valence space from the C. LIRIS-ACCEDE database are provided at the second level for each movie. To include a suitable amount of temporal information on spectators' physiological and behavioural reactions as well as audio-video movie content with affective cues, we used a 5 second sliding window with a 4 second overlap between neighbouring windows to extract all features.

5.1.1 Movie audience reaction based features

To take into account the fact that induced emotions are subjective, we included two audience reaction based features, namely statistical features of physiological and behavioural signals. We assume that each person within a movie audience can display similar behaviours and have similar physiological responses when they are watching a movie together because [112, 114, 115, 139, 140]:

- the aesthetic and emotional design of movie scenes are made by filmmakers to evoke specific emotional reactions and aesthetic experiences (e.g., adding special effects and music in the background, empathy and compassion toward a main character, etc.).
- watching a movie together causes movie audience's affective reactions to be synchronized through emotional contagion.

EDA and ACC signals of spectators were filtered by a third order low-pass Butterworth filter with cut-off frequency at 0.3 Hz to remove noise before feature extraction. The statistical features are mean, median, standard deviation, minimum and maximum value as well as minimum and maximum ratio over sliding windows of a signal and its first and second derivatives. In particular, the statistical features were computed over sliding windows of EDA and ACC measurements collected from sensors attached to the spectators' limbs. Then, the same features of each spectator were concatenated into one feature vector. These feature vectors describe changes and their dynamics in physiological and behavioural responses of all spectators while watching movies. It is important to mention that these physiological and behavioural data were collected from a different group of movie experiment participants than those who annotated induced emotions [12, 125].

5.1.2 Movie content based features

Audio-Visual features

We extracted features from audio-visual movie content by means of the OpenSMILE toolkit [69]. In fact, we computed 1582 InterSpeech2010 Paralinguistic Challenge Low-Level Descriptor audio features [165] and 1793 visual features [68] for each sliding window. The visual features are histograms of Local Binary Pattern, HSV (hue, saturation, and value) color representation, and optical flows of each image region. These audio-visual features are benchmark features used in various emotion recognition tasks [153].

Dimensionality reduction was required due to the small number of available instances for model training. This results in less model parameters to tune. To reduce the number of features, we used the ReliefF algorithm and ranked the discriminative power of each feature for emotion recognition by means of performing regression with 20 nearest neighbours [157]. To do it, we created ReliefF based feature ranking over the remaining 22 movies of the C. LIRIS-ACCEDE database. These movies are different from the 8 movies on which we conducted emotion recognition experiments. This guarantees that selected audio-visual features are relevant to emotion recognition and testing instances were not included during feature selection since the ReliefF is a supervised feature selection algorithm.

We selected the most discriminative 100 audio and visual feature sets for arousal and valence prediction. This reduced the number of model parameters and balanced the dimensionality between multimodal features. As a result, overfitting of recognition models was prevented. In addition, we tested different feature engineering approaches, such as selecting more features or performing feature selection on a combined audio-visual feature set. We also applied dimensionality reduction instead of feature selection. We used a linear principal component analysis, a nonlinear principal component analysis with a Gaussian kernel and diffusion maps [45]. In all the cases, the first 100 components were sufficient to describe 99% of the total data variance. However, these did not result in any significant performance improvement.

Lexical features

This subsection on lexical features is based on the work of Leimin Tian [142, 192]. It has been shown that lexical features are discriminative for speaker emotion recognition in spontaneous dialogues [190]. The lexical features DIS-NV and CSA features. The former are extracted from manual annotations of DIS-NVs in movie dialogues. The latter are crowd-sourced annotations (CSA) of arousal, valence, and power ratings of 13,915 English lemmas [190, 198]. To extract CSA features, we firstly removed stop words (commonly used

words such as "the", "and", "a", etc.) from the movie transcript. Then, we lemmatized the words (e.g., transform "beginning" to "begin") that remain by means of the Natural Language Toolkit [17]. These are a standard part of pre-processing in Natural Language Processing studies. To compute the feature values, we searched for the lemmas in each sliding window in the dictionary of Warriner et al. [198]. Each dictionary entry contained 63 statistics that were calculated over the collected arousal, valence, and power ratings. The statistics are means, standard deviations and the number of contributing ratings over all the raters and over 6 subsets of raters: male, female, older, younger, high education, and low education, resulting in 21 (3 statistics for the whole set of raters and its subsets) statistics for each emotion dimension. Sums of each of the 63 statistics for all the lemmas in the sliding window are the 63 lexical features.

The six DIS-NV features were computed as the total duration of manual annotations of each DIS-NV category, including the general lexicons (see Section 3.4.2) in each sliding window divided by the window length (5 seconds). We did not apply stop word removal or lemmatization for computing the DIS-NV features because these features are based on the duration of words.

Aesthetic movie highlights

The aesthetic movie highlights that are associated with the occurrences of meaningful scenes defined with respect to art form and content by film experts [112] can be considered as high level affective features characterizing aesthetic and affective movie content. Using aesthetic highlight annotations as features is supported by the fact that these highlights are designed by moviemakers to evoke specific emotional reactions (see Sections 1.2 and 4.3.1). These results show that aesthetic highlights in movies elicit a various range of emotions in spectators. Thus, their occurrence can indicate a specific emotion elicitation in order to improve induced emotion recognition. In general, there are two main categories defined: Form and Content, as shown in Figure 3.1. The former is split into 2 subcategories: Spectacular highlights (H1) and Subtle highlights (H2), while the latter is divided into 3 subcategories: Character development highlights (H3), Dialogue highlights (H4), and Theme development highlights (H5).

We also proposed to create highlights H6 that indicate the occurrences of any highlight categories that mentioned above. Then, we used the annotations of all these 6 highlight categories at the window level as the high-level knowledge-inspired affective features for induced emotion recognition. These features are more abstract than the audio-visual movie content features. Also, it is important to mention that aesthetic highlight features are sparse, because aesthetic highlights are rare events in movies (see Section 3.3).

Perceived emotions

The motivation to use perceived emotion annotations as as high level affective features is that we assume that there is a certain relationship between perceived and induced emotions of movie audiences. When we are able to extract features that describe perceived emotions of movie audiences, then we only have to find a model that maps them to induced emotions. Thus, the annotations of perceived emotions of movie audiences were used as the high-level affective features to recognize induced emotions. To do so, the scores in the arousal-valence-power space were averaged and then normalized to interval $[-1,1]$. Sliding windows were applied to the emotional scores to align them with the features of movie content and movie audience reactions. Dialogues are not very frequent events in movies (please Section 3.3) that is why these features are sparse like aesthetic highlight features.

5.2 Recognition models

In this section we detail LSTM-RNN models and their hierarchical architecture to fuse multimodal signals for induced emotion recognition. Also, SVR and DBN models are described as baseline emotion recognition models. We proposed the hierarchical architecture of LSTM-RNN models for fusion of multimodal information because we assume that there is a complex temporal relationship between induced and perceived emotions. This is why we extracted different sets of features that describe affective movie content as well as spectators' physiological and behavioural reactions. We selected LSTM-RNN models because of three reasons [143, 190]:

- LSTM-RNN models are able to learn long range dependencies between two time series and are able to capture temporal information. This is required because movies and spectators' reactions to movie content have sequential structures.
- LSTM-RNN models can learn a new representation of data. It is desired since multimodal information is encoded in many noisy features with different temporal dynamics.
- LSTM-RNN models allow multimodal features to be incorporated in different model layers. The hierarchical structure is designed based on both the temporal characteristics and the abstraction level of features.

However, it is important to mention that building a deep structure (multiple layers) of the LSTM-RNNs would require us to have access to massive labelled data. We compared our proposed LSTM-RNN models to SVR models that are the baseline emotion recognition models [143]. The big advantage of using SVR models is that a small number of training

instances is required to find their optimal parameters. However, these SVR models are not able to capture temporal information. Besides SVR models, we compared the proposed LSTM-RNN models to DBN models that are able to learn a new representation of data and complex dependencies between them [99]. Nevertheless, temporal information is omitted by the DBN models. Also, a large number of instances is needed to train these models properly.

5.2.1 Long short-term memory recurrent neural networks

Section 5.2.1 is based on the work of Leimin Tian [142, 192]. Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) are recurrent neural networks with multiple hidden layers. This structure allows LSTM-RNN models to capture temporal information. It has been shown that a 3 hidden layer hierarchical structure of LSTM-RNN models for fusion of multiple modalities improved emotion recognition in spoken dialogues [190]. Moreover, the LSTM-RNN model outperformed state of the art algorithms to classify voicing or silence with noise in movies [67].

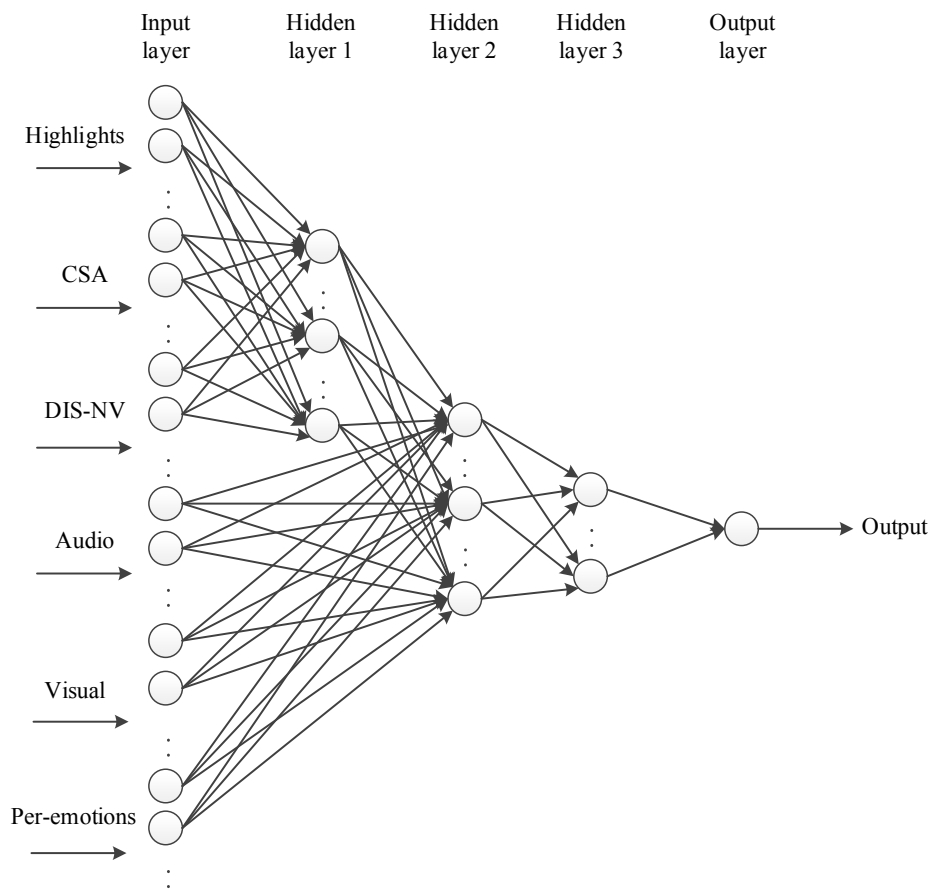


Fig. 5.1 Structure of HL model using movie based features [142, 192].

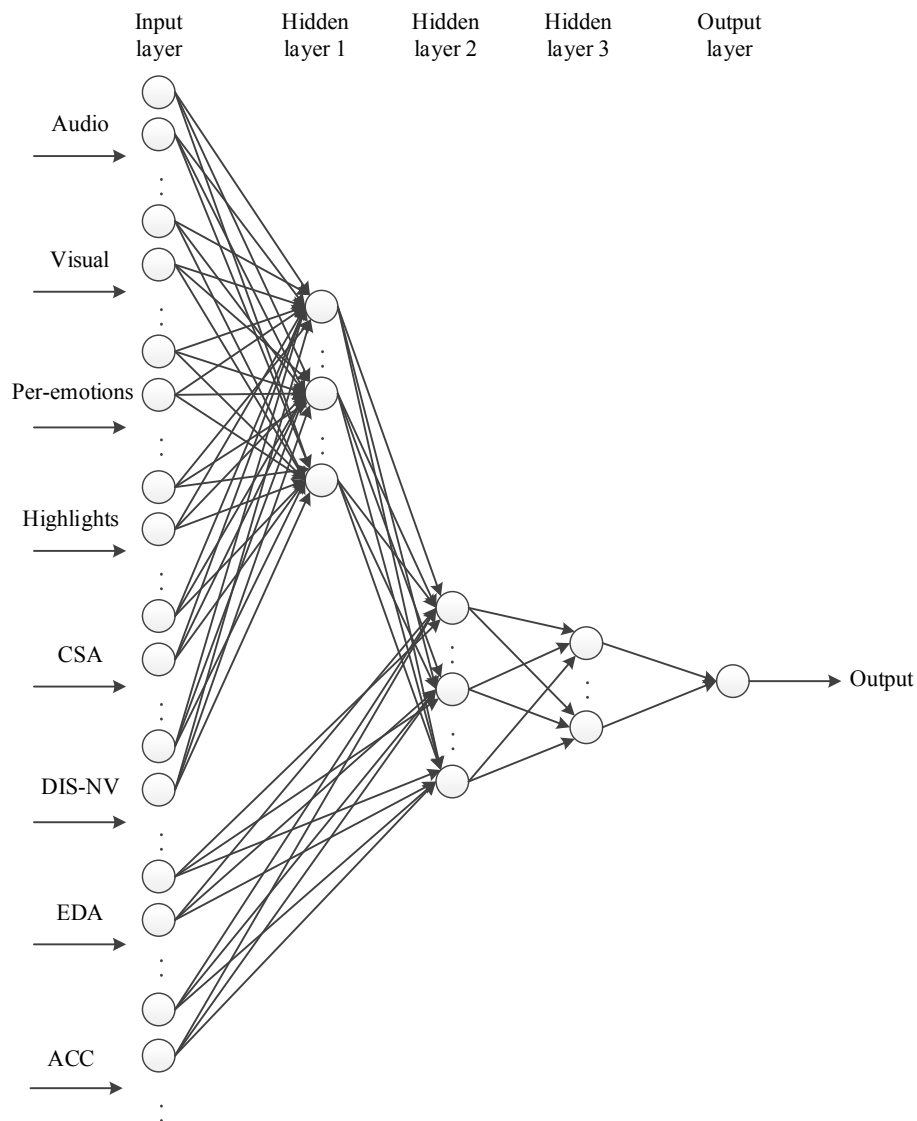


Fig. 5.2 Structure of HL model using all features [142, 192].

We built LSTM-RNN models using the Keras library for induced emotion recognition [38]. All the LSTM-RNN models had 3 hidden layers with 64 , 32, and 16 neuron units from bottom to top. This architecture was already applied to emotion recognition [190] with success. To avoid overfitting, we used dropout in the first hidden layer with a rate of 0.5 and set the maximum training iteration to 50 epochs with an early stopping tolerance of 10 epochs. The size of mini-batches is 10 due to computational efficiency of training. Other sizes that varied from 3 to 36 were tested. In fact, performance was not influenced by the size selection.

We used and evaluated three fusion strategies: Feature-Level (FL) fusion, Decision-Level (DL) fusion, and Hierarchical (HL) fusion for multimodal emotion recognition [190]. All multimodal features are concatenated in a vector before feeding recognition models, when the FL fusion is used. While applying the DL fusion, unimodal recognition models for each feature set are built and their outputs are incorporated in a decision making module that is another LSTM-RNN model. The HL fusion strategy incorporates different features in different levels of its hierarchy, e.g., aesthetic highlight and perceived emotion annotation based features with noise are incorporated in lower layers of the LSTM-RNN models, while more abstract features, e.g., audio and video features are incorporated in their higher layers.

All features are used at the input layer of the LSTM-RNN model for the multimodal FL fusion. Nevertheless, predictions of unimodal LSTM-RNN models are input to another LSTM-RNN model for the multimodal DL fusion. Furthermore, input neurons of low-level features are connected to the first hidden layer, while input neurons of high-level features are directly connected to the second hidden layer for the multimodal HL fusion.

We built multimodal models combining only movie content based features as well as movie content based features with spectators' reactions. As we can see in Figure 5.1, the former model uses the descriptors of audio-video content at a higher layer than noisy affective clues because in-domain knowledge was enhanced during feature selection of audio and visual features. The latter model uses features of physiological and behavioural signals at a higher layer than movie based features, as shown in Figure 5.2, because movie audiences' reactions are characterized by different dynamics of changes.

5.2.2 Deep belief networks

Deep Belief Networks (DBNs) improved emotion recognition performance [106]. It has been shown that two hidden layer DBNs are able to learn a new representation of audio-visual features, capturing complex non-linear dependencies between them. Also, these DBNs are capable of reducing the high dimensionality of the original audio-visual feature space. The structure of DBNs is a stack of multiple Restricted Boltzmann Machines (RBMs). Moreover, the RBMs have drawn increasing attention in current machine learning research because these stochastic graphical models have improved performance in many applications, such as speech recognition and emotion recognition [106, 179].

A basic Bernoulli-Bernoulli RBM (BBRBM) assumes that the input data comes from a binary distribution. This is a crucial limitation. Thus, a RBM assuming that the data are derived from a Gaussian distribution was proposed in [92]. In this paper we only used a Gaussian-Bernoulli RBM (GBRBM) that is a RBM which uses Gaussian distributions for the visible units and binary distributions for the hidden units [203]. Furthermore, a deep

belief network (DBN) is a stack of multiple RBMs. The hidden units of a learned RBM are used as the visible units of the following RBM. The DBNs are able to learn a high level representation from a large amount of unlabelled instances. Then, relatively small number of labelled data is required for the fine-tuning of the model.

We selected a GBRBM for the input layer with respect to the distributions of physiological and behavioural signals that are better fitted to the Gaussian distribution than the pseudo binary distribution. Other layers were BBRBMs. We learned the DBNs with only 2 hidden layers with 50 and 15 neuron units, respectively, as a result of the limited number of training instances available. The size of mini-batch is the number of features divided by 4 due to computational efficiency. The initial learning rate and its upper bound are set to 0.002 for pre-training and the weight-updating ratio is set to 0.1. The cross entropy is used as a loss function. We also applied gradient decent based supervised fine tuning with maximum 100 iterations to find optimal parameters for the whole DBNs. To avoid overfitting on the limited training set, we used a dropout with a ratio of 0.5 for each hidden layer.

5.2.3 Support vector regression

Support Vector Regression models have demonstrated high performance for affect prediction [5, 12, 13, 33]. In this work we used a nonlinear ν -support vector regression (SVR) with a Gaussian kernel as a baseline model for induced emotion recognition [30]. The optimal scaling parameter $\gamma \in \{2^3, \dots, 2^{-15}\}$ of the radial basis function, the optimal regularization parameter $C \in \{2^{-5}, \dots, 2^{15}\}$ and the optimal parameter $\nu \in (0, 1]$ that controls the number of support vectors were identified by grid search.

5.3 Experimental results

5.3.1 Perceived and induced emotions

In this section we answer our research question RQ3 on the relationship between perceived and induced emotions of movie audiences. Please take into account the fact that the induced emotions were annotated at the second level (annotations for each second of a movie) while the perceived emotions were annotated at the utterance-level (annotations for each utterance of movie dialogues). This implies that the perceived emotion annotations are generally longer than one second. To align the annotations, we calculated the mean values of induced arousal-valence ratings over each movie utterance. This provided us the utterance-level induced emotion annotation. Then, we independently calculated the CC between each pair of perceived and induced emotional dimensions for each movie. Finally, we used a fixed-effect

model [161] to analyze the CC between perceived and induced emotion dimensions described by CC values. To do so, we computed the weighted average of the CC over all 8 movies, as shown in Figure 5.3. To measure the statistical significance of the correlation between

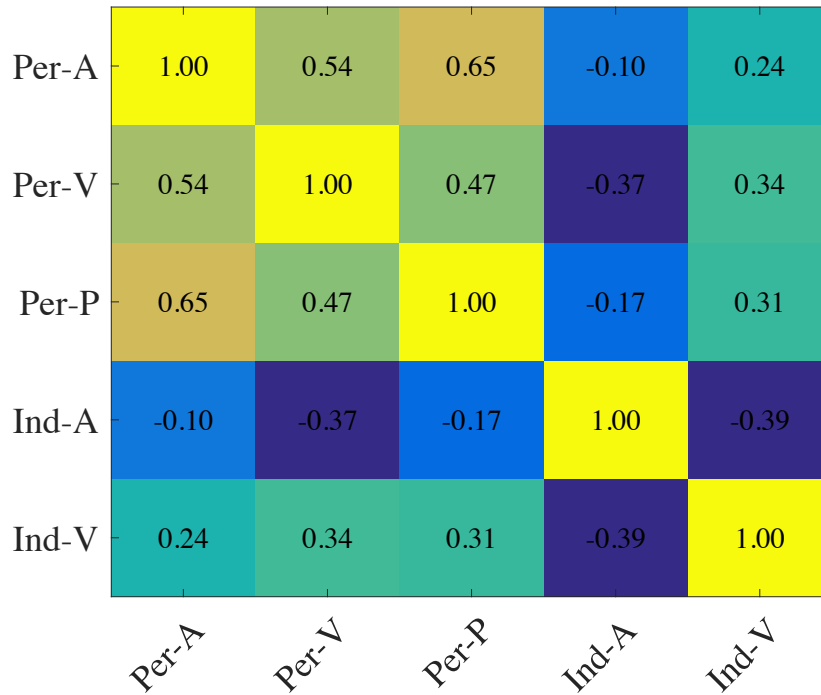


Fig. 5.3 The weighted average of the CC between perceived (Per) and induced (Ind) emotional dimensions of movie audience (yellow and purple color indicate strong correlation and anti-correlation, respectively) [142, 192].

induced and perceived emotional dimensions, we interpret absolute values of the weighted mean of CC values (absolute overall CC values) that are around 0.1, 0.3, and 0.5 as small, medium, and large effect sizes, respectively, following Cohen's model [42].

Consequently, we observe that perceived arousal, valence, and power dimensions are highly positively correlated (two large effects and one medium effect) with each other, as shown in Figure 5.3. Induced arousal and valence dimensions are moderately negatively correlated with each other. This suggests that the dependencies between the perceived dimensions and the dependencies between the induced dimensions are not the same. The negative correlation between induced arousal and valence dimensions is in line with previous research on induced emotions. This suggests that induced negative emotions may have the stronger effect than induced positive emotions regarding arousal.

Nevertheless, we can not make general conclusions due to the small values of the overall CC and the small sample size. Moreover, the induced valence dimension and all

the perceived emotion dimensions have moderately positive correlations while the induced arousal dimension and all the perceived emotion dimensions are negatively correlated at the weak or moderate level. In particular, perceived arousal and induced arousal are weakly negatively correlated. It seems that watching too many exciting, pleasant, and dominating scenes in movies may evoke boredom in movie audiences. Nevertheless, movie audiences can feel displeasure during watching movie scenes in which main characters are dominated by dramatic events.

These results suggest that there is the discrepancy between perceived and induced emotion annotations. Consequently, we show here a significant difference between perceived emotional movie content and felt emotions by movie audiences. Emotion induction and emotional responses of movie audiences can be influenced by many various factors other than the affective movie content, such as personality, life experience as well as movie and art preferences. Our analysis proves that the assumption that perceived and induced emotions of movie audiences are consistent is not entirely accurate and thus researchers have to take into account this result when designing experiments for affective content analysis research on movies.

5.3.2 Perceived and induced emotions vs. aesthetic highlights

In this section we respond to our research questions RQ3 and RQ4. We investigated the relationship between the intensity of induced and perceived emotions and the occurrences of aesthetic highlights. We considered the 8 movies from the C. LIRIS-ACCEDE database as a set of empirical experiments about the given topic. We related the intensity level of induced and perceived emotions of movie audiences with the occurrences of different aesthetic highlight categories in these movies (see Section 3.3). To do so, we calculated an effect size over individual movies. The effect size is the standardized mean difference that is defined as the difference between mean values of continuous emotion annotations of highlight and non-highlight intervals divided by their pooled standard deviation. Positive values indicate a higher level of induced/perceived emotions of highlight scenes in comparison with non-highlight scenes, whereas negative values of the effect size indicate a lower level.

To combine the effect sizes, statistical analysis requires the weighting of each effect size estimate as a function of its precision assuming a fixed-effect model [21]. In this thesis we follow Cohen's benchmarks for the practical significance of the weighted average effect size. We assume that the values around 0.2, 0.5, and 0.8 can be interpreted as the small, medium, and large effect sizes, respectively [42].

We report the weighted average effect size of induced/perceived emotional dimensions for the 8 movies from the C. LIRIS-ACCEDE database in Table 5.1. Strong emotional

Table 5.1 Dependencies between aesthetic highlights and perceived and induced emotions of movie audience (small, medium, and large magnitudes of the overall effect in bold) [142].

H	Per-A	Per-V	Per-P	Ind-A	Ind-V
H1	0.33	-0.22	-0.38	0.48	-0.26
H2	-0.03	-0.84	-0.52	0.17	0.07
H3	0.24	-0.18	0.06	0.15	0.22
H4	-0.20	-0.47	-0.02	0.09	-0.24
H5	0.17	-0.22	-0.24	0.29	0.29

reactions may be associated with the occurrences of spectacular highlights H1 in movies, such as adding special effects, changes in saturation of colors, lightening, and camera location. A small positive effect size of induced and perceived arousal dimension and a small negative effect size of induced and perceived valence dimension are observed for spectacular highlights H1. Moreover, a small negative effect size of perceived power dimension is found. It is important to point out that the directions of effects for both induced and perceived arousal/valence dimensions are only consistent during highlights H1.

Slow movements of cameras, lightening, shadowing, environmental noise, and playing music in the background during subtle highlights H2 are not expected to elicit strong emotional responses among movie audiences. Nevertheless, there are a large negative effect of perceived valence dimension and a medium negative effect of perceived power dimension for highlights H2.

The main characters' development and tensions among them that are included in character development highlights H3 could influence emotional and physiological states of movie audiences. We observe a small positive effect of perceived arousal and induced valence dimension.

Specific dialogues among main characters (highlights H4) can affect emotional and physiological states of movie audiences. We find a small negative effect of perceived and induced valence dimension as well as perceived arousal dimension. It is worth noting that the direction of the effect for perceived and induced valence dimension is the same. This means that emotions, such as anger, sadness, joy, and pleasures perceived from dialogues evoke similar emotional states in movie audiences, e.g., empathy toward the main characters.

Theme development highlights H5 partially overlap with other categories of aesthetic highlights, for example, spectacular highlights H1 and character development highlights H3. In particular, the development of a theme is often associated with some changes in emotional states of main characters as their reactions to dramatic events presented in a spectacular or sublime manner. We observe a small negative effect of perceived valence and perceived power

dimension. Also, we find a small positive effect of induced arousal and valence dimension. A related point to consider is the incoherence of the effect directions for perceived and induced valence dimension. It means that perceiving negative valence (unpleasantness) by movie audiences can evoke pleasure in them. Essentially, we find aesthetic highlights as high level aesthetic cues that include information on perceived and induced emotions of movie audiences regardless of the discrepancies between them.

5.3.3 Induced emotion recognition

Section 5.3.3 is based on our joint work with Leimin Tian [142, 192]. We propose an approach to recognize induced emotions of movie audiences from multimodal signals, answering the research question RQ4. We used the average arousal-valence scores over each window of length 5 seconds as the gold-standard induced emotion annotations. Also, we removed the end movie credits because the spectators started to touch and remove the wearable sensors. This introduced noise and outliers in the physiological and behavioural signals. Eventually, 7103 data instances were available for induced emotion recognition.

We performed leave-one-movie-out cross-validation and reported the unweighted average of MSE and the absolute value of CC and CCC for arousal (A) and valence (V) prediction. For example, A-MSE refers to the average MSE over leave-one-movie-out cross-validation for arousal prediction. The MSE and CC are the most commonly reported evaluation metrics in related work on emotion recognition (see Tables 2.1 and 2.2). A high value of the CC represents a strong linear relationship between the values of emotion predictions and annotations. This means that general value changes (increase/decrease trends) in both signals co-occur. A low value of the MSE implies that values of both signals are similar to each other and corresponds to the high quality training of a predictive model. The CCC combines the CC with the square difference between the mean of the two compared time series, which makes it sensitive to bias and scaling factors [104]. This measure is commonly applied to multiple unambiguous annotation predictions, for example, induced emotions [104] (see Figure 3.7). A large value of the CCC describes a high agreement between values of predictions and annotations. This means that prediction and annotations values are similar to each other and general trend changes in both signals are the same.

We used the following validation to investigate the statistical significance of the results. In order to show that our models performed better than a random prediction model, we generated arousal and valence prediction scores at random. Then, we compared predictions of two models with highest CC or CCC values for each experiment to random predictions of arousal and valence scores, respectively. Finally, we compared the predictions of these pairs of models that did not perform randomly (e.g., the SVM models fed by EDA and

audio features, respectively, for arousal prediction). All the statistical comparisons were made by means of two-sample Wilcoxon test at 0.05 significance level. When we report results for each experiment, numbers in bold italics indicate significantly best performance with (p-value $\ll 0.0001$) and numbers in bold indicate significantly best performance with p-value < 0.05 .

Influence of temporal information on induced emotion recognition performance

Previous work on emotion recognition has shown that human emotions are context dependent and do not significantly evolve over short time intervals [153]. Nevertheless, the suitable amount of temporal context for predicting movie induced emotions is unknown and is task-dependent.

Firstly, we had to determine the suitable amount of temporal context to predict induced emotions. To do so, we tested LSTM-RNN models with different time steps. In particular, we used statistical features of EDA measurements to feed the LSTM-RNN models. These features could capture dynamic changes in the movie audience's physiological and behavioural reactions [114]. Our experiments [192, 142] show that including features for the past 3 time steps gives better recognition performance than shorter or longer time steps. Consequently, all our LSTM-RNN models used a time step of 3 in this thesis. This means that all the LSTM-RNN models include 8 seconds of temporal context because all features were extracted over a 5 second sliding window with a 4 second overlap.

Unimodal induced emotion recognition

The results of our unimodal induced emotion recognition experiments are shown in Table 5.2 in which we report the average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction. As we can see for arousal and valence prediction, the SVR model achieved the best performance on physiological features and perceived emotion features measured by the CC and CCC, respectively. This means that physiological signals and perceived emotions provide discriminative information on induced emotions.

Moreover, the SVM is able to capture the dependencies between changes in physiology and emotional states of spectators. As shown in Table 5.2, the SVM can only predict an increase or decrease of arousal and valence intensity from EDA signals with respect to the CC values. Besides, the values of the CCC suggest that the same SVR model is able to predict induced emotions from perceived emotion annotations in terms of upward and downward

trends and values as well. Nevertheless, the large values of MSE suggest that there is a need to improve learning of this model for these emotion recognition tasks.

Table 5.2 Performance of unimodal induced emotion recognition using SVR, DBN, and LSTM-RNN models [142].

Features	A-MSE	A-CC	A-CCC	V-MSE	V-CC	V-CCC
SVR model						
EDA	0.260	0.229	0.002	0.326	0.216	0.003
ACC	0.259	0.168	0.001	0.325	0.109	0.001
Audio	0.260	0.185	0.002	0.325	0.133	0.001
Visual	0.260	0.154	0.002	0.326	0.173	0.002
CSA	0.399	0.075	0.006	1.575	0.058	0.023
DIS-NV	1.924	0.060	0.016	1.225	0.062	0.020
Highlights	0.258	0.134	0.008	0.325	0.093	0.000
Per-emotions	0.709	0.138	0.104	0.743	0.090	0.056
DBN model						
EDA	0.065	0.074	0.008	0.082	0.144	0.016
ACC	0.064	0.112	0.009	0.081	0.086	0.008
Audio	0.066	0.217	0.026	0.081	0.194	0.022
Visual	0.065	0.111	0.010	0.082	0.148	0.014
CSA	0.063	0.016	0.000	0.076	0.052	0.003
DIS-NV	0.065	0.059	0.002	0.081	0.071	0.002
Highlights	0.065	0.143	0.019	0.084	0.148	0.027
Per-emotions	0.064	0.102	0.008	0.079	0.077	0.008
LSTM-RNN model						
EDA	0.047	0.190	0.044	0.066	0.432	0.072
ACC	0.049	0.183	0.082	0.064	0.129	0.054
Audio	0.054	0.218	0.055	0.069	0.134	0.033
Visual	0.060	0.126	0.018	0.090	0.152	0.025
CSA	0.050	0.085	0.029	0.071	0.060	0.014
DIS-NV	0.049	0.124	0.010	0.069	0.115	0.011
Highlights	0.049	0.153	0.042	0.070	0.056	0.006
Per-emotions	0.049	0.145	0.024	0.065	0.159	0.038

The average of the MSE as well as the CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction are calculated (A/V-MSE: the average of the MSE for arousal/valence prediction, A/V-CC/CCC: the average of the CC/CCC absolute values for arousal/valence prediction).

To prove the statistical significance of the results, we first referred the predictions of two SVR models with the highest performance to predictions of a random prediction model for each experiment. As a result, we showed that SVR predictions were significantly different

from random predictions (p-value $\ll 0.0001$). Then, we compared the arousal and valence predictions of these SVR models. We found that all of them were significantly different (p-value $\ll 0.0001$), except for the CC of valence prediction from EDA and visual features (p-value = 0.7584). As shown in Table 5.2, the DBN model best performed induced emotion recognition using audio features of movie content with regard to the values of CC. This means that trends in arousal and valence intensity over time are easily captured. Moreover, the values of the CCC suggest that the DBN is also able to accurately predict the values of arousal scores. However, this is not the case for valence prediction. The DBN achieved the highest values of the CCC for valence prediction from aesthetic highlight annotations.

Firstly, we referred the predictions of two DBN models with the highest performance to random arousal and valence predictions for each experiment. We showed that these DBN models performed significantly different from a random prediction model (p-value $\ll 0.0001$). Then, we compared arousal and valence predictions of these DBN models. We found that all of them were significantly different with p-value $\ll 0.0001$.

The LSTM-RNN model could predict induced arousal from audio features with regard to the CC values, as shown in Table 5.2. However, the values of the CCC suggest that the features of behavioural signals are the most discriminative at least for induced arousal prediction. Moreover, the LSTM-RNN model best performed valence prediction from the physiological signals. The values and trends of valence intensity were captured by the LSTM-RNN model fed by the EDA features. This is confirmed by the high values of the CC and CCC, respectively.

To validate the results of two LSTM-RNN models with the highest performance, we first compared their predictions to random arousal and valence predictions for each experiment. We proved that the predictions of these LSTM-RNN models performed significantly better than random predictions (p-value $\ll 0.0001$). We then compared the predictions of these LSTM-RNN models. As a result, we observed that all of them were significantly different with p-value $\ll 0.0001$. However, there was an exception for the CC of valence prediction based on EDA signals and perceived emotion annotations (p-value = 0.4782).

It is important to mention that our results are not directly comparable with previous emotion recognition research on the C. LIRIS-ACCEDE database (see Table 2.1) due to different data processing procedures, such as the use of the overlapping window and different settings of cross-validation, e.g., the number of folds and the size of training and testing sets. Nevertheless, we can see that we outperformed the state of the art recognition models¹ for valence prediction by means of the LSTM-RNN models with the statistical features of EDA signals (a CC of 0.432).

¹the best reported CC for arousal is 0.337, for valence is 0.296 [12]

Multimodal induced emotion recognition

We report the average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction. Tables 5.3 and 5.4 present the results of multimodal induced emotion recognition experiments. We consider fusion of all the audio-video features with high-level affective clues, such as audio, video, CSA and DIS-NV features as well as aesthetic highlight and perceived emotion annotation based features. Moreover, we investigated the fusion of all the movie content based features mentioned above with physiological and behavioural responses of movie spectators. We compared the proposed hierarchical fusion (LSTM-RNN-HL) architecture of LSTM-RNN models to baseline fusion strategies for LSTM-RNN models, such as feature-level fusion (LSTM-RNN-FL) and decision-level fusion (LSTM-RNN-DL) (see Section 5.2). Also, we examined the recognition performance of SVM and DBN models when the FL fusion was applied.

Table 5.3 Performance of multimodal induced emotion recognition from movie content based features using SVR, DBN, and LSTM-RNN models [142].

Model	A-MSE	A-CC	A-CCC	V-MSE	V-CC	V-CCC
SVR	0.260	0.189	0.004	0.325	0.105	0.002
DBN	0.065	0.195	0.022	0.081	0.113	0.013
LSTM-RNN-FL	0.054	0.218	0.056	0.071	0.110	0.038
LSTM-RNN-DL	0.045	0.144	0.011	0.057	0.186	0.033
LSTM-RNN-HL	0.060	0.111	0.070	0.074	0.061	0.031

The average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction are calculated (A/V-MSE: the average of the MSE for arousal/valence prediction, A/V-CC/CCC: the average of the CC/CCC absolute values for arousal/valence prediction).

As seen in Table 5.3, the LSTM-RNN model with the FL fusion best performed induced arousal recognition from movie content based features with respect to the CC values. It means that trend changes in arousal intensity could be easily captured by this model. Nevertheless, the values of CCC suggest that the proposed hierarchical fusion architecture of the LSTM-RNN model could best predict induced arousal in terms of trends and values. Besides, the LSTM-RNN-HL did not succeed in recognizing induced valence. The LSTM-RNN-DL reached the highest value of the CC. Actually, the LSTM-RNN-FL outperformed the other fusion strategies and predictive models and could the most accurately predict the values and trend fluctuations for induced valence according to the CCC values. Generally, all LSTM-RNN models outperformed SVR and DBN models for induced emotion recognition from movie based features.

Table 5.4 Performance of multimodal induced emotion recognition from audience reaction and movie content based features using SVR, DBN, and LSTM-RNN models [142].

Model	A-MSE	A-CC	A-CCC	V-MSE	V-CC	V-CCC
SVR	0.260	0.251	0.005	0.326	0.179	0.004
DBN	0.065	0.092	0.008	0.081	0.115	0.009
LSTM-RNN-FL	0.055	0.247	0.085	0.070	0.135	0.052
LSTM-RNN-DL	0.043	0.199	0.025	0.076	0.161	0.038
LSTM-RNN-HL	0.076	0.178	0.111	0.087	0.266	0.143

The average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction are calculated (A/V-MSE: the average of the MSE for arousal/valence prediction, A/V-CC/CCC: the average of the CC/CCC absolute values for arousal/valence prediction).

As shown in Table 5.4, the SVR model could be the most accurate predictor of trend changes in induced arousal intensity from fusion of both movie content features and movie audience reactions. However, the large value of the MSE indicates that the SVR model was not able to predict arousal values as well as slight increases and decreases in trends. Furthermore, the LSTM-RNN-HL achieved the highest value of the CCC. This means that the LSTM-RNN-HL could accurately predict downward/upward trend changes in induced arousal intensity as well as its values. Also, the LSTM-RNN-HL best performed induced valence recognition that is confirmed by values of the CC and CCC, respectively. The results that are obtained suggest that the proposed hierarchical architecture of LSTM-RNN models for fusion of movie content features and movie audience reactions is well designed to predict the intensity of induced arousal and valence.

To prove the statistical significance of the results obtained from multimodal fusion, we referred the arousal and valence predictions of two multimodal fusion models with the highest performance to predictions of a random prediction model. We observed that all of them performed significantly different with p-value $\ll 0.0001$. Next, we compared arousal and valence predictions of these pairs of the multimodal fusion models fed by movie content based features as well as movie content based features and statistical features of audience reactions, respectively. We remarked that all of them were significantly different with p-value $\ll 0.0001$.

5.4 Discussion

In this Section, we discuss the limitations of our work and present the open issues regarding the choice of modalities, the sample size and the algorithm selection.

5.4.1 Limitations of our study

Induced emotions can be expressed through different multimodal channels. The amount of information that is passed through each channel is not the same. Some multimodal signals have more discriminative power than others. For example, facial expressions of spectators can convey more information on the induced emotions than their body movements.

Different spectators can have different physiological and behavioural responses to the same stimuli. These can be affected by ambient temperature, body postures, gestures as well as attention and mental effort. Furthermore, induced emotions can vary from one person to another due to many factors e.g., personal life experience. Recording and combining multimodal signals of a group of participants still remain a big challenge due to a lack of access to non-obstructive and reliable sensors. This limits the feasibility of running a large scale experiment in a cinema theater. Measurements of physiological and behavioural signals are often corrupted due to electrode contact noise and sensor device failures during data collection. This results in incomplete data.

Besides, there are many other factors that influence induced emotions in movie audiences, such as personal interest, movie preferences, aesthetic taste, and personality. Also, spectators' emotions are often affected by their recent emotions.

5.4.2 Available modalities and sample size

In our studies, we only analyzed 8 movies from the C. LIRIS-ACCEDE database that come from 4 movie genre. In total, this results in 118 minutes of movies and 7103 labelled instances. Although our conclusions are supported by the magnitudes of effect sizes, we cannot generalize about all movie genres based on such a small number of movies.

Since spectators were watching movies in a darkened cinema theater, only the EDA and ACC measurements of each spectator could be collected. Signals, such as FE, EG, and SG were not recorded due to technical constraints. Our unimodal experiments on induced emotion recognition confirm that spectators have similar physiological responses and display similar behaviours during watching movies. However, the features of ACC measurements are less discriminative than the features of EDA measurements for induced emotion recognition. This outcome might be influenced by the placement of sensors. The sensors were attached to spectators' hands when the experiment was conducted. We do not observe that spectators often make some limb movements when they are watching movies.

The inter-annotation agreement for induced and perceived emotions is low. To reduce this variability in the gold standard, the dynamics of changes in annotations could be considered

instead of emotion intensity. Moreover, some outlier annotations might be removed, and identifying and correcting annotators' biases can be applied.

5.4.3 Model selection

The results that we obtained show that the small amount of labelled instances available for emotion recognition can significantly limit the performance of the system. Model selection is strictly associated with the amount of available data that are recorded and annotated, as well as evaluation metrics. The CC could be selected when the goal is only to capture trend changes in induced emotions by using models. However, the CCC is a more suitable measure to evaluate the quality of models since it describes if models are able to capture changes in trends and values of emotion intensity.

When physiological and behavioural reactions are not recorded and high-level affective cues are not annotated, it is recommended that induced emotions should be recognized by DBN models fed by audio movie features. If physiological or behavioural measurements are available, the results suggest that LSTM-RNN models should be applied due to their capabilities of capturing long term dependencies in movie audience reactions. Besides, when it is only possible to run crowdsourcing annotation experiments, SVM models should be learned on high-level affective cues, such as annotations of aesthetic highlights in movies or perceived emotions of movie audience (see Section 5.3.3).

Our multimodal experiments on induced emotion recognition show that our LSTM-RNN models benefit from including temporal information and combining knowledge-inspired affective cues with audio-visual movie content and movie audience responses. Nevertheless, there is a need to work on LSTM-RNN architectures to incorporate high-level affective cues with audio-visual movie content features since the proposed hierarchical fusion did not improve induced valence recognition (see Section 5.3.3).

The SVM and DBN models could not capture consecutive emotional states and reactions of spectators because they do not take into account temporal information. This is why the LSTM-RNN models could outperform them. Also, feature fusion by means of these baseline models does not allow multimodal features to be incorporated at different stages of modelling. Thus, multilevel fusion is desired to fuse features with different temporal dynamics, e.g., audio-video features of movie content and statistical features of spectators' physiological and behavioural reactions.

The last but not least limitation is that these basic models cannot deal with noisy features and temporal evolution of the probability distribution of movie content features and statistical features of movie audience reactions. The probability distribution varies from one movie to another because measurements of physiological and behavioural signals are corrupted by

electrode contact noise and they are subject-dependent. Furthermore, audio-video features are contaminated with movie background noise. On the contrary, the LSTM-RNN models are able to operate on different scales of time which limits the influence of variability of spectators' physiological and behavioural signals and movie content. Also, noisy features can be filtered out by learning a new representation in the first layer of LSTM-RNN models.

5.5 Conclusion

This work clarifies the difference between perceived and induced emotions of movie audiences and may serve as a reference for future affective content analysis studies. We extend the annotations of the C. LIRIS-ACCEDE database. We find that perceived and induced emotions of movie audiences are not always positively correlated, responding to our **third research question (RQ3)**. Although the inconsistency was observed on a small movie data set, it should be taken into account while selecting emotional stimuli. There is more to be considered than simply assuming that the perceived emotions of the stimuli are consistent with the emotions induced in spectators. To expand our understanding of perceived and induced emotions and answer our **fourth research question (RQ4)**, we used perceived emotions to predict induced emotions. Moreover, perceived and induced emotions of the movie audiences are associated with the occurrence of aesthetic highlights in movies. These highlights are considered to be high level affective cues for induced emotion recognition.

The improvement of performance of LSTM-RNN models by means of multimodal hierarchical fusion leads to the conclusion that adding other modalities, such as FE, HR, and EEG signals of spectators could effect an even larger increase of performance. Also, our promising model can be scalable to a larger movie set and benefit from a larger number of training instances. Nevertheless, there is a need to deeply study at which layer of the model audio-video features and affective cues should be incorporated.

Inspired by audio-visual features benefiting from including in-domain knowledge, we will be studying the advantages of using transfer learning between different emotion recognition tasks. The pretrained models on other emotion recognition challenges, e.g., emotion recognition of individuals watching short videos could be applied to induced emotion recognition of movie audiences. Improved performance may be achieved by learning new feature representations that reduce inter/intra-person variability of physiological and behavioural responses.

In addition, we plan on conducting further investigations into how emotions and affective cues differ from one movie genre to another, e.g., action, crime, epics, historical, horror, etc. Studies on a wide range of movie emotions may make a major contribution to cinematography

research as well as help moviemakers to design affective content with better alignment of intended and induced emotions.

Chapter 6

Exploring aesthetic emotions of movie audiences

In this Chapter, we focus on the relationship between the occurrences of aesthetic highlights in movies and aesthetic emotions felt by spectators (see Sections 3.3 and 3.5). In particular, we investigate whether or not aesthetic highlights evoke a wide range of aesthetic emotions beyond "everyday" emotions, e.g., anger, fear, sadness, happiness, joy, and surprise. Also, we study the dependencies between "everyday" emotions and aesthetic emotions, such as *awe*, *boredom*, *disgust*, *being touched*, and *wonder*. In particular, we examine whether or not a continuous arousal-valence space is sufficient to accurately represent movie aesthetic emotions. We argue that personality can influence aesthetic preferences and the intensity of aesthetic emotions felt while watching movies. It is important to mention that we do not study the effect of mood on the intensity of aesthetic emotions. Moreover, we attempt to recognize movie aesthetic emotions based on spectators' reactions. Our modality selection is motivated by work of Tarvainen et al. [187, 188] in which the prediction of movie aesthetic attributes from audio-visual features failed. We extract different features of EDA and ACC measurements to investigate which characteristics of these signals are the most discriminative for aesthetic emotion recognition. Our work on aesthetic emotions can help movie industry to design movie content that can be personalized and evoke different emotions in individuals on demand.

To quantitatively study dependencies between aesthetic highlights, aesthetic emotions, "everyday" emotions, personality, and spectators' physiological and behavioural reactions, we address the following research questions (see a full list of research questions in Section 1.5):

1. Is there a direct relationship between "everyday" emotions and aesthetic emotions?
(RQ5)

- Is an arousal-valence space sufficient to accurately represent aesthetic emotions?
2. Are there dependencies among aesthetic highlights, aesthetic emotions, personality, and spectators' physiological and behavioural reactions? (RQ6)
- Do aesthetic highlights elicit aesthetic emotions beyond "*everyday*" emotions?
 - Does personality influence aesthetic preferences?
 - Is it possible to predict aesthetic emotions from physiological and behavioural responses of spectators?

We emphasize the contributions of our work below, highlighting the novelty compared to the state of the art research (to the best of our knowledge):

- We carry out the first quantitative analysis of the relationship between "*everyday*" emotions and aesthetic emotions.
- We are first to discover the relationship between aesthetic emotions evoked in movie audiences and the occurrences of aesthetic highlights in movies.
- We quantitatively investigate the influence of personality on aesthetic preferences.
- We successfully use EDA and ACC signals of spectators to recognize aesthetic emotions.
- We create one of the largest databases of aesthetic emotion annotations that will allow researchers to carry out research on film aesthetic experience. This database contains aesthetic emotion annotations of 30 full-length movies derived from 9 movie genres: action, adventure, animation, comedy, documentary, drama, horror, romance, and thriller.

6.1 Physiological and behavioural feature extraction

To remove signal artifacts, we filtered all signals which are EDA and ACC measurements by means of a third order low-pass Butterworth filter with cut-off frequency at 0.3 Hz. Then, we extracted 3 different sets of features from EDA and ACC signals of spectators, namely statistical, wavelet, and synchronization features. To capture a suitable amount of temporal information conveyed by the multimodal signals for feature extraction, we used a 5 second sliding window with a 4 second overlap between neighbouring windows to compute all features. The same features of each spectator then are concatenated into one feature vector.

6.1.1 Statistical features

Statistical features are mean, median, standard deviation, minimum and maximum value as well as minimum and maximum ratio over the sliding windows of original signals of their first and second derivatives [114]. Consequently, the same features were computed over sliding windows of EDA and ACC measurements collected from sensors attached to the spectators' hands (see Section 5.1.1).

6.1.2 Wavelet features

Wavelet transforms are considered to be appropriate for analysis of non-stationary signals with low and high frequency components [3]. We used the Wavelet Packet Transform (WPT) to extract features of EDA and ACC signals that capture trend changes, spikes, and drifts. The WPT can be considered as a tree of subspaces, where $\Omega_{0,0}$ represents a space of the original signal and is the root node of the tree. In general, $\Omega_{j,k}$ that corresponds to the node with the scale index j and subband index k . It is decomposed into two orthogonal subspaces $\Omega_{j+1,2k}$ and $\Omega_{j+1,2k+1}$. This is obtained by means of splitting the orthogonal basis $\{\phi_j(t - 2^j k)\}_{k \in Z}$ of $\Omega_{j,k}$ into two new orthogonal basis $\{\phi_{j+1}(t - 2^{j+1} k)\}_{k \in Z}$ of $\Omega_{j+1,2k}$ and $\{\psi_{j+1}(t - 2^{j+1} k)\}_{k \in Z}$ of $\Omega_{j+1,2k+1}$, where $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are the scaling wavelet functions expressed as follows [105]

$$\phi_{j,k}(t) = \frac{1}{\sqrt{|2^j|}} \phi\left(\frac{t - 2^j k}{2^j}\right), \psi_{j,k}(t) = \frac{1}{\sqrt{|2^j|}} \psi\left(\frac{t - 2^j k}{2^j}\right), \quad (6.1)$$

where 2^j is the dilation factor and represents scaling and $2^j k$ is the translation parameter and indicates the time location. The decomposition process can be iteratively repeated to generate a binary wavelet packet tree in which nodes represent subspaces with different frequency. To extract features, we calculate energy, variance, waveform length, and entropy of each the WPT tree subspaces.

6.1.3 Synchronization features

We use pairwise synchronization measures to estimate dependencies between spectators' physiological and behavioural signals during watching movies [139]. Let us suppose there are two time windows $x_i(l)$ and $x_j(l)$ for $l = 1, \dots, N$, where N is the total number of sliding windows.

Dynamic time warping

Dynamic Time Warping (DTW) distance between two time windows $x_i(l)$ and $x_j(l)$ is the total cost of the warping path W^* , and it is expressed by [16]

$$d_{DTW}(x_i(l), x_j(l)) = c_{W^*}(x_i(l), x_j(l)), \quad (6.2)$$

where W^* is a warping path that has a minimal total cost among all possible warping paths (see Section 4.2.1).

Windowed mutual information

Windowed Mutual Information (WMI) is able to capture nonlinear dependencies between signals in terms of synchronization. We estimate the WMI by means of calculating the joint entropy $H(x_i(l), x_j(l))$ and marginal entropies $H(x_i(l))$, $H(x_j(l))$, respectively [116]. The normalized WMI is defined as follows

$$I(x_i(l), x_j(l)) = \frac{H(x_i(l)) + H(x_j(l)) - H(x_i(l), x_j(l))}{\sqrt{H(x_i(l))H(x_j(l))}}. \quad (6.3)$$

Phase locking value

Although the amplitudes of signals are statistically independent, their instantaneous phases can be strongly synchronized. That corresponds to phase synchronization [120]. Phase Locking Value (PLV) is defined as

$$PLV(x_i(l), x_j(l)) = \left| \frac{1}{m} \sum_{n=1}^m e^{i\Delta\theta(t_n)} \right|, \quad (6.4)$$

where $\Delta\theta(t_n)$ is a phase difference between two signals on sliding window interval, and m is a number of samples inside the sliding window. If there is a strong phase synchronization between signals, the relative phase is small and thus the PLV value is close to 1.

6.2 Recognition model for aesthetic emotions

Restricted Boltzmann Machines (RBMs) can be used to learn a representation of unimodal and multimodal signals in many applications, such as image recognition, speech recognition, and emotion recognition [179]. A RBM assumes that the input data comes from a binary distribution. This is the critical limitation of the model. Thus, the RBM assuming that the

data are derived from a Gaussian distribution was proposed in [92]. We chose a Gaussian-Bernoulli RBM (GBRBM) for aesthetic emotion recognition. A GBRBM is the RBM with a Gaussian distribution for the visible units and binary distribution for the hidden units [203]. The visible units and hidden units in the GBRBM are fully connected. Also, the visible units of the input layer correspond to features feeding the model.

We selected the GBRBM because this model could take into account the fact that the distributions of physiological and behavioural signals are better fitted to the Gaussian distribution than the pseudo binary distribution. This was confirmed by using the Dip test for unimodality [89]. A DBN is a stack of multiple RBMs. The hidden units of a learned RBM are used as the visible units of the following RBM. The DBNs are able to learn a high level representation from unlabelled instances. Each layer is supposed to represent the data at a higher level of abstraction. Then, relatively small number of labelled data is needed to fine-tune the model. We selected 2 hidden layer DBNs with first layers composed of GBRBMs because real values of features come from unimodal distributions.

We learned the DBN with only 2 hidden layers with 50 and 15 neuron units, respectively, due to the limited number of training instances. Cross entropy was used to define a loss function. The size of mini-batch was the number of features divided by 4 for computational efficiency, while the initial learning rate and its upper bound were set to 0.002 for pre-training and the weight-updating ratio was set to 0.1. Also, gradient descent based supervised fine tuning with maximum 100 iterations was applied to find optimal parameters for the whole DBN [15]. To avoid overfitting on the limited training set, a dropout with the ratio of 0.5 for hidden layers was used.

6.3 Results

6.3.1 Dependencies between aesthetic and "everyday" emotions

Many studies on affective content analysis [12, 125, 177, 180] assume that emotions elicited by multimedia content can be sufficiently represented in a continuous arousal-valence space [63]. In this section we wonder whether or not movie aesthetic emotions are "everyday" (basic) emotions (RQ5). In particular, we aim to show how many emotional dimensions are required to accurately describe aesthetic emotions felt by movie audiences. To address these questions and uncover underlying dimensional structure of aesthetic emotions, we conducted a linear principal component analysis. Although we considered each movie as one separate experiment, we calculated principal components of aesthetic emotion annotations on all 30

movies from the C. LIRIS-ACCEDE database. We find that 4 principal components explain around 90% aesthetic annotation variance.

To identify the underlying dimensions of aesthetic emotions, we attempt to uncover dependencies between induced arousal and valence dimensions and these 4 principal components. We used the CC as the effect size to measure the dependencies. To integrate the results from different movies and obtain an overall effect size, we selected a fixed-effect model [21]. We interpret a weighted average effect size of the CC that is around 0.1, 0.3, and 0.5 as a small, medium, and large effect size, respectively [43].

Firstly, we calculated the weighted average effect size on the whole C. LIRIS-ACCEDE database, as presented in Table 6.1. We observe that only the 4th principal component is strongly positively correlated (large effect) with the arousal dimension, as well as the 1st and 4th principal components are strongly positively correlated (large effects) with the valence dimension. This means that the rest of the principal components are not associated with emotions described in the arousal-valence space. Furthermore, this suggests that at least two more dimensions beyond arousal and valence are required to characterize aesthetic emotions of movie audiences. For example, understanding and self-consciousness might be considered as extra dimensions. The analysis leads us to the conclusion that the arousal-valence space only partially represents movie aesthetic emotions. Thus, the basic assumption that emotions evoked by multimedia content can be sufficiently describe in the continuous arousal-valence space is not accurate when we study aesthetic emotions of movie audiences.

Table 6.1 The weighted average effect size of the CC (fixed-effect model) between 4 Principal Components (PCs) and induced arousal and valence dimensions calculated over the C. LIRIS-ACCEDE database.

PC	Arousal	Valence
1st	0.22	0.69
2nd	-0.03	0.10
3rd	0.27	0.09
4th	0.69	0.79

To investigate whether or not movie genre influences a dimensional representation of movie aesthetic emotions, we conducted the same meta analysis for each of 9 movie genres (see Section 3.2). Tables 6.2 and 6.3 present the weighted average effect size of the CC. We measured dependencies between 4 principal components and the induced arousal dimension as well as the induced valence dimension per movie genre. The directions of effects varies from one movie genre to another. We only considered magnitudes of effects regardless of

Table 6.2 The weighted average effect size of the CC (fixed-effect model) between 4 Principal Components (PCs) and induced arousal dimension calculated per movie genre.

PC	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
1st	0.38	0.01	-0.60	0.01	0.42	0.56	-0.09	-0.36	0.29
2nd	-0.40	-0.25	-0.58	0.56	-0.14	0.52	0.06	-0.33	-0.14
3rd	-0.42	0.22	0.53	0.30	0.63	0.58	0.23	0.44	-0.06
4th	-0.62	-0.01	0.93	0.16	0.17	0.03	0.29	0.36	-0.04

Table 6.3 The weighted average effect size of the CC (fixed-effect model) between 4 Principal Components (PCs) and induced valence dimension calculated per movie genre.

PC	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
1st	-0.52	0.08	-0.35	0.54	0.47	0.94	-0.27	0.62	-0.20
2nd	0.24	0.18	-0.41	0.56	-0.32	-0.16	-0.27	-0.35	0.16
3rd	-0.82	0.07	0.04	0.29	0.36	0.47	-0.04	-0.57	0.12
4th	0.86	-0.16	0.58	0.21	0.16	0.90	0.15	-0.62	-0.06

their directions for an interpretation of the results because we focused on discovering the principal components that could be associated with induced arousal and valence dimension.

As we can see in Table 6.2, there are the dependencies (medium and large effects) between all 4 principal components and the arousal dimension for drama, thriller, and documentary movies. Also, we can find movie genres, such as action and comedy as well as romance for which either 2 or 3 principal components are strongly correlated (medium and large effects) with the induced arousal dimension. This suggests that movie aesthetic emotions represented by 4 principal components can be linked to arousal intensity for these movie genres only. Nevertheless, we observe that none of 4 principal components is correlated with the arousal dimension for animations, adventures, and horrors. This means that aesthetic emotions elicited by these movies are not represented by the arousal dimension. Thus, emotional annotations in terms of the arousal dimension do not encompass aesthetic emotions felt movie audiences for these movie genres.

As shown in Table 6.3, we find that at least 3 principal components are strongly correlated (medium and large effects) with the induced valence dimension for documentaries, dramas, thrillers, comedies, and romances only. As a result, movie aesthetic emotions are associated with the induced valence dimension for these movie genres. Furthermore, 2 principal components are linked to valence when we analyze action movies. However, we see that none of 4 principal components is significantly correlated with the valence dimension for dramas, animations, adventures, and horror similarly to the arousal dimension. This confirms that emotional annotations in terms of valence do not cover movie aesthetic emotions.

To sum up, we uncover that aesthetic emotions felt by movie audiences cannot be represented by two dimensions only: arousal and valence. These two dimensions do not provide sufficient representations of movie aesthetic emotions. Hence, the affective computing and multimedia community should take into account the fact that a new dimensional representation of movie aesthetic emotions is required to accurately examine film aesthetic experience. Moreover, emotional annotations in an arousal-valence space largely omit any aesthetic emotions for animations, adventures, and horrors. This means that there are a few categories of an aesthetic emotion elicited in movie audiences. These categories are enhanced by watching only particular movies. Furthermore, this suggests that any aesthetic emotion should be annotated individually when analysis of film aesthetic experience is conducted for these movie genres.

6.3.2 Dependencies between aesthetic highlights and aesthetic emotions

In this Section we investigate dependencies between aesthetic highlights in movies and aesthetic emotions of movie audiences, answering our research question RQ6. We attempt to confirm that aesthetic highlights in movies can evoke aesthetic emotions beyond "everyday" emotions (see Section 4.3.1). We associated the occurrences of aesthetic highlights in movies with the intensity of *awe*, *boredom*, *disgust*, *being touched*, and *wonder*. In order to measure the influence of aesthetic highlights on affective states of movie audiences and film aesthetic experience, we used a meta analysis. Each movie was considered as one separate experiment due to different content, duration, and the amount of aesthetic highlights in movies from the C. LIRIS-ACCEDE database (see Section 3.3).

To evaluate it, effect size estimators were computed over each movie. We selected the standardized mean difference as an effect-size. It is expressed by the difference between mean values of a given emotion intensity over highlight and non-highlight intervals normalized by their pooled standard deviation. Positive values reveal a higher level of an emotion intensity during highlight scenes than non-highlight scenes while negative values mean a lower level of intensity.

We combined the effect-sizes using a fixed-effect model that weights each effect size estimate based on its precision [21]. To interpret the statistical significance of the dependencies between aesthetic highlights and aesthetic emotions, we followed Cohen's benchmarks [43]. We consider the values close to 0.2, 0.5, and 0.8 as small, medium, and large effect sizes, respectively. Table 6.4 presents the weighted average effect size of each of the five aesthetic emotions on the whole C. LIRIS-ACCEDE database. A small negative effect size of awe is observed for highlights H1. Also, a small positive effect size of boredom is found for highlights H3 and H4. Moreover, a small positive and negative effect size of disgust is

Table 6.4 The weighted average effect size (fixed-effect model) of the five aesthetic emotions during aesthetic highlights over all the C. LIRIS-ACCEDE database.

Emotions \ Highlights	H1	H2	H3	H4	H5
Awe	-0.10	-0.11	0.03	0.13	-0.25
Boredom	0.09	-0.02	0.25	0.42	-0.06
Disgust	-0.03	-0.05	0.10	0.22	-0.27
Being touched	0.15	0.04	0.17	0.34	0.03
Wonder	0.22	-0.22	0.25	0.11	0.26

identified for highlights H4 and H5. Furthermore, a small positive effect of being touched is retrieved for highlights H4. In addition, a small negative effect size of wonder is revealed for highlights H2 while a small positive effect size is reported for highlights H1, H3, and H5. It is important to mention that we do not observe any effect sizes of medium and large magnitudes.

We can only suppose that highlights H1 increase wonder intensity because the subjects are presented in a spectacular way by means of technical choices and special effects, e.g., adding artificial fog. Subtle highlights H2 do not definitely elicit emotions themselves while highlights H3 can make movie audiences feel bored or wonder. As expected, although movie dialogues can evoke boredom and disgust, some part of them can also cause that spectators are touched by movie events. Furthermore, theme development elicits wonder that is composed of fear, surprise, and joy changing over time.

To study how movie genre affects aesthetic emotion elicitation, we carried out the same meta analysis for each of 9 movie genres. As we can see in Tables 6.5, 6.6, 6.7, 6.8, and 6.9, the direction of the effect size varies from one movie genre to another for a given aesthetic emotion. We expect that spectacular highlights H1 elicit strong emotional reactions

Table 6.5 The weighted average effect size (fixed-effect model) of awe intensity during aesthetic highlights calculated per movie genre.

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	-0.64	-0.36	-0.57	-0.98	0.82	-0.54	-0.25	0.25	-0.03
H2	-0.18	-0.50	0.05	0.17	0.05	0.56	-0.43	0.26	-0.13
H3	0.23	0.15	-0.46	0.07	0.12	0.56	0.57	0.26	-0.06
H4	0.37	-0.50	0.02	0.04	0.23	-0.42	0.37	-	0.05
H5	-0.47	-0.01	-0.40	-0.09	0.29	-0.12	0.25	-0.77	-0.48

in spectators. Technical choices and special effects are supposed to intensify film aesthetic experience and thus aesthetic emotions as well. We find a large positive effect size of awe for

Table 6.6 The weighted average effect size (fixed-effect model) of boredom intensity during aesthetic highlights calculated per movie genre.

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	0.23	0.07	0.01	0.01	0.70	0.77	-0.11	0.23	-0.02
H2	0.26	-0.49	0.12	-0.32	-0.44	0.70	0.11	0.24	0.18
H3	-0.28	0.12	-0.06	0.02	0.46	0.82	0.30	0.24	0.31
H4	-0.32	0.31	-0.37	-0.37	0.51	-0.45	0.56	-	0.90
H5	0.32	0.42	-0.22	-0.02	0.56	-0.26	0.15	-0.76	-0.37

Table 6.7 The weighted average effect size (fixed-effect model) of disgust intensity during aesthetic highlights calculated per movie genre.

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	-0.04	0.07	-0.32	-0.99	-0.41	-0.33	-0.17	0.38	0.18
H2	-0.15	-0.50	-0.28	0.05	0.12	0.06	-0.06	0.28	0.19
H3	0.43	-0.37	-0.55	0.37	0.11	0.30	-0.43	0.32	0.09
H4	0.05	0.36	0.46	0.14	-0.21	0.26	0.30	-	0.57
H5	-0.06	0.07	-0.45	-0.09	-0.30	0.01	0.29	-0.62	-0.43

Table 6.8 The weighted average effect size (fixed-effect model) of being touched intensity during aesthetic highlights calculated per movie genre.

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	-0.25	-0.17	-0.22	0.75	0.05	-0.79	0.01	0.10	0.25
H2	-0.19	-0.58	0.38	0.55	0.13	0.48	-0.09	0.02	0.25
H3	0.42	-0.14	-0.17	0.34	0.11	0.23	0.08	0.02	0.16
H4	0.31	-0.16	-0.32	0.13	0.44	0.41	0.47	-	0.37
H5	-0.33	0.18	-0.49	0.17	0.01	-0.08	0.14	0.01	0.10

Table 6.9 The weighted average effect size (fixed-effect model) of wonder intensity during aesthetic highlights calculated per movie genre.

H \ Genre	Drama	Animat.	Thrill.	Action	Comed.	Roman.	Advent.	Docum.	Horror
H1	-0.04	-0.37	-0.60	-0.83	-0.48	0.65	-0.30	0.58	0.67
H2	0.17	-0.70	-0.51	-0.08	-0.27	-0.42	-0.04	0.28	-0.14
H3	-0.16	0.09	-1.0	-0.01	-0.02	-0.27	-0.33	0.37	0.60
H4	-0.69	-0.07	0.40	-0.13	0.52	0.65	0.13	-	0.11
H5	-0.11	0.10	-0.53	0.03	0.11	0.02	0.15	-0.25	0.50

comedies while observing a large negative effect of awe for action movies and a medium negative size effect for dramas, thrillers, and romances, as shown in Table 6.5.

Regarding boredom intensity during different aesthetic highlights presented in Table 6.6, we reveal a medium positive effect size for comedies and romances. Also, we can see a large negative effect size of disgust for action movies in Table 6.7. Moreover, intensity of being touched increases during spectacular scenes for action movies, as shown in Table 6.8. This is described by a medium positive effect size. Furthermore, we find a large and medium negative effect size of wonder for action and thriller movies, respectively, while observing a medium positive effect for romances, documentaries, and horrors, as presented in Table 6.9. It is worth mentioning that the intensity of awe, disgust, and wonder strongly decreases during spectacular scenes in action movies, unlike the intensity of being touched. This means that spectacular highlights clarify movie stories and increase the spectators' engagements with movies.

Subtle highlights H2 that include usage of cameras, lightening, and music do not elicit aesthetic emotions, except for romances. We suppose that subtle scenes are essential to create and express a romantic atmosphere of a movie story. This is supported by a medium positive effect size of awe and boredom for the romantic movies, as shown in Tables 6.5 and 6.6. Also, we find a negative medium effect size of awe, disgust, being touched, and wonder for animations, as presented in Tables 6.5, 6.7, 6.8, and 6.9. This can be explained by the fact that some part of movie audiences could have difficulties with perceiving animated worlds. It means that subtle scenes in animated movie stories, such as mirroring the face of main characters in the water or adding shadows in the animated world could be omitted by spectators because these scenes are not directly associated with the real world.

Highlights H3 that consist of the main characters' development and tensions among them evoke aesthetic emotions in movie audiences only for some movie genres, such as romance, adventure, and horror movies. We can see a medium positive effect size of awe for romance and adventure movies in Table 6.5. This means that the main characters' development in these two movie genres is designed to elicit awe that is a mixture of surprise and fear. Also, we observe a large positive effect of boredom for romances, as shown in Table 6.6. This suggests the main characters' stories in these movies are too long and spectators do not enjoy watching them. Moreover, we find a medium negative effect of disgust for thriller movies in Table 6.7. Thus, the main characters' development in thrillers is not disgusting for spectators. Furthermore, we reveal that the characters' development in horrors increase wonder intensity among spectators. This is described by a medium positive effect size in Table 6.9. Thus, horrors are made in a such way that the uncertainty of the main characters' stories increases over time and spectators feel surprise and joy at the end of the movies.

Highlights H4 that are defined as dialogues among main characters evoke aesthetic emotions if they occur in particular movie genres, such as comedy, adventure, horror, and romance. We find a medium positive effect size of boredom for comedy and adventure movies as well as a large positive effect size of boredom for horrors, as shown in Table 6.6. This suggests that dialogues in these movies are too long and thus spectators lose the interest in watching.

Also, we observe a medium positive effect size of disgust for horrors in Table 6.7. This means that horror dialogues are written in a very peculiar way that shows the tensions among the main characters when their lives are threatened by danger. Furthermore, we observe a medium positive effect size of wonder for comedies and romances in Table 6.9. It is related to the fact that the comedy and romance dialogues are intended to make spectators feel curious and joyful.

Besides finding these positive effect sizes, we observe a medium negative effect size of awe for animations as well as a medium negative effect size of wonder for dramas, as shown in Tables 6.5 and 6.9, respectively. We can suppose that animation and drama dialogues occur when movie stories in these movies are already clarified. This means that movie audiences are enough familiar with the stories to predict the next movie events.

Highlights H5 cover theme development that often co-occurs with different categories of aesthetic highlights, e.g., highlights H3. The reason is that main characters' development could be a part of theme development. The meta analysis enhances a medium positive effect size of wonder for horrors, as presented in Table 6.9. As expected, theme development in horror movies is designed to evoke a combination of fear, curiosity, and joy in movie audiences with their periodical accumulation. As we can see in Tables 6.6 and 6.7, there is a medium negative effect of boredom and disgust for documentary movies. This can be explained by the fact that theme development of these movies has a special structure in which the subjects and events are objectively presented without emotional tones.

6.3.3 Personality and aesthetic preferences

Emotional states of individuals are influenced by their personalities [37]. Personality describes individual characteristics of behaviour and cognition [132]. The Big Five Factor Model characterizes human personality by means of five dimensions (traits), namely *extraversion* (sociable vs. reserved), *agreeableness* (compassionate vs. dispassionate), *conscientiousness* (dutiful vs. easy-going), *neuroticism* (nervous vs. confident) and *openness* (curious vs. conservative) [150]. Personality in terms of these five dimensions can be measured by means of self-completion questionnaires. In our studies, the big five personality dimensions were quantified by an online questionnaire (self-report big five inventory test with 44 questions

rated in a 5 point scale [97]) that were filled in by annotators before starting aesthetic emotion annotations.

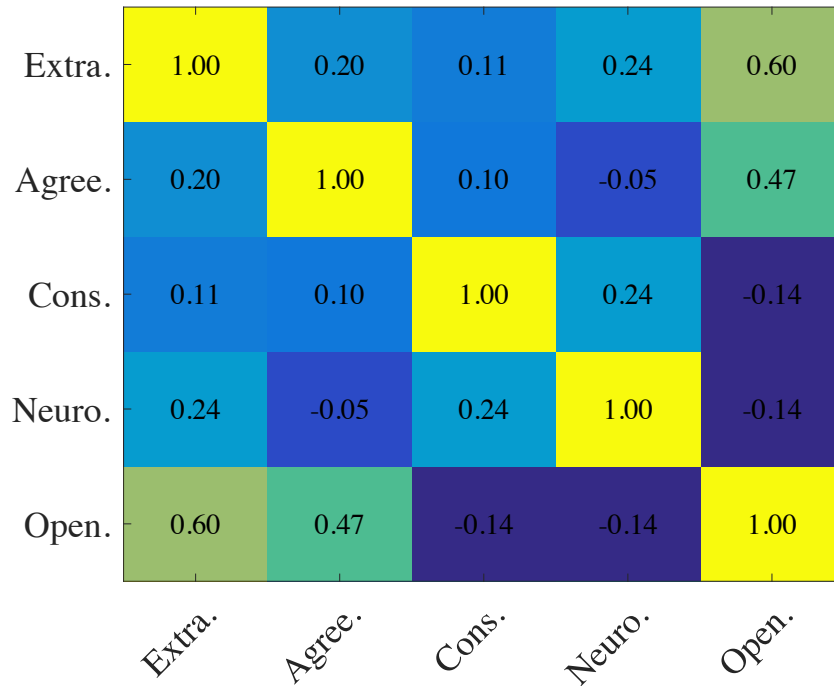


Fig. 6.1 The weighted average of the CC between the big five personality traits: extraversion, agreeableness, conscientiousness, neuroticism, and openness (yellow and purple color indicate strong correlation and anti-correlation, respectively).

We selected the CC as the effect size to measure dependencies among the personality dimensions of aesthetic emotion annotators. To aggregate dependencies between annotators' personality scores for each movie and aesthetic emotion, we used a fixed-effect model [21]. To interpret the statistical significance of the dependencies between personality dimensions, we followed Cohen's benchmarks [43]. We consider the values close to 0.1, 0.3, and 0.5 as small, medium, and large effect sizes, respectively. Figure 6.1 presents the weighted average of the CC between the big five personality traits of all aesthetic emotion annotators. We observe that the openness dimension is strongly positively correlated (large effect) with the extraversion dimension, as well as is moderately positively correlated (medium effect) with the agreeableness dimension. This is in line with previous work on personality [180]. This suggests that annotators who are open to new experiences (e.g., creative and unconventional) are also energetic, social, modest, and altruistic.

To respond to our research question RQ6, we then conducted a similar analysis of relationships between personality and aesthetic ratings for awe, boredom, disgust, being

touched, and wonder, as shown in Table 6.10. In particular, we examined the weighted average of the CC between personality traits and the average annotation score of each annotator. The extraversion dimension is strongly positively correlated (almost large effect)

Table 6.10 The weighted average of the CC between the Big-Five Personality Traits and average annotation ratings of aesthetic emotions.

Emotions \ Personality traits	Extra.	Agree.	Cons.	Neuro.	Open.
Awe	-0.02	0.11	-0.06	-0.06	0.22
Boredom	0.48	0.11	0.03	-0.11	-0.09
Disgust	-0.19	-0.12	-0.25	-0.16	-0.05
Being touched	-0.06	-0.08	-0.12	-0.12	-0.01
Wonder	-0.02	-0.04	-0.01	0.04	-0.03

with boredom scores and weakly negative correlated (small effect) with disgust scores. On the one hand, extraverts often feel boredom when watching movies and thus their movie expectations are very high. On the other hand, disgust can be easily elicited in introverts by movie scenes.

As we can see in Table 6.10, the agreeableness dimension is weakly positively correlated (small effect) with awe and boredom ratings and weakly negatively correlated (small effect) with disgust ratings. Thus, a mixture of surprise and fear can be easily evoked in agreeable groups of annotators by movie scenes. However, the groups of disagreeable annotators who are stubborn and suspicious feel disgusted while watching some movie scenes. Moreover, we observe that conscientiousness is weakly negative correlated (small effect) with disgust and being touched scores. Conscientious annotators tend to deliberate on the whole movie and all movie events in order to deeply understand the movie story. They consider various movie scenes as pleasant rather than offensive for movie audiences.

Furthermore, we find that the neuroticism dimension is weakly negatively correlated (small effect) with boredom, disgust, and being touched ratings. Neurotic annotators are moody and interpret ordinary movie events as difficult life situations for main characters. This limits the engagement with movie stories. In addition, openness dimension is weakly positively correlated (small effect) with awe ratings. Open annotators who have active imagination and intellectual curiosity can be engaged with various movie stories and movie characters. This allows them to feel surprise and fear while movie stories develop.

To sum up, we studied the differences in aesthetic scene ratings with respect to personality characteristics. This suggests personality can affect a level of spectators' engagement with movie story development. Thus, film aesthetic experience can be also influenced by

personality. For example, neurotic spectators cannot enjoy watching movies because they strongly overreact to ordinary movie events due to being mentally unbalanced.

6.3.4 Aesthetic emotion recognition

To address our research question RQ6 and prove that it is possible to recognize aesthetic emotions evoked in movie audiences based on physiological and behavioural (ACC) responses, we extract different feature sets from EDA and ACC signals of spectators, as described in Section 6.1. To align aesthetic annotations with extracted features, we averaged aesthetic emotions scores over each sliding window as the gold-standard annotations. We also discarded the end credits of each movie because participants started to remove the wearable sensors that added distortions to the measurements.

Table 6.11 Performance of unimodal aesthetic emotion recognition using DBNs.

Features	EDA			ACC		
	MSE	CC	CCC	MSE	CC	CCC
Awe						
Statistical	0.011	0.157	0.013	0.011	0.152	0.012
Wavelet	0.011	0.032	0.001	0.011	0.034	0.001
Synchronization	0.011	0.103	0.005	0.011	0.131	0.007
Boredom						
Statistical	0.012	0.243	0.020	0.012	0.145	0.015
Wavelet	0.012	0.027	0.001	0.012	0.039	0.002
Synchronization	0.012	0.098	0.007	0.012	0.120	0.006
Disgust						
Statistical	0.009	0.219	0.018	0.009	0.144	0.023
Wavelet	0.009	0.032	0.001	0.009	0.025	0.001
Synchronization	0.009	0.072	0.003	0.009	0.125	0.006
Being touched						
Statistical	0.012	0.226	0.027	0.012	0.101	0.010
Wavelet	0.012	0.038	0.001	0.012	0.029	0.002
Synchronization	0.012	0.057	0.002	0.012	0.113	0.005
Wonder						
Statistical	0.015	0.177	0.016	0.015	0.116	0.010
Wavelet	0.015	0.027	0.001	0.015	0.031	0.001
Synchronization	0.016	0.054	0.006	0.015	0.138	0.005

The average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for aesthetic emotion prediction are calculated.

Then, we used the DBN models to separately recognize each aesthetic emotion. We performed leave-one-movie-out cross-validation and we report the unweighted average of the MSE, the absolute CC, and the absolute CCC. A high value of the CC corresponds to a strong linear dependency between trend changes in emotion predictions and emotion annotations while a low value of the MSE is associated with high quality prediction of emotion intensity in terms of intensity values. The CCC merges the CC with the square difference between the means of prediction and annotations scores and their variances. A CCC value around 1 means that there is a strong linear relationship between trend changes in emotion predictions and emotion annotations and the values of predictions and annotations are similar to each other.

In Table 6.11, we illustrate the prediction performance of each aesthetic emotion based on different feature sets extracted from either EDA or ACC signals. To show that DBNs with different features of EDA and ACC measurements performed better than a random prediction model, we generated prediction scores for each aesthetic emotion at random. We then compared predictions of the two models with highest CC or CCC values for each experiment to random predictions of awe, boredom, disgust, being touched, and wonder, respectively. We evaluated the significance differences of the performance by means of two-sample Wilcoxon tests at the significant level of 0.05. When we report results for each experiment, numbers in bold italics indicate significantly best performance regarding CC and CCC values (p-value $\ll 0.0001$) while numbers in bold indicate only highest values of the CC and CCC.

We showed that two the best models for each emotion prediction based on EDA or ACC measurements were significantly different from random prediction (p-value $\ll 0.0001$). Also, we found that all pairs of these models were significantly different with p-value $\ll 0.0001$, except for the comparison between statistical features and synchronization features of the ACC signals for awe prediction (p-value = 0.1799), as shown in Table 6.11. The small values of the MSE suggest that all the models were able to predict awe, boredom, and disgust intensity from EDA and ACC signals, as shown in Table 6.11. As a result, we observe that the DBN models with statistical features of EDA signals achieved the best performance for awe, boredom, disgust, being touched, and wonder prediction measured by the mean value of the absolute CC and CCC. Moreover, we can see that statistical features of ACC signals are also discriminative for boredom and disgust prediction, as shown in Table 6.11.

With reference to small values of MSE in Table 6.11, all the DBN models fed by different features (statistical and synchronization features) of ACC measurements could predict the intensity of being touched and wonder. The highest performance was achieved by the DBNs

with statistical features of EDA signals for boredom, disgust and being touched prediction regarding the mean value of the absolute CC and the absolute CCC, respectively.

In general, we observe that the DBN models with different descriptors of spectators' physiological and behavioural reactions were able to predict the intensity of each aesthetic emotion. Nevertheless, only the DBN fed by statistical features of EDA and ACC signals could best capture trend changes in aesthetic ratings, except for prediction of being touched and wonder based on ACC signals. This shows that the analysis of dynamic changes in spectators' physiological and behavioural responses is critically important for movie aesthetic emotion recognition. Furthermore, the results suggest that the DBN with synchronization features could be used when spectators' reactions are synchronized by watching strongly emotional scenes.

6.4 Discussion and conclusions

In this work, we investigate aesthetic emotions evoked in movie audiences. We extend the C. LIRIS-ACCEDE database by crowdsourcing annotations of five aesthetic emotions, namely awe, boredom, disgust, being touched, and wonder. This new annotation database makes a contribution to the limited resources currently existing for movie emotion research.

Responding to our **fifth research question** (RQ5), we discover that aesthetic emotions felt by movie audiences cannot be described by two emotional dimensions only: arousal and valence. Thus, aesthetic emotions are more complex than "*everyday*" emotions and require at least four emotional dimensions to be characterized accurately. We can suppose that aesthetic emotions are mixtures of a few "*everyday*" emotions: anger, fear, happiness, surprise, sadness, etc. For example, awe is considered as a combination of surprise and fear. Answering our **sixth research question** (RQ6), we show that aesthetic highlights in movies evoke aesthetic emotions in spectators and the intensity of these emotions depends on aesthetic highlight category and movie genre. For example, theme development in horror movies is designed in a such way to gradually elicit fear, curiosity, and joy in spectators. Also, we find that aesthetic scene ratings of annotators are influenced by their personality. This suggests that personality characteristics moderate spectators' engagement with the movie story and the intensity of felt emotions. Spectators who are characterized by openness can feel more curiosity when movie story develops over time.

Moreover, we showed that movie aesthetic emotions could be predicted based on spectators' physiological and behavioural reactions. Also, we identify that the statistical features that describe dynamic changes in spectators' responses are crucial for aesthetic emotion recognition. Furthermore, we obtained the strongest correlation between aesthetic ratings

and prediction scores of DBN models for boredom, being touched, and disgust intensity using statistical features of EDA signals. The experiments that we conducted can serve as a baseline for future research on movie aesthetic emotion recognition.

Our future work includes extracting features of movie content to find a representation of movie aesthetic attributes changing over time. Moreover, a comprehensive approach to movie aesthetic emotion recognition will require to explore different fusion strategies to combine multimodal signals of spectators with descriptors of aesthetic movie content. Combining personality characteristics of spectators with their physiological and behavioural reactions could improve emotion recognition since personality provides information on a characteristic set of individuals' behaviours.

Chapter 7

Conclusions and perspectives

In this thesis we focused on the quantitative analysis of film aesthetic experience. We aimed to determine the discriminative power of spectators' physiological and behavioural reactions (EDA and ACC signals) or movie content based features regarding aesthetic highlight detection in movies, induced emotion recognition, and aesthetic emotion recognition. Firstly, we explored a wide range of emotions evoked by aesthetic highlights in full-length movies and detected aesthetic highlights based on synchronized spectators' EDA and ACC signals. Secondly, we studied the differences between induced and perceived emotions of movie audiences and recognized induced emotions from movie content features as well as EDA and ACC signals. Finally, we investigated the relationship between "*everyday*" emotions and aesthetic emotions as well as dependencies among aesthetic highlights, aesthetic emotions, personality, and spectators' physiological and behavioural reactions. In particular, we recognized aesthetic emotions felt by spectators from EDA and ACC signals.

7.1 Outcomes of the research

Chapter 1 introduced film aesthetic experience and our motivation to study it. Firstly, "*everyday*" and aesthetic emotions were defined and a clear distinction between induced and perceived emotions of movie audiences were made. Secondly, several emotion representations and emotion elicitation were discussed. Thirdly, the concept of interpersonal synchronization with regard to watching movies together was accounted for. Finally, research questions on film aesthetic experience were formulated, and the contribution of this thesis was summarized.

Chapter 2 provided an extensive literature review on aesthetic and affective content video analysis. Existing work on aesthetic and affect recognition as well as highlight detection from video content and spectators' reactions were presented and described with respect to different

criteria. The main limitations of the existing work were then discussed. It was highlighted that much research omitted movie context during stimulus selection as an important factor for emotion elicitation. It was also pointed out that emotion recognition models were trained and tested on small datasets without taking account the fact that emotions are gradually evoked over time. The quality of the emotional ground truth collected from small numbers of annotators was discussed. Consequently, it was emphasized that a large variance in the results was observed and thus fair comparisons of these emotion recognition models could not be made. Finally, approaches for overcoming all these limitations were proposed.

Chapter 3 described and discussed existing aesthetic and affective multimedia databases in details. The continuous LIRIS-ACCEDE database that was created to study film emotional experience in a movie theater was selected to study film aesthetic experience. Then, protocols for collecting annotations of aesthetic highlights in movies, perceived emotions and aesthetic emotions felt by movie audiences were described. Also, the statistical analysis of the annotations was provided. The results suggest that complex emotions like aesthetic emotions should be annotated by at least several annotators following an adapted protocol to reduce variance of annotations.

In Chapter 4, it was shown that aesthetic highlights in movies could elicit a wide range of emotions. The amount of these emotions (a level of arousal and valence intensity) strongly depends on the aesthetic highlight category and on the movie genre. Also, methodology and results of aesthetic highlight detection based on the level of synchronization among spectators' EDA and ACC measurements were presented. The results suggest that the level of synchronization among spectators' EDA and ACC signals is discriminative in the context of watching movies together. In particular, pairwise synchronization measures are stable measures of synchronization and achieved the best performance of aesthetic highlight detection independently of movie genre and highlight categories. Nevertheless, the level of synchronization among EDA measurements is more indicative of aesthetic highlights than the level of synchronization among ACC signals. This can be justified by the fact that aesthetic experience is mainly associated with a high level of arousal that can be observed in EDA measurements. Also, it is in line with our findings on aesthetic highlights evoked a wide range of emotions described by a high level of arousal in movie audiences. Synchronized behavioural reactions of spectators were supposed to be observed at a higher level. Movement ACC measurements only convey part of spectators' behavioural reactions. Behavioural signals, such as eye gaze, facial expression, and shoulder gesture can provide more information on spectators' synchronized behavioural responses. However, these behavioural signals could not be recorded because spectators watched movies in a darkened amphitheater.

In Chapter 5, the relationship between induced and perceived emotions of movie audiences was studied. As a result, an inconsistency in induced and perceived emotion annotations was observed. In particular, it was found that induced and perceived emotions of movie audiences were not always positively correlated. The complex dependencies among emotional dimensions were discovered. For example, perceived arousal and induced arousal were weakly negatively correlated. This suggests that watching many exciting scenes in movie can make spectators bored. This should be taken into account when stimuli for emotion induction are selected. Then, the relationship between the intensity of induced and perceived emotions of movie audiences, and the occurrences of aesthetic highlights in movies was studied. As a main result, we found that aesthetic highlights conveyed information on both perceived and induced emotions and could be used as high level affective features. Finally, induced emotions were recognized from spectators' EDA and ACC signals as well as movie content. LSTM-RNN models outperformed SVR and DBN models because their ability to take into account temporal information and hierarchically combine knowledge-inspired affective cues with audio-visual movie content and movie audience responses. The improvement of performance using multimodal hierarchical fusion leads us to the conclusion that adding other behavioural and physiological signals, e.g., facial expression, heart rate, and electroencephalogram signals of spectators could boost emotion recognition performance.

In Chapter 6, it was shown that aesthetic highlights in movies could evoke aesthetic emotions in movie audiences. Aesthetic emotions, such as awe, boredom, disgust, being touched, and wonder that are felt by movie audiences are associated with the category of aesthetic highlights as well as the movie genre. Aesthetic highlights elicit aesthetic emotions in movie audiences that are beyond "*everyday*" emotions. In fact, four emotional dimensions were found for movie aesthetic emotions. That is why movie aesthetic emotions cannot be accurately represented in the arousal-valence space like "*everyday*" emotions. Furthermore, the influence of personality on aesthetic emotions was discovered by measuring the differences in aesthetic scene ratings with regard to personality traits. Also, aesthetic emotions could be predicted based on spectators' EDA and ACC signals. Thus, spectators' reactions are discriminative for movie aesthetic emotion recognition.

7.2 Responding to research questions

We respond to the research questions described in Section 1.5, as follows:

- We show that aesthetic highlights evoke a wide range of emotions in movie audiences, studying the direct link between emotional dimensions (arousal-valence space) and the occurrences of aesthetic highlights (RQ1).

- We investigate several approaches to synchronization estimation to measure the amount of synchronization among multiple spectators' reactions and detect aesthetic highlights in movies. We show that pairwise synchronization measures are the most reliable to efficiently detect aesthetic highlights (RQ2).
- To the best of our knowledge, we carry out the first quantitative analysis of the relationship between perceived and induced emotions of movie audiences. We find that perceived and induced emotions of movie audiences are not always consistent and the dependencies among them are complex. We then use movie audiences' perceived emotions to predict their induced emotions (RQ3, RQ4).
- We determine that LSTM-RNN models outperform SVR and DBN models because their ability to take into account temporal information and hierarchically combine knowledge-inspired affective cues with audio-visual movie content and movie audience responses (RQ4).
- We show that aesthetic highlights elicit aesthetic emotions in movie audiences that are beyond "*everyday*" emotions. In fact, movie aesthetic emotions cannot be accurately represented in the arousal-valence space like "*everyday*" emotions (RQ5, RQ6).
- We find that the characteristics of personality influence aesthetic preferences (RQ6).

7.3 Lessons learned

We learned many lessons while carrying out this research on film aesthetic experience. We summarize them below:

- Understanding film aesthetic experience is an extremely difficult and challenging task since it is a weakly defined concept in art and is considered to correspond to subjective feelings of being engaged with a film. There is no one universal approach to analyze film aesthetic experience. However, it is important that any research should investigate together movie content with spectators' physiological and behavioural responses.
- Choosing a sufficiently large amount of emotional stimuli is essential to correctly elicit various emotions. Stimuli should be full-length movies to provide spectators movie context and thus evoke subtle and complex emotions. These movies should represent several different movie genres.
- Selecting an emotion representation for aesthetic and affective video content analysis is critically important and requires special attention when an experimental protocol

is designed. An emotion representation should be selected based on the goal of the studies. When the aim of research is to investigate a wide range of emotions, a dimensional representation is recommended. The number of emotional dimensions that accurately represent these emotions should be validated in preliminary studies. For example, aesthetic emotions felt by spectators cannot be entirely characterized in the arousal-valence space.

- Collecting emotional ground truth is one of the main challenges in aesthetic and affective video content analysis. It has to be specified which emotions annotators should report while performing data collection. In particular, a clear distinction between self-reports of induced emotions and self-reports of perceived emotions should be made in the experimental protocol.
- Intra/inter-spectator variability of physiological and behavioural reactions is observed. Taking into account this variability is required to build a robust individual-independent emotion recognition system.
- Multimodal fusion is needed since spectators' physiological and behavioural reactions as well as high-level affective movie content cues, such as aesthetic highlights, perceived emotions, and lexical features of movie dialogues are discriminative for induced emotion recognition.
- Recognizing emotions can benefit from taking into account temporal information because spectators' physiological and behavioural reactions to consecutive movie scenes have sequential structures.

7.4 Conclusions and perspectives

All these outcomes of our studies suggest that film aesthetic experience is complex and subjective. Film aesthetic experience is affected by many various factors, such as personality, life experience, mood, and interest that are difficult to objectively quantify. The conclusion can be made that film aesthetic experience cannot be investigated without taking into account multimodal reactions of movie audiences in naturalistic conditions, e.g., watching movies together in a movie theater.

As presented in this thesis, physiological and behavioural signals of spectators and video content are informative to study film aesthetic experience. There are several research questions on aesthetic and affect video content analysis that have arisen during the studies. These research questions can be summarized, as follows:

- Emotions can be evoked and recognized in different contexts. Emotions induced by controlled stimuli in laboratory conditions (watching short videos in a laboratory without any context) are different from naturalistic emotions elicited in ecological situations (watching full-length movies in a movie theater). Thus, there is a need to develop emotion recognition systems that are able to include these emotion elicitation contexts to improve emotion recognition performance. For example, when people interact with one another they have tendencies to reach similar emotion states through emotion contagion and feeling empathy. To determine emotions of one individual, affective states and behaviours of other individuals can be investigated.
- The relationship between induced and perceived emotions of individuals is very complex and high-level affective video cues contain information on both these types of emotions. That is why the automatic extraction of video content based features, for example, lexical features (e.g., CSA and DIS-NV features) of movie dialogues should be developed.
- Intra/inter-spectator variability of physiological and behavioural responses to the same stimuli is observed. It is important to learn a new representation of physiological and behavioural signals that reduces this variability and improve individual-independent emotion recognition systems.
- There is a lack of labelled instances to train advance machine learning. To overcome this limitation, transfer learning between different emotion recognition tasks could be investigated. Pretrained models on one emotion recognition task can be applied to other emotional challenges, e.g., induced emotion recognition of movie audiences can take the advantage of emotion recognition of individuals watching short videos.
- Fusion of different physiological and behavioural signals can be beneficial for emotion recognition because information on humans' emotional states is passed through multimodal channels. However, modality fusion at feature and decision level is not sufficient for emotion recognition since all multimodal signals do not have the same discriminative power and the dynamics of changes. Different strategies of hierarchical fusion that allows multimodal features to be incorporated at different stages of modelling are required.
- The improvement of existing annotation collection protocols is needed because there is the large variance of emotional annotations. Some emotions are extremely difficult to be annotated since they are very complex and subjective. The existence of the one absolute ground truth is questionable and it should be replaced by modelling ground

truth distribution. In particular, the emotional ground truth can be represented by the first four statistical moments of annotation distributions.

- Current approaches to emotion recognition mainly focus on the analysis of individuals' physiological and behavioural signals without including psychological factors, e.g., interest, personality, and mood. Personal interest and personality can influence the felt emotions of individuals and the perception of other individuals' emotions. Taking into account dependencies among emotional states, physiological and behavioural reactions, personality, mood, and interest can improve performance of emotion recognition systems.

References

- [1] Abadi, M. K., Subramanian, R., Kia, S. M., Avesani, P., Patras, I., and Sebe, N. (2015). DECAF: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222.
- [2] Acar, E., Hopfgartner, F., and Albayrak, S. (2014). Understanding affective content of music videos through learned representations. In *International Conference on Multimedia Modeling*, pages 303–314. Springer.
- [3] Akay, M. (1997). Wavelet applications in medicine. *IEEE spectrum*, 34(5):50–56.
- [4] AlZoubi, O., D’Mello, S. K., and Calvo, R. A. (2012). Detecting naturalistic expressions of nonbasic affect using physiological signals. *IEEE Transactions on Affective Computing*, 3(3):298–310.
- [5] Anastasia, T. and Leontios, H. (2016). AUTH-SGP in MediaEval 2016 emotional impact of movies task. In *MediaEval2016*.
- [6] Ancona, N., Marinazzo, D., and Stramaglia, S. (2004). Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, 70(5):056221.
- [7] Arnold, M. B. (1960). *Emotion and personality*. Columbia University Press.
- [8] Aviyente, S. (2005). A measure of mutual information on the time-frequency plane. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP’05). IEEE International Conference on*, pages IV–481. IEEE.
- [9] Bard, G. V. (2007). Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric. In *ACSW2007*, volume 68, pages 117–124. Australian Computer Society, Inc.
- [10] Baveye, Y., Bettinelli, J.-N., Dellandréa, E., Chen, L., and Chamaret, C. (2013). A large video database for computational models of induced emotion. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 13–18. IEEE.
- [11] Baveye, Y., Chamaret, C., Dellandréa, E., and Chen, L. (2017). Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing*.
- [12] Baveye, Y., Dellandréa, E., Chamaret, C., and Chen, L. (2015a). Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 77–83. IEEE.

- [13] Baveye, Y., Dellandrea, E., Chamaret, C., and Chen, L. (2015b). Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55.
- [14] Bazin, A. (2004). *What is cinema?* University of California Press.
- [15] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160.
- [16] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.
- [17] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."
- [18] Blinowska, K. J., Kuś, R., and Kamiński, M. (2004). Granger causality and information flow in multivariate processes. *Physical Review E*, 70(5):050902.
- [19] Blood, A. J. and Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences*, 98(20):11818–11823.
- [20] Bordwell, D., Thompson, K., and Ashton, J. (1997). *Film art: an introduction*. McGraw-Hill New York.
- [21] Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R. (2009). *Introduction to Meta-analysis*. John Wiley & Sons, Inc.
- [22] Bradley, A. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.
- [23] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.
- [24] Canini, L., Benini, S., and Leonardi, R. (2013). Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):636–647.
- [25] Carmeli, C., Knyazeva, M. G., Innocenti, G. M., and De Feo, O. (2005). Assessment of eeg synchronization based on state-space analysis. *Neuroimage*, 25(2):339–354.
- [26] Carvalho, S., Leite, J., Galdo-Álvarez, S., and Gonçalves, Ó. F. (2012). The emotional movie database (emdb): A self-report and psychophysiological study. *Applied psychophysiology and biofeedback*, 37(4):279–294.
- [27] Cavell, S. (1979). *The world viewed: Reflections on the Ontology of Film*. Harvard University Press.
- [28] Chanel, G., Ansari-Asl, K., and Pun, T. (2007). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 2662–2667. IEEE.

- [29] Chanel, G., Kronegg, J., Grandjean, D., and Pun, T. (2006). Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. In *International workshop on multimedia content representation, classification and security*, pages 530–537. Springer.
- [30] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.
- [31] Chaspari, T., Timmons, A. C., Baucom, B. R., Perrone, L., Baucom, K. J., Georgiou, P., Margolin, G., and Narayanan, S. S. (2017). Exploring sparse representation measures of physiological synchrony for romantic couples. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 267–272. IEEE.
- [32] Chen, L.-H., Hsu, H.-W., Wang, L.-Y., and Su, C.-W. (2011). Violence detection in movies. In *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pages 119–124. IEEE.
- [33] Chen, S. and Jin, Q. (2016). RUC at MediaEval 2016 emotional impact of movies task: Fusion of multimodal features. In *MediaEval2016*.
- [34] Chen, Y., Rangarajan, G., Feng, J., and Ding, M. (2004). Analyzing multiple nonlinear time series with extended granger causality. *Physics Letters A*, 324(1):26–35.
- [35] Chênes, C., Chanel, G., Soleymani, M., and Pun, T. (2013). Highlight detection in movie scenes through inter-users, physiological linkage. In *Social Media Retrieval*, pages 217–237.
- [36] Chepushtanova, S., Kirby, M., Peterson, C., and Ziegelmeier, L. (2015). An application of persistent homology on grassmann manifolds for the detection of signals in hyperspectral imagery. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, pages 449–452. IEEE.
- [37] Chevalier, P., Martin, J.-C., Isableu, B., and Tapus, A. (2015). Impact of personality on the recognition of emotion expressed via human, virtual, and robotic embodiments. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*, pages 229–234. IEEE.
- [38] Chollet, F. (2015). Keras. <https://github.com/fchollet/keras>.
- [39] Cirelli, L. K., Einarson, K. M., and Trainor, L. J. (2014). Interpersonal synchrony increases prosocial behavior in infants. *Developmental science*, 17(6):1003–1011.
- [40] Clark, H. H. (1996). *Using language*. Cambridge university press.
- [41] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- [42] Cohen, J. (1988a). *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge, 2 edition.
- [43] Cohen, J. (1988b). *Statistical power analysis for the behavioral sciences*. erlbaum. Hillsdale, NJ.

- [44] Coifman, R. R. and Lafon, S. (2006a). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30.
- [45] Coifman, R. R. and Lafon, S. (2006b). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30.
- [46] Cooper, J. M. and Silvia, P. J. (2009). Opposing art: Rejection as an action tendency of hostile aesthetic emotions. *Empirical Studies of the Arts*, 27(1):109–126.
- [47] Cover, T. M. (1991). Elements of information theory thomas m. cover, joy a. thomas copyright© 1991 john wiley & sons, inc. print isbn 0-471-06259-6 online isbn 0-471-20061-1.
- [48] Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety*. Jossey-Bass.
- [49] Csikszentmihalyi, M. (2014). *Toward a psychology of optimal experience*. Springer.
- [50] Cui, D., Liu, X., Wan, Y., and Li, X. (2010). Estimation of genuine and random synchronization in multivariate neural series. *Neural Networks*, 23(6):698–704.
- [51] Cupchik, G. C. (1992). From perception to production: A multilevel analysis of the aesthetic process. *Emerging visions of the aesthetic process: Psychology, semiology, and philosophy*, pages 61–81.
- [52] Cupchik, G. C. (1995). Emotion in aesthetics: Reactive and reflective models. *Poetics*, 23(1-2):177–188.
- [53] Cupchik, G. C., Vartanian, O., Crawley, A., and Mikulis, D. J. (2009). Viewing artworks: contributions of cognitive control and perceptual facilitation to aesthetic experience. *Brain and cognition*, 70(1):84–91.
- [54] Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., and Lang, P. J. (2000). Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biological psychology*, 52(2):95–111.
- [55] Dauwels, J., Vialatte, F., Musha, T., and Cichocki, A. (2010). A comparative study of synchrony measures for the early diagnosis of alzheimer’s disease based on eeg. *NeuroImage*, 49(1):668–693.
- [56] Dauwels, J., Vialatte, F. B., Rutkowski, T. M., and Cichocki, A. (2007). Measuring neural synchrony by message passing. In *NIPS*, pages 361–368.
- [57] David, B. and Thompson, K. (1994). *Film History: An Introduction*. New York: MacGraw-Hill.
- [58] Davidson, R. J., Sherer, K. R., and Goldsmith, H. H. (2009). *Handbook of affective sciences*. Oxford University Press.
- [59] Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365.

- [60] Deleuze, G. (1989). *Cinema 2: The time-image*, trans. hugh tomlinson and robert galeta. *London: Athlone*.
- [61] Deleuze, G., Tomlinson, H., and Habberjam, B. (1986). *The Movement-Image*. University of Minnesota.
- [62] Demarty, C.-H., Penet, C., Gravier, G., and Soleymani, M. (2012). A benchmarking campaign for the multimodal detection of violent scenes in movies. In *European Conference on Computer Vision*, pages 416–425. Springer.
- [63] Dietz, R. and Lang, A. (1999). Affective agents: Effects of agent affect on arousal, attention, liking and learning. In *Proceedings of the Third International Cognitive Technology Conference, San Francisco*.
- [64] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., et al. (2007). The humane database: addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer.
- [65] Ekman, P. (1999). Basic emotions. *handbook of cognition and emotion*, vol. 98.
- [66] Eyben, F., Weninger, F., Lehment, N., Schuller, B., and Rigoll, G. (2013a). Affective video retrieval: Violence detection in hollywood movies by large-scale segmental feature extraction. *PloS one*, 8(12):e78506.
- [67] Eyben, F., Weninger, F., Squartini, S., and Schuller, B. (2013b). Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 483–487. IEEE.
- [68] Eyben, F., Weninger, F., Wöllmer, M., and Schuller, B. (2016). open-source media interpretation by large feature-space extraction.
- [69] Eyben, F., Wöllmer, M., and Schuller, B. (2010). OpenSMILE: the munich versatile and fast open-source audio feature extractor. In *ICMI2010*, pages 1459–1462. ACM.
- [70] Feldman, R. (2007). Parent–infant synchrony: Biological foundations and developmental outcomes. *Current directions in psychological science*, 16(6):340–345.
- [71] Ferrer, E. and Helm, J. L. (2013). Dynamical systems modeling of physiological coregulation in dyadic interactions. *International Journal of Psychophysiology*, 88(3):296–308.
- [72] Fleureau, J., Guillotel, P., and Huynh-Thu, Q. (2012). Physiological-based affect event detector for entertainment video applications. *IEEE Transactions on Affective Computing*, 3(3):379–385.
- [73] Fleureau, J., Guillotel, P., and Orlac, I. (2013). Affective benchmarking of movies based on the physiological responses of a real audience. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 73–78. IEEE.

- [74] Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.
- [75] Frijda, N. H. (1989). *Aesthetic emotions and reality*. American Psychological Association.
- [76] Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1_suppl):123–147.
- [77] Gendron, M. and Feldman Barrett, L. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review*, 1(4):316–339.
- [78] Ghaemmaghami, P., Abadi, M. K., Kia, S. M., Avesani, P., and Sebe, N. (2015). Movie genre classification by exploiting meg brain signals. In *International Conference on Image Analysis and Processing*, pages 683–693. Springer.
- [79] Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75.
- [80] Gninkoun, G. and Soleymani, M. (2011). Automatic violence scenes detection: A multi-modal approach. In *In Working Notes Proceedings of the MediaEval 2011 Workshop*.
- [81] Gong, Y., Han, M., Hua, W., and Xu, W. (2004). Maximum entropy model-based baseball highlight detection and classification. *Computer Vision and Image Understanding*, 96(2):181–199.
- [82] Gordon, P. C. and Holyoak, K. J. (1983). Implicit learning and generalization of the "mere exposure" effect. *Journal of Personality and Social Psychology*, 45(3):492.
- [83] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438.
- [84] Greenwald, M. K., Cook, E. W., and Lang, P. J. (1989). Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of psychophysiology*, 3(1):51–64.
- [85] Gross, J. J. (1998). Antecedent-and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology*, 74(1):224.
- [86] Gunduz, A. and Principe, J. C. (2009). Correntropy as a novel measure for nonlinearity tests. *Signal Processing*, 89(1):14–23.
- [87] Hamm, J. and Lee, D. D. (2008). Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, pages 376–383. ACM.
- [88] Hanjalic, A. and Xu, L.-Q. (2005). Affective video content representation and modeling. *IEEE transactions on multimedia*, 7(1):143–154.
- [89] Hartigan, P. (1985). Algorithm as 217: Computation of the dip statistic to test for unimodality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3):320–325.

- [90] Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1994). *Emotional contagion*. Cambridge university press.
- [91] Helm, J. L., Sbarra, D., and Ferrer, E. (2012). Assessing cross-partner associations in physiological responses via coupled oscillator models. *Emotion*, 12(4):748.
- [92] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.
- [93] Hsu, L. (2009). *Visible and expression. study on the intersubjective relation between visual perception, aesthetic feeling, and pictorial form*. PhD thesis, Ph. D. dissertation, Ecole des Hautes Etudes en Sciences Sociales (EHESS).
- [94] Irie, G., Satou, T., Kojima, A., Yamasaki, T., and Aizawa, K. (2010). Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE Transactions on Multimedia*, 12(6):523–535.
- [95] Jalili, M., Barzegaran, E., and Knyazeva, M. G. (2014). Synchronization of eeg: Bivariate and multivariate measures. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(2):212–221.
- [96] Jelles, B., Scheltens, P., Van der Flier, W., Jonkman, E., da Silva, F. L., and Stam, C. (2008). Global dynamical analysis of the eeg in alzheimer’s disease: frequency-specific changes of functional interactions. *Clinical Neurophysiology*, 119(4):837–841.
- [97] John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.
- [98] Joho, H., Staiano, J., Sebe, N., and Jose, J. M. (2011). Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications*, 51(2):505–523.
- [99] Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM.
- [100] Kallinen, K. and Ravaja, N. (2006). Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 10(2):191–213.
- [101] Kang, H.-B. (2003). Affective content detection using hmms. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 259–262. ACM.
- [102] Karvonen, A., Kykyri, V.-L., Kaartinen, J., Penttonen, M., and Seikkula, J. (2016). Sympathetic nervous system synchrony in couple therapy. *Journal of marital and family therapy*, 42(3):383–395.
- [103] Katti, H., Yadati, K., Kankanhalli, M., and Tat-Seng, C. (2011). Affective video summarization and story board generation using pupillary dilation and eye gaze. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 319–326. IEEE.

- [104] Keren, G., Kirschstein, T., Marchi, E., Ringeval, F., and Schuller, B. (2017). End-to-end learning for dimensional emotion recognition from physiological signals. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 985–990. IEEE.
- [105] Khushaba, R. N., Kodagoda, S., Lal, S., and Dissanayake, G. (2011). Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Transactions on Biomedical Engineering*, 58(1):121–131.
- [106] Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3687–3691. IEEE.
- [107] Kipp, M. (2014). *Anvil: The video annotation research tool. Handbook of Corpus Phonology*. Oxford University Press, pages 420–436.
- [108] Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379.
- [109] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.
- [110] Koelstra, S., Yazdani, A., Soleymani, M., Mühl, C., Lee, J.-S., Nijholt, A., Pun, T., Ebrahimi, T., and Patras, I. (2010). Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos. In *International Conference on Brain Informatics*, pages 89–100. Springer.
- [111] Koestler, A. (1970). *The act of creation (Revised Danube Edition ed.)*. London: Pan Books.
- [112] Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2015a). Dynamic time warping of multimodal signals for detecting highlights in movies. In *Proceedings of the 1st Workshop on Modeling INTERPERSONAL Synchrony And influence*, pages 35–40. ACM.
- [113] Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2015b). Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*.
- [114] Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2015c). Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals. In *QoMEX2015*, pages 1–6. IEEE.
- [115] Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2017). Films, affective computing and aesthetic experience: identifying emotional and aesthetic highlights from multimodal signals in a social setting. *Frontiers in ICT*, 4:11.
- [116] Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

- [117] Kreitler, S. and Kreitler, H. (1984). Meaning assignment in perception. *Psychological processes in cognition and personality*, pages 173–191.
- [118] Kroupi, E., Vesin, J.-M., and Ebrahimi, T. (2013). Phase-amplitude coupling between eeg and eda while experiencing multimedia content. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 865–870. IEEE.
- [119] Kubovy, M. (1999). On the pleasures of the mind. *Well-being: The foundations of hedonic psychology*, pages 134–154.
- [120] Lachaux, J.-P., Rodriguez, E., Martinerie, J., Varela, F. J., et al. (1999). Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208.
- [121] Leder, H., Belke, B., Oeberst, A., and Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology*, 95(4):489–508.
- [122] Lench, H. C., Flores, S. A., and Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitation.
- [123] Levenson, R. W. and Gottman, J. M. (1983). Marital interaction: physiological linkage and affective exchange. *Journal of personality and social psychology*, 45(3):587.
- [124] Li, J., Wang, T., Hu, W., Sun, M., and Zhang, Y. (2006). Soccer highlight detection using two-dependence bayesian network. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1625–1628. IEEE.
- [125] Li, T., Baveye, Y., Chamaret, C., Dellandréa, E., and Chen, L. (2015). Continuous arousal self-assessments validation using real-time physiological responses. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 39–44. ACM.
- [126] Lickley, R. J. (2015). 20 fluency and disfluency. *The handbook of speech production*, page 445.
- [127] Liu, S., Rovine, M. J., Cousino Klein, L., and Almeida, D. M. (2013). Synchrony of diurnal cortisol pattern in couples. *Journal of Family Psychology*, 27(4):579.
- [128] Malandrakis, N., Potamianos, A., Evangelopoulos, G., and Zlatintsi, A. (2011). A supervised approach to movie emotion tracking. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2376–2379. IEEE.
- [129] Mandler, G., Mandler, J. M., Kremen, I., and Sholiton, R. D. (1961). The response to threat: Relations among verbal and physiological indices. *Psychological Monographs: General and Applied*, 75(9):1.
- [130] Marković, S. (2012). Components of aesthetic experience: aesthetic fascination, aesthetic appraisal, and aesthetic emotion. *i-Perception*, 3(1):1–17.
- [131] Maslow, A. H. (2013). *Toward a psychology of being*. Simon and Schuster.
- [132] Matthews, G., Deary, I. J., and Whiteman, M. C. (2003). *Personality traits*. Cambridge University Press.

- [133] Matthews, G., Jones, D. M., and Chamberlain, A. G. (1990). Refining the measurement of mood: The uwest mood adjective checklist. *British journal of psychology*, 81(1):17–42.
- [134] Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., and Gross, J. J. (2005). The tie that binds? coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175.
- [135] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.
- [136] Miles, L. K., Nind, L. K., and Macrae, C. N. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of experimental social psychology*, 45(3):585–589.
- [137] Miranda-Correa, J. A., Abadi, M. K., Sebe, N., and Patras, I. (2017). AMIGOS: A dataset for mood, personality and affect research on individuals and groups. *arXiv preprint arXiv:1702.02510*.
- [138] Müller, M. (2007). *Information retrieval for music and motion*, volume 2. Springer.
- [139] Muszynski, M., Kostoulas, T., Chanel, G., Lombardo, P., and Pun, T. (2015). Spectators’ synchronization detection based on manifold representation of physiological signals: Application to movie highlights detection. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 235–238. ACM.
- [140] Muszynski, M., Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2016). Synchronization among groups of spectators for highlight detection in movies. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 292–296. ACM.
- [141] Muszynski, M., Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2018). Aesthetic highlight detection in movies based on synchronization of spectators’ reactions. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(3):68.
- [142] Muszynski, M., Tian, L., Lai, C., Moore, J., Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2019). Recognizing induced emotions of movie audiences from multimodal information. *IEEE Transactions on Affective Computing*, in press.
- [143] Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.
- [144] Nunez, P. L. and Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA.
- [145] Penet, C., Demarty, C.-H., Gravier, G., and Gros, P. (2012). Multimodal information fusion and temporal integration for violence detection in movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2393–2396. IEEE.
- [146] Penet, C., Demarty, C.-H., Gravier, G., and Gros, P. (2015). Variability modelling for audio events detection in movies. *Multimedia Tools and Applications*, 74(4):1143–1173.

- [147] Perea, J. A., Deckard, A., Haase, S. B., and Harer, J. (2015). Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC bioinformatics*, 16(1):257.
- [148] Perea, J. A. and Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics*, 15(3):799–838.
- [149] Pereda, E., Quiroga, R. Q., and Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in neurobiology*, 77(1):1–37.
- [150] Perugini, M. and Di Blas, L. (2002). Analyzing personality related adjectives from an eticemic perspective: the big five marker scales (bfms) and the italian ab5c taxonomy. *Big Five Assessment*, pages 281–304.
- [151] Plantinga, C. (2012). Art moods and human moods in narrative cinema. *New Literary History*, 43(3):455–475.
- [152] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1(3-31):4.
- [153] Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.
- [154] Quiroga, R. Q., Arnhold, J., and Grassberger, P. (2000). Learning driver-response relationships from synchronization patterns. *Physical Review E*, 61(5):5142.
- [155] Ray, R. D. (2007). Emotion elicitation using films. *Handbook of emotion elicitation and assessment*, pages 9–28.
- [156] Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30.
- [157] Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, 53(1-2):23–69.
- [158] Rulkov, N. F., Sushchik, M. M., Tsimring, L. S., and Abarbanel, H. D. (1995). Generalized synchronization of chaos in directionally coupled chaotic systems. *Physical Review E*, 51(2):980.
- [159] Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.
- [160] Saito, N., Kuginuki, T., Yagyu, T., Kinoshita, T., Koenig, T., Pascual-Marqui, R. D., Kochi, K., Wackermann, J., and Lehmann, D. (1998). Global, regional, and local measures of complexity of multichannel electroencephalography in acute, neuroleptic-naive, first-break schizophrenics. *Biological psychiatry*, 43(11):794–802.
- [161] Sánchez-Meca, J. and Marín-Martínez, F. (2015). Meta-analysis in psychological research. *International Journal of Psychological Research*, 3(1):150–162.

- [162] Saxbe, D. and Repetti, R. L. (2010). For better or worse? coregulation of couples' cortisol levels and mood states. *Journal of personality and social psychology*, 98(1):92.
- [163] Schaefer, A., Nils, F., Sanchez, X., and Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172.
- [164] Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.
- [165] Schuller, B. W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., Narayanan, S. S., et al. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Interspeech*, volume 2010, pages 2795–2798.
- [166] Shah, R. R., Yu, Y., and Zimmermann, R. (2014). Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 607–616. ACM.
- [167] Sigari, M.-H., Soltanian-Zadeh, H., and Pourreza, H.-R. (2015). Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference. *International Journal of Computer Graphics*, 6(1).
- [168] Silvia, P. J. (2009). Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 3(1):48.
- [169] Silvia, P. J. and Brown, E. M. (2007). Anger, disgust, and the negative aesthetic emotions: Expanding an appraisal model of aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, 1(2):100.
- [170] Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., Schedl, M., Demarty, C.-H., and Chen, L. (2015). The mediaeval 2015 affective impact of movies task. In *MediaEval*.
- [171] Soleymani, M., Asghari-Esfeden, S., Fu, Y., and Pantic, M. (2016). Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28.
- [172] Soleymani, M., Chanel, G., Kierkels, J. J., and Pun, T. (2008). Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 228–235. IEEE.
- [173] Soleymani, M., Chanel, G., Kierkels, J. J., and Pun, T. (2009a). Affective characterization of movie scenes based on content analysis and physiological changes. *International Journal of Semantic Computing*, 3(02):235–254.
- [174] Soleymani, M., Kierkels, J. J., Chanel, G., and Pun, T. (2009b). A bayesian framework for video affective representation. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE.

- [175] Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012a). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.
- [176] Soleymani, M. and Pantic, M. (2013). Multimedia implicit tagging using eeg signals. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE.
- [177] Soleymani, M., Pantic, M., and Pun, T. (2012b). Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223.
- [178] Song, P., Ou, S., Zheng, W., Jin, Y., and Zhao, L. (2016). Speech emotion recognition using transfer non-negative matrix factorization. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5180–5184. IEEE.
- [179] Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., and Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5688–5691. IEEE.
- [180] Subramanian, R., Wache, J., Abadi, M., Vieriu, R., Winkler, S., and Sebe, N. (2016). Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*.
- [181] Sun, K. and Yu, J. (2007). Video affective content representation and recognition using video affective tree and hidden markov models. In *International Conference on Affective Computing and Intelligent Interaction*, pages 594–605. Springer.
- [182] Sun, M., Farhadi, A., and Seitz, S. (2014). Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pages 787–802. Springer.
- [183] Takens, F. (1981). Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, 898:366–381.
- [184] Tan, E. S.-H. (1995). Film-induced affect as a witness emotion. *Poetics*, 23(1-2):7–32.
- [185] Tarvainen, J., Laaksonen, J., and Takala, T. (2017). Computational and perceptual determinants of film mood in different types of scenes. In *Multimedia (ISM), 2017 IEEE International Symposium on*, pages 185–192. IEEE.
- [186] Tarvainen, J., Laaksonen, J., and Takala, T. (2018). Film mood and its quantitative determinants in different types of scenes. *IEEE Transactions on Affective Computing*.
- [187] Tarvainen, J., Sjöberg, M., Westman, S., Laaksonen, J., and Oittinen, P. (2014). Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments. *IEEE Transactions on Multimedia*, 16(8):2085–2098.
- [188] Tarvainen, J., Westman, S., and Oittinen, P. (2013). Stylistic features for affect-based movie recommendations. In *International Workshop on Human Behavior Understanding*, pages 52–63. Springer.

- [189] Tellegen, A. and Atkinson, G. (1974). Openness to absorbing and self-altering experiences ("absorption"), a trait related to hypnotic susceptibility. *Journal of abnormal psychology*, 83(3):268.
- [190] Tian, L., Moore, J., and Lai, C. (2016). Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. *SLT2016*.
- [191] Tian, L., Moore, J. D., and Lai, C. (2015). Emotion recognition in spontaneous and acted dialogues. In *ACII2015*, pages 698–704. IEEE.
- [192] Tian, L., Muszynski, M., Lai, C., Moore, J. D., Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2017). Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same? In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 28–35. IEEE.
- [193] Timmons, A. C., Chaspari, T., Han, S. C., Perrone, L., Narayanan, S. S., and Margolin, G. (2017). Using multimodal wearable technology to detect conflict among couples. *Computer*, 50(3):50–59.
- [194] Timmons, A. C., Margolin, G., and Saxbe, D. E. (2015). Physiological linkage in couples and its implications for individual and interpersonal functioning: A literature review. *Journal of Family Psychology*, 29(5):720.
- [195] Vidrascu, L. and Devillers, L. (2005). Detection of real-life emotions in call centers. In *Ninth European Conference on Speech Communication and Technology*.
- [196] Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79.
- [197] Wang, H. L. and Cheong, L.-F. (2006). Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704.
- [198] Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.
- [199] Wiley, N. (2003). Emotion and film theory. In *Studies in Symbolic Interaction*, pages 169–187. Emerald Group Publishing Limited.
- [200] Wundt, W. (1893). *Grundzüge der physiologischen psychologie leipzig*, 1893.
- [201] Xu, M., Jin, J. S., Luo, S., and Duan, L. (2008). Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 677–680. ACM.
- [202] Yadati, K., Katti, H., and Kankanhalli, M. (2014). Cavva: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 16(1):15–23.
- [203] Yamashita, T., Tanaka, M., Yoshida, E., Yamauchi, Y., and Fujiyoshii, H. (2014). To be Bernoulli or to be Gaussian, for a Restricted Boltzmann Machine. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1520–1525. IEEE.

- [204] Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., and Guo, B. (2015). Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4633–4641.
- [205] Yao, T., Mei, T., and Rui, Y. (2016). Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990.
- [206] Zeki, S. and Nash, J. (1999). *Inner vision: An exploration of art and the brain*, volume 415. Oxford University Press Oxford.
- [207] Zhang, S., Huang, Q., Jiang, S., Gao, W., and Tian, Q. (2010). Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 12(6):510–522.
- [208] Zhao, S., Yao, H., Sun, X., Jiang, X., and Xu, P. (2013). Flexible presentation of videos based on affective content analysis. In *MMM (1)*, pages 368–379.
- [209] Zhou, F., Qu, X., Jiao, J. R., and Helander, M. G. (2014). Emotion prediction from physiological signals: A comparison study between visual and auditory elicitors. *Interacting with computers*, 26(3):285–302.

Appendix A

List of publications

Journal papers

1. **Muszynski, M.**, Tian, L., Lai, C., Moore, J., Kostoulas, T., Lombardo, P., Pun T., and Chanel, G. (2019). Recognizing induced emotions of movie audiences from multimodal information. *IEEE Transactions on Affective Computing*, (in press).
2. **Muszynski, M.**, Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2018). Aesthetic highlight detection in movies based on synchronization of spectators' reactions. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14.3:68.
3. Kostoulas, T., Chanel, G., **Muszynski, M.**, Lombardo, P., and Pun, T. (2017). Films, affective computing and aesthetic experience: Identifying emotional and aesthetic highlights from multimodal signals in a social setting. *Frontiers in ICT* 4:1-11.

International conferences with refereed papers

1. Tian, L., **Muszynski, M.**, Lai, C., Moore, J., Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2017). Recognizing induced emotions of movie audiences: are induced and perceived emotions the same? (**Best Paper Nominee**). *7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 23-26 October 2017, San Antonio, USA.
2. **Muszynski, M.**, Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2016). Synchronization among groups of spectators for highlight detection in movies. (**Best**

- Paper Nominee**). *ACM Multimedia 2016*, 15-19 October 2016, Amsterdam, The Netherlands.
3. **Muszynski, M.**, Kostoulas, T., Chanel, G., Lombardo, P., and Pun, T. (2015). Spectators' synchronization detection based on manifold representation of physiological signals: application to movie highlights detection. *17th International Conference on Multimodal Interaction*, 9-13 November 2015, Seattle, Washington, USA.
 4. Kostoulas, T., Chanel, G., **Muszynski, M.**, Lombardo, P., and Pun, T. (2015). Dynamic time warping of multimodal signals for detecting highlights in movies. *17th International Conference on Multimodal Interaction: First International Workshop on Modeling INTEPERsonal SynchrONy*, 13 November 2015, Seattle, Washington, USA.
 5. Kostoulas, T., Chanel, G., **Muszynski, M.**, Lombardo, P., and Pun, T. (2015). Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals. *7th International Workshop on Quality of Multimedia Experience*, 26-29 May 2015, Costa Navarino, Messinia, Greece.

Invited presentations

1. Recognition of movie audiences' induced emotions using multimodal machine learning models. Lawrence Berkeley National Laboratory, University of California, 3 November 2017, Berkeley, CA, USA.
2. How do we feel about what we see?: multimodal machine learning based models for recognizing induced emotions of movie audiences. Institute for Computational and Mathematical Engineering, Stanford University, 2 November 2017, Stanford, CA, USA.
3. Induced and perceived emotions in film viewing. Interdisciplinary Workshop on Literature, Film and the Emotions, University of Geneva, 19 July 2017, Geneva, Switzerland.

Other conferences

1. **Muszynski, M.**, Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2018). Film, aesthetics and emotions - interdisciplinary studies. *Annual Research Forum*, 7 February 2018, Swiss Center for Affective Sciences, University of Geneva, Switzerland (poster).

2. **Muszynski, M.**, Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2017). Evaluation of synchronization measures for aesthetic highlight detection in movies. *NCCR Affective Sciences - International Conference on Emotions*, 18-19 May 2017, Swiss Center for Affective Sciences, University of Geneva, Switzerland (poster).
3. **Muszynski, M.**, Kostoulas, T., Chanel, G., Lombardo, P., and Pun, T. (2015). Spectators' synchronization detection based on physiological signals: application to movie highlights detection. *Annual Research Forum*, 3-4 March 2016, Swiss Center for Affective Sciences, University of Geneva, Switzerland (poster, same as the poster in September 2015).
4. **Muszynski, M.**, Kostoulas, T., Chanel, G., Lombardo, P., and Pun, T. (2015). Spectators' synchronization detection based on manifold representation of physiological signals: application to movie highlights detection. *BBL/CIBM Research day*, 28 September 2015, Brain-and-Behavior Laboratory, University of Geneva, Switzerland. *Inauguration of the Swiss Doctoral School in Affective Sciences*, 15 September 2015, University of Geneva, Switzerland (poster).
5. Kostoulas, T., Chanel, G., **Muszynski, M.**, Lombardo, P., and Pun, T. (2015). Identifying aesthetic highlights in movies from physiological and behavioral synchronization. *International Society for Research on Emotions (ISRE) Conference 2015*, 7-11 July 2015, Geneva, Switzerland (abstract, poster).

