Automatic prediction of emotions induced by movies

Yoann Baveye

Thèse

pour obtenir le grade de

DOCTEUR DE L'ECOLE CENTRALE DE LYON
spécialité "Informatique"

préparée au
LIRIS

Automatic Prediction of Emotions Induced by Movies

Ecole Doctorale InfoMaths

Thèse soutenue le 12/11/2015 par

Yoann BAVEYE

devant le jury composé de :

Prof.	Björn Schuller	University of Passau	(Rapporteur)
Prof.	Thierry Pun	Université de Genève	(Rapporteur)
Prof.	Patrick Le Callet	Université de Nantes	(Examinateur)
Dr.	Mohammad Soleymani	Université de Genève	(Examinateur)
Prof.	Liming Chen	Ecole Centrale de Lyon	(Directeur de thèse)
Dr.	Emmanuel Dellandréa	Ecole Centrale de Lyon	(Co-encadrant)
M^me^	Christel Chamaret	Technicolor	(Co-encadrante)

ÉCOLE CENTRALE LYON

UNIVERSITÉ DE LYON

Acknowledgements

First of all, I wish to thank Christel Chamaret for giving me the opportunity, after two internships, to work on this thesis in Technicolor. She gave me her trust and has always been there to support me. Thanks to Liming Chen and Emmanuel Dellandréa, who also supervised this thesis, for their guidance and the numerous discussions that we had, among others, during the video conferences.

I am grateful to the members of my jury, Prof. Björn Schuller, Prof. Thierry Pun, Prof. Patrick Le Callet, and Dr. Mohammad Soleymani, for accepting to be part of the committee, and for their feedback regarding my thesis.

I would also like to thank all my colleagues in the numerous teams I have been involved with at Technicolor and the Ecole Centrale de Lyon, for all the stimulating discussions, but also for the fun times. Speaking about fun times, I especially want to thank Baptiste, Jean-Noël and Laetitia, Jérémie, Lucille, Marianne and Thomas (and their little Chloé). Thanks also to Xingxian and Ting who helped me during this thesis. Of course, I would also like to thank all film-makers sharing their amazing works under Creative Commons licenses.

On a more personal note, I would like to thank all my friends who shared with me these years (and for some of them, for decades!). Je tiens donc à vous remercier, en français, pour tous ces incroyables moments que j'ai passés avec vous et qui m'ont aidés à supporter ces trois années intensives. Merci à Amel (Big bowl) et Louis, Anaëlle, Anaïs et Adrien, Arnaud, Brandy, Carine et Solène (et notamment son aide linguistique), Charlenn et Glenn, Clément, Clémence, Fanny et Flavien, Mélie et Léo (l'escargot), sans oublier Neric (avec un né et un ric) et Laurie !

Je voudrais aussi évidemment remercier ma famille, notamment ma mère, mon père, ma soeur, ma grand-mère, et Jérémy qui ont eu le courage de me supporter, et sans qui rien de ceci n'aurait été possible.

Pour finir, je voudrais dédier cette thèse à Matthieu (avec deux "t", n'est-ce pas? ;), qui est parti brutalement, et bien trop tôt.

Merci.

Abstract

Never before have movies been as easily accessible to viewers, who can enjoy anywhere the almost unlimited potential of movies for inducing emotions. Thus, knowing in advance the emotions that a movie is likely to elicit to its viewers could help to improve the accuracy of content delivery, video indexing or even summarization. However, transferring this expertise to computers is a complex task due in part to the subjective nature of emotions. The present thesis work is dedicated to the automatic prediction of emotions induced by movies based on the intrinsic properties of the audiovisual signal.

To computationally deal with this problem, a video dataset annotated along the emotions induced to viewers is needed. However, existing datasets are not public due to copyright issues or are of a very limited size and content diversity. To answer to this specific need, this thesis addresses the development of the LIRIS-ACCEDE dataset. The advantages of this dataset are threefold: (1) it is based on movies under Creative Commons licenses and thus can be shared without infringing copyright, (2) it is composed of 9,800 good quality video excerpts with a large content diversity extracted from 160 feature films and short films, and (3) the 9,800 excerpts have been ranked through a pair-wise video comparison protocol along the induced valence and arousal axes using crowdsourcing. The high inter-annotator agreement reflects that annotations are fully consistent, despite the large diversity of raters' cultural backgrounds.

Three other experiments are also introduced in this thesis. First, affective ratings were collected for a subset of the LIRIS-ACCEDE dataset in order to cross-validate the crowdsourced annotations. The affective ratings made also possible the learning of Gaussian Processes for Regression, modeling the noisiness from measurements, to map the whole ranked LIRIS-ACCEDE dataset into the 2D valence-arousal affective space. Second, continuous ratings for 30 movies were collected in order develop temporally relevant computational models. Finally, a last experiment was performed in order to collect continuous physiological measurements for the 30 movies used in the second experiment. The correlation between both modalities strengthens the validity of the results of the experiments.

Armed with a dataset, this thesis presents a computational model to infer the emotions induced by movies. The framework builds on the recent advances in deep learning and takes into account the relationship between consecutive scenes. It is composed of two fine-tuned Convolutional Neural Networks. One is dedicated to the visual modality and uses as input crops of key frames extracted from video segments, while the second one is dedicated to the audio modality through the use of audio spectrograms. The activations of the last fully connected layer of both networks are con-

catenated to feed a Long Short-Term Memory Recurrent Neural Network to learn the dependencies between the consecutive video segments. The performance obtained by the model is compared to the performance of a baseline similar to previous work and shows very promising results but reflects the complexity of such tasks. Indeed, the automatic prediction of emotions induced by movies is still a very challenging task which is far from being solved.

Keywords: Computational emotion modeling; Induced emotion; Video dataset; Affective computing; Crowdsourcing.

Résumé

Jamais les films n'ont été aussi facilement accessibles aux spectateurs qui peuvent profiter de leur potentiel presque sans limite à susciter des émotions. Savoir à l'avance les émotions qu'un film est susceptible d'induire à ses spectateurs pourrait donc aider à améliorer la précision des systèmes de distribution de contenus, d'indexation ou même de synthèse des vidéos. Cependant, le transfert de cette expertise aux ordinateurs est une tâche complexe, en partie due à la nature subjective des émotions. Cette thèse est donc dédiée à la détection automatique des émotions induites par les films, basée sur les propriétés intrinsèques du signal audio-visuel.

Pour s'atteler à cette tâche, une base de données de vidéos annotées selon les émotions induites aux spectateurs est nécessaire. Cependant, les base de données existantes ne sont pas publiques à cause de problèmes de droit d'auteur ou sont de taille restreinte. Pour répondre à ce besoin spécifique, cette thèse présente le développement de la base de données LIRIS-ACCEDE. Cette base a trois avantages principaux: (1) elle utilise des films sous licence Creative Commons et peut donc être partagée sans enfreindre le droit d'auteur, (2) elle est composée de 9,800 extraits vidéos de bonne qualité qui proviennent de 160 films et courts métrages, et (3) les 9,800 extraits ont été classés selon les axes de "valence" et "arousal" induits grâce un protocole de comparaisons par paires mis en place sur un site de crowdsourcing. L'accord inter-annotateurs élevé reflète la cohérence des annotations malgré la forte différence culturelle parmi les annotateurs.

Trois autres expériences sont également présentées dans cette thèse. Premièrement, des scores émotionnels ont été collectés pour un sous-ensemble de vidéos de la base LIRIS-ACCEDE dans le but de faire une validation croisée des classements obtenus via crowdsourcing. Les scores émotionnels ont aussi rendu possible l'apprentissage d'un processus gaussien par régression, modélisant le bruit lié aux annotations, afin de convertir tous les rangs liés aux vidéos de la base LIRIS-ACCEDE en scores émotionnels définis dans l'espace 2D valence-arousal. Deuxièmement, des annotations continues pour 30 films ont été collectées dans le but de créer des modèles algorithmiques temporellement fiables. Enfin, une dernière expérience a été réalisée dans le but de mesurer de façon continue des données physiologiques sur des participants regardant les 30 films utilisés lors de l'expérience précédente. La corrélation entre les annotations physiologiques et les scores continus renforce la validité des résultats de ces expériences.

Equipée d'une base de données, cette thèse présente une modèle algorithmique afin d'estimer les émotions induites par les films. Le système utilise à son avantage les récentes avancées dans le domaine de

l'apprentissage profond et prend en compte la relation entre des scènes consécutives. Le système est composé de deux réseaux de neurones convolutionnels ajustés. L'un est dédié à la modalité visuelle et utilise en entrée des versions recadrées des principales frames des segments vidéos, alors que l'autre est dédié à la modalité audio grâce à l'utilisation de spectrogrammes audio. Les activations de la dernière couche entièrement connectée de chaque réseau sont concaténées pour nourrir un réseau de neurones récurrent utilisant des neurones spécifiques appelés "Long-Short-Term-Memory" qui permettent l'apprentissage des dépendances temporelles entre des segments vidéo successifs. La performance obtenue par le modèle est comparée à celle d'un modèle basique similaire à l'état de l'art et montre des résultats très prometteurs mais qui reflètent la complexité de telles tâches. En effet, la prédiction automatique des émotions induites par les films est donc toujours une tâche très difficile qui est loin d'être complètement résolue.

Mots-clés : Modèle d'estimation des émotions ; Emotions induites ; Base de donnée de vidéos ; Affective computing ; Crowdsourcing.

Publications

The work conducted in this thesis has resulted in a number of peer-reviewed publications in journals and conference proceedings. Patents have also been filed and are currently examined by patent offices.

Journal articles

- **Y. Baveye**, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, pp. 43–55, Jan 2015.

Peer-reviewed international conferences and workshops

- **Y. Baveye**, J.-N. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret, "A large video data base for computational models of induced emotion," in *Affective Computing and Intelligent Interaction (ACII)*, pp. 13–18, 2013.
- **Y. Baveye**, E. Dellandréa, C. Chamaret, and L. Chen, "From crowdsourced rankings to affective ratings," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, Jul. 2014.
- **Y. Baveye**, C. Chamaret, E. Dellandréa, and L. Chen, "A protocol for cross-validating large crowdsourced data: The case of the LIRIS-ACCEDE affective video dataset," in *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia, ser. CrowdMM '14*, pp. 3–8, 2014.
- M. Sjöberg, **Y. Baveye**, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen, "The MediaEval 2015 Affective Impact of Movies Task," in *MediaEval 2015 Workshop*, 2015.
- T. Li, **Y. Baveye**, C. Chamaret, E. Dellandréa, and L. Chen, "Continuous Arousal Self-assessments Validation Using Real-time Physiological Responses," in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia (ASM '15)*, pp. 39–44, 2015.
- **Y. Baveye**, E. Dellandréa, C. Chamaret, and L. Chen, "Deep Learning *vs.* Kernel Methods: Performance for Emotion Prediction in Videos," in *Affective Computing and Intelligent Interaction (ACII)*, 2015.

Contents

List of Figures

List of Tables

Notations

ACCEDE	Annotated Creative Commons Emotional DatabasE
ANOVA	ANalysis Of VAriance
AV	AudioVisual
CC	Creative Commons
CNN	Convolutional Neural Network
CPM	Component Process Model
EEG	ElectroEncephaloGram
GP	Gaussian Process
GSR	Galvanic Skin Response
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
MSE	Mean-Square Error
MCD	Minimum Covariance Determinant
MLDA	Multi-Level Data-Augmentation
MFCC	Mel-Frequency Cepstral Coefficients
PAD	Pleasure-Arousal-Dominance
PCA	Principal Component Analysis
RBF	Radial Basis Function
RNN	Recurrent Neural Network
RVM	Relevance Vector Machine
SAM	Self-Assessment Manikin
SECs	Stimulus Evaluation Checks
SD	Standard Deviation
SRCC	Spearman's Rank Correlation Coefficient
SVM	Support Vector Machine
SVR	Support Vector Regression
VA	Valence-Arousal

Introduction

1

Contents

First and foremost, I would like to start this introduction with a citation from G. M. Smith who reminds us in [7] that:

> "Films do not "make" people feel. A better way to think of filmic emotions is that films extend an invitation to feel in particular ways. Individuals can accept or reject the invitation. Those who accept the invitation can accept in a variety of ways, just as people invited to a party can participate in very different activities."

1.1 Context and motivation

According to legend, when "L'arrivée d'un train en gare de La Ciotat" directed and produced by Auguste and Louis Lumière was first screened in 1896 (Figure 1.1), the audience was so terrified at the sight of the oncoming locomotive that people screamed and tried to hide under their seats. Nowadays, film-lovers are more hardened but still enjoy the almost unlimited potential of movies for inducing emotions. And with the impressive movie collection available online through popular on-demand streaming media websites such as Netflix [1] or M-GO [2], increasing day after day, film spectators can feel everywhere a large variety of emotions.

Under these circumstances, knowing in advance the emotions that a movie is likely to elicit to its viewers is highly beneficial; not only to improve the accuracy for video indexing, and summarization (*e.g.* [8], [9]), but also for mood-based personalized content delivery [10]. While major

1. https://www.netflix.com/
2. http://www.mgo.com/

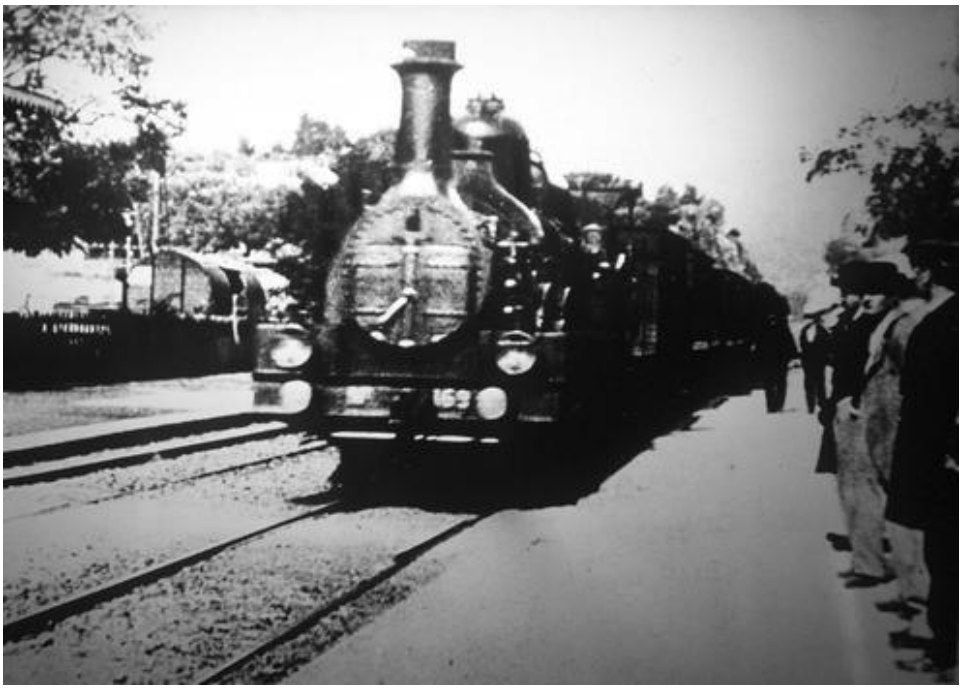


Figure 1.1 – *L'arrivée d'un train en gare de La Ciotat/Arrival of a Train at La Ciotat directed and produced by Auguste and Louis Lumière (1895)*

progress has been achieved in computer vision for visual object detection, scene understanding and high level concept recognition, a natural further step is modeling and recognition of affective concepts. This explains why the affective video content analysis research topic has emerged and attracted more and more attention during these last years.

1.2 Novelty and challenges

Affective video content analysis aims at automatic recognition of emotions elicited by videos. This has received increasing interest from research communities, *e.g.*, computer vision, machine learning, with an overall goal of endowing computers with human-like perception capabilities. Among the three perspectives dealing with affect in multimedia (namely, intended, induced, and expected emotions, each related to specific emotion detection [3]), this thesis focuses on the induced, *i.e.*, felt emotions. The induced emotions are the emotions that arise as a result of the content in most of its audience.

However, while human affective perception is highly subjective, machine-based affective modeling and recognition require large amounts of reliable ground truth data for training and testing. Unfortunately, the subjective nature of "induced emotions" makes it hard to collect consistent and large volumes of affective annotations suitable for the use as ground truth, while the copyright issues concerning video clips prevent free distribution of existing annotated datasets. Most state of the art work uses a private dataset of a very limited size and content diversity, thus making

3. A more comprehensive description of these three perspectives can be found in Section 3.1.

fair comparisons and results reproducibility impossible, and preventing achievement of major strides in the field. To overcome the limitations of the existing affective video datasets and foster research in affective video content analysis, our first objective is to create a dataset consisting of a large number of good quality video excerpts, with a large content diversity, and that can be freely distributed without copyright issues.

Armed with a dataset, computational models of induced emotions can be learned. Most of approaches for affective video content analysis have so far featured a standard architecture. Low-level audiovisual features that are known to be related to the emotions induced by movies are first extracted. Then, the features are used to feed and train a machine learning model to finally predict the desired affective score. The performance of state of the art emotion prediction models stagnates at moderate levels. On the other hand, mainly due to the advances of deep learning, the performances in scene and object recognition have been recently progressing intensively. Our second objective is to outperform the performance of standard baselines by benefiting from the recent breakthroughs in deep learning.

1.3 Approach and contributions

As discussed above, affective video content analysis faces two challenges. Creating a large and robust dataset that can be shared among researchers is the first challenge to solve in order to be able to improve the performance of the computational models to estimate induced emotions, which is the second challenge. Our contributions mainly focus on these two aspects and can be summarized as follows.

1. Our first contribution lies in the public release[4] of the LIRIS-ACCEDE dataset [11, 12]. LIRIS-ACCEDE has been designed to bypass the size and scope of related limitations of existing datasets for affective video content analysis.
   - We have extracted 9,800 good quality video excerpts with a large content diversity from 160 feature films and short films. All excerpts are shared under Creative Commons licenses and can thus be freely distributed without copyright issues. The 9,800 short segmented video clips last between 8 and 12 seconds and have been automatically segmented using a robust cut and fade in/out detection so that it is very likely that each segment be perceived by users as semantically coherent.
   - We have collected affective annotations for the induced arousal and valence axes using crowdsourcing through a pair-wise video comparison protocol, thereby ensuring that annotations are fully consistent, as testified by a high inter-annotator agreement, despite the large diversity of raters' cultural backgrounds.
   - To enable fair comparison and landmark progresses of future affective computational models using the LIRIS-ACCEDE dataset in different ways, we have provided four reproducible experimental protocols (*i.e.*, predefined training/validation/test sets,

4. Available at: http://liris-accede.ec-lyon.fr/.

leave-one-movie-out, leave-one-genre-out, and performance per movie).
- We have designed another experiment to cross-validate the crowdsourced annotations and to map the ranked database into the 2D valence-arousal affective space. The converted affective ratings make possible the comparison of the excerpts with other video clips annotated with absolute valence and arousal values.
- We have collected continuous ratings for 30 movies to develop psychologically relevant computational models taking into account the fact that previous scenes may reasonably influence the emotion inference of future ones.
- Finally, we have also designed an experiment to collect physiological annotations to extend the range of applications of the dataset.

2. Our second contribution concerns the development of a computational model to estimate the emotions induced by movies.
   - We introduced and evaluated a baseline, similar to what can be found in the state of the art, for the prediction of induced emotions using a large set of both visual and audio features. The features described include, but are not limited to, colorfulness, harmonization energy, length of scene cuts, audio zero-crossing rate, and Mel-Frequency Cepstral Coefficients (MFCC).
   - We have also developed and evaluated an audiovisual spatio-temporal model in order to outperform the performance of the baseline. The architecture of this model is based on two Convolutional Neural Networks (CNN), one dedicated to the visual modality and the other to the audio modality through spectrograms. Finally, a Long-Short-Term-Memory Recurrent Neural Network is used to model the dynamic of emotions induced by movies which is a key aspect reflecting the essential nature of emotions.

1.4 Organization

The remaining of this thesis is organized as follows:

Chapter 2 introduces the main concepts used throughout this thesis. In particular, this chapter defines the concepts of aesthetic emotions and emotion induction, as well as the Valence-Arousal (VA) dimensional representation of emotions.

Chapter 3 presents previous discrete and continuous affective video content analysis work, including video emotion recognition and violence detection, and shows the limitations of current existing affective video databases.

Chapter 4 describes the LIRIS-ACCEDE dataset created to overcome the limitations of existing datasets and foster research in affective video content analysis.

Chapter 5 evaluates the reliability of the ratings of the LIRIS-ACCEDE dataset through a cross-validation with affective scores collected in a new experiment.

Chapter 6 describes another experimental protocol in order to collect continuous affective self-assessments to make possible the learning of models for long movies where previous scenes may reasonably influence the emotions induced by future ones.

Chapter 7 gives an introduction to physiological measurements and in particular to the galvanic skin response.

Chapter 8 presents baselines for discrete and continuous affective movie analysis and describes reproducible protocols using the LIRIS-ACCEDE dataset to enable fair comparison between future work.

Chapter 9 develops the spatio-temporal framework to automatically estimate the continuous affective curves of movies. This framework is composed of two Convolutional Neural Networks dedicated to specific modalities and a Recurrent Neural Network to learn the temporal dependencies between consecutive scenes.

The last Chapter finally concludes the thesis, portrays the use of LIRIS-ACCEDE dataset by the research community, and explores directions for future work.

Part I

State of the art

"Remember ! Souviens-toi, prodigue ! Esto memor !
(Mon gosier de métal parle toutes les langues.)"

— Charles Baudelaire, *Les Fleurs du mal*

The Emotions: A Psychological Insight

2

Contents

As stated by Scherer [13]:

> "The number of scientific definitions [of the concept of emotions] proposed has grown to the point where counting seems quite hopeless (Kleinginna and Kleinginna already reviewed more than one hundred in 1981)."

Kleinginna and Kleinginna proposed a consensus by suggesting the following definition considered as one of the most comprehensive definitions of emotions [14]:

> "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."

2.1 Main psychological approaches

Decades of research driven by psychologists interested in the comprehension of emotions lead to three theoretical major approaches: "basic", "appraisal", and "psychological constructionist" [15].

2.1.1 Basic emotions: Facial expression and emotion

Basic emotion theorists have been originally inspired by the ideas expressed by Darwin. In *The Expression of the Emotions in Man and Animals*, Darwin demonstrated the universal nature of facial expressions. Thus, basic emotion models assume that several emotions are automatically triggered by objects and situations in the same way everywhere in the world. This idea of universality was reused by Ekman and Friesen to develop the Facial Action Coding System, which is a technique to score all observable facial movements [16]. Ekman considers emotions to be discrete states that are associated to facial expressions. Consequently, he postulates that there are a fixed number of emotions. The list of basic emotion defined by Ekman is often used in the literature to represent categorically the emotions. The basic emotions defined by Ekman in [17] include fear, anger, sadness, disgust, joy, and surprise.

2.1.2 Appraisal processes: the Component Process Model

Appraisal models assume that emotions are triggered by the interpretation of stimulus events and thus can be seen as relevance detectors [18].

The Component Process Model (CPM) is Scherer's major contribution to the study of emotions. The CPM postulates that the emotion process is a psychological construct driven by subjective appraisal and is the consequence of synchronized changes in five components corresponding to five distinctive functions [1]: cognitive appraisal (evaluation of objects and events), physiological activation (system regulation), motivational tendencies (preparation and direction of action), motor expression (communication of reaction and behavioral intention), and subjective feeling state (monitoring of internal state and external environment). Figure 2.1 shows the general architecture of the CPM organized into three modules: appraisal, response patterning, and integration/categorization.

The CPM considers the experience of emotions as the result of the recursive multilevel sequential evaluation checking of an emotional stimulus or event. It is a response to the evaluation of a stimulus relevant to major concerns of the organism. The first module, *i.e.*, the appraisal module, is the most important element in the CPM. It determines if an emotion episode is elicited and its characteristics based on four major appraisal objectives:

- Relevance — scan for salient events requiring attention
- Implication — estimation of implication, consequences
- Consequences — assessment of coping potential, *i.e.*, actions
- Significance — compatibility with social norms and self-concept

These objectives are evaluated based on stimulus evaluation checks (SECs) defined as novelty, intrinsic pleasantness, relevance to goals and needs, cause, probable outcomes, failure to meet expectations, conduciveness to

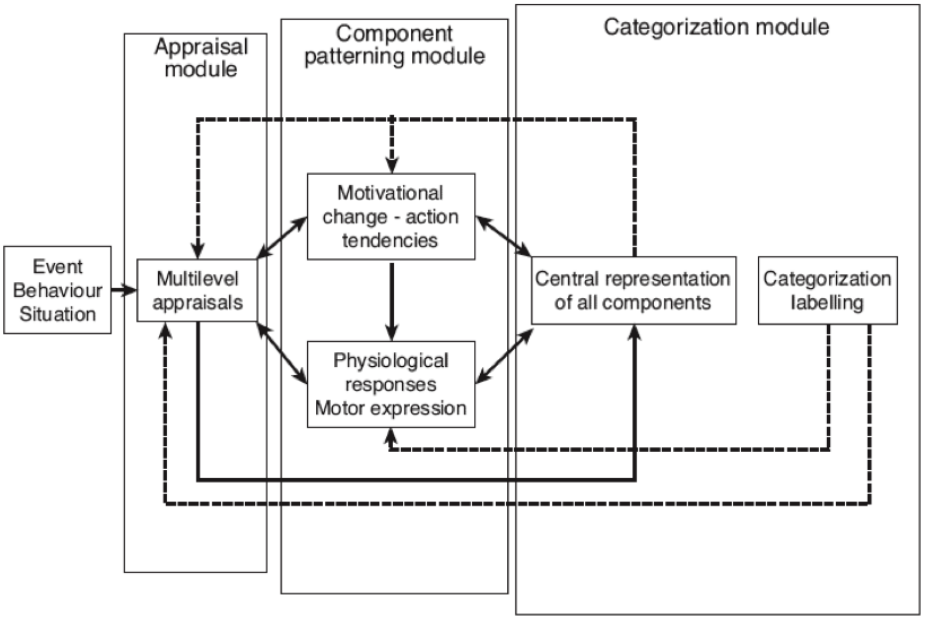


Figure 2.1 – *Basic architecture of Scherer's Component Process Model of emotion (originally published in [1])*

goals and need, urgency, control, power, adjustment, internal standards, and external standards. The results of the SECs are highly subjective and are biased by individual differences, moods, cultural values, group pressures, or other context factors [19]. This is why a stimulus may evoke different emotions to different people at different places: it is the evaluation of the events, rather than the events, which determines the characteristics of an emotional episode.

The result of the appraisal updates the motivational state that existed before the occurrence of the emotional stimulus or event. Both the appraisal results and motivational changes affect the automatic nervous system and the somatic nervous system. All these components are continuously updated as events and appraisal change. Scherer claims that, even if most of these components are examined unconsciously through parallel processing, some components may be evaluated consciously to allow more controlled regulation processes [1].

2.1.3 A psychological constructionist model: the Core Affect

Psychological constructionist models presume that emotions can be broken down into primitives that are also involved in other mental states [15]. Contrarily to appraisal models which assume that it is the evaluation of the stimulus that determines the characteristics of an emotional episode, psychological constructionist models assume that an emotion emerges when one's internal state is consciously understood in relation to an event.

Russell describes emotions as a part of a general component process called Core Affect [20]. Core affect is a primitive, universal and simple neurophysiological state that controls moods, when core affect is experienced as free-floating, and emotions, when core affect can be attributed

to some cause, whether based on reality or fiction. In the core affect model, the raw feelings (the conscious experience) is a blend of two dimensions: pleasure–displeasure (called pleasure or valence), and activation–deactivation (called arousal or energy). Core affect is a continuous introspection, mental but not cognitive or reflective.

Russell claims that an emotional episode is an event that counts as a member of an emotion category. The prototypical emotional episode begins with an antecedent event perceived in terms of its affective quality. The core affect is altered and attributed to the antecedent to make possible the perceptual cognitive processing of the antecedent with various manifestations: instrumental action, physiological and expressive changes, subjective conscious experiences, and emotional meta-experience. The emotional meta-experience is a self-perception: the emotional episode is noticed by the person who categorizes the emotional episode. Emotion categories are subjective since they are mental representation of emotions.

For Russell, emotions respond to the continuous flow of events (whether based on reality or fiction) and are influenced at the same time by the background environment (weather, odors, noise) and the social environment [21].

2.1.4 Summary of the main psychological approaches

The three main psychological approaches introduced in the previous sections, namely the basic, appraisal, and psychological constructionist models, share common ideas.

All the three approaches emphasize that emotions are constructed from universal basic biological or psychological parts. However, psychological constructionist and appraisal models differ from basic models to the extent that they assume that the social context of the situation and/or cultural differences have an effect on the experience of emotions [15].

Psychological constructionist models and appraisal models both consider emotion as an act of making meaning. However, the meaning analysis differs for both approaches. Psychological constructionist models, including Russell's Core Affect [20], assume that an emotion arises when a person's internal state is understood. For appraisal models, emotions are intentional states created by the evaluation of an original stimulus, and not by the internal state of the body. The internal state is only affected by this meaning analysis.

2.2 Emotions in multimedia

The theories introduced in the previous section model the emotions regardless of the significant stimulus that provoked the emotions. The stimulus could either be a natural phenomenon, the behavior of other people or animals, or even one's own behavior [13]. However, we are interested in this thesis in the emotions that are induced by movies. Thus, we focus in this section on a particular type of emotions: the aesthetic emotions.

2.2.1 Aesthetic emotions

An aesthetic emotion is an emotional response to a work of art (paintings, pictures, but also songs, movies,...) which can be described with three characteristics [22]:

- Persons are involved, in the state of intense attention engagement, and are strongly focused on a particular object.
- The viewer appraises the aesthetic objects as parts of virtual reality and finally, has a strong feeling of unity with the object of aesthetic appraisal.
- Aesthetic emotions are not oriented towards the satisfaction of bodily needs.

In the literature, the aesthetic information processing models share a common architecture: the emotional experience starts with a stimulus input, continues with the appraisal of the stimulus (from low level characteristics to deeper memorial instances) and ends with the evaluative judgment of the stimulus and the emotional response.

Linking aesthetics and emotion, Leder *et al.* proposed a model of aesthetic and emotional experience of modern art in an artistic context [2]. The model, depicted in Figure 2.2, is composed of five stages resulting in two distinct outputs: an aesthetic judgement which is the result of the evaluation of the cognitive mastering stage and an aesthetic emotion based on the affective state of satisfaction. Like Scherer, the model proposed by Leder postulates both automatic and deliberate processing.

Each of the five stages is concerned with different cognitive analyses. Processing during the first two stages is automatic and implicit. Features such as complexity, contrast, symmetry, order, and grouping are perceptually analyzed on the perceptual analysis level. With the second stage, an unconscious recall based on past experience begins. This implicit memory integration stage is also influenced by features such as familiarity, prototypicality, and peak-shift principle. Then, during the last three stages, processing takes place consciously and forms a feedback-loop. The explicit classification and cognitive mastering, based on the analysis of the style, content, and interpretation of the work of art, are also influenced by previous evaluations that have not been subjectively experienced as successful. Finally, the evaluation stage guides the aesthetic processing by measuring its success and enters into social interaction discourses which is going to be the input of another artistic evaluation.

However, the model of Leder is too simplified and only considers aesthetic experience as affectively positive and self-rewarding (successful processing). It does not take into account the negative aspect, the antinomy, and the variety of art experience [23]. Several works extended Leder's model to offer a more precise model of aesthetic processing. For example, Marković created a model in which the aesthetic experience is closer to arousal, *i.e.*, the interest for the work of art, than other dimensions of subjective experience [22]. In this multimodal model, the narrative content and the composition form both influence the aesthetic emotion (films being more focused on the narrative content).

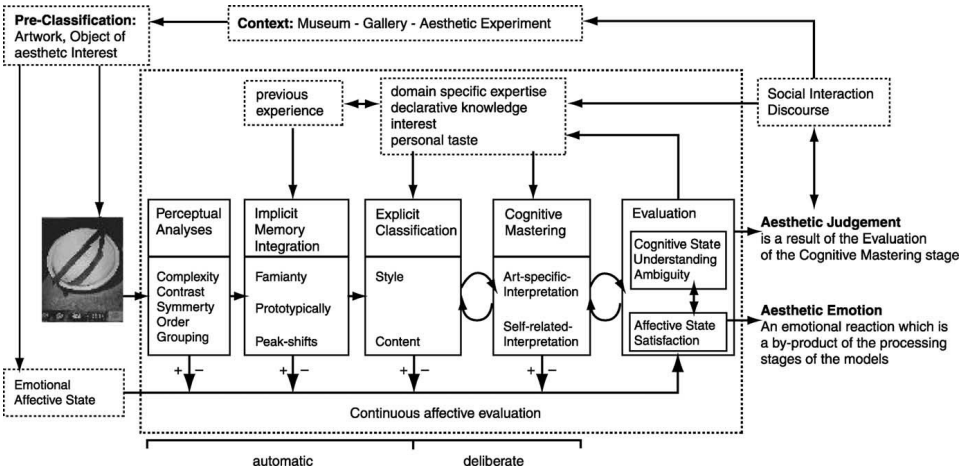


Figure 2.2 – *Leder's model of aesthetic and emotional experience (originally published in [2])*

2.2.2 A sociological perspective

Aesthetic emotions, and more specifically, emotions provoked by movies, can also be analyzed from a sociological perspective. Hochschild was one of the first to study the sociology of emotions and how shared norms can influence the way we want to try to feel emotions in given social relations [24].

Wiley discussed the differences between the emotions people experience watching a movie, called "movie emotions" and those occurring in our everyday life [25]. He explained that viewers split attention between the movie and the physical environment which helps regulate distance from the movie and creates an aura of safety and that the effects of movie emotions tend to end with the movie or at least decrease. Furthermore, the viewer is not the subject of the movie emotions, they happen because an identification is built with the character. For Wiley, movie emotions come with clear labels since narratives are written with precise emotional scripts. They are included in the movie with dialogues, clearly structured situations, transparent tendencies and musical cues. Movie emotions can be anticipated since films tend to follow the usual stability-instability-stability rule and because the music is geared to place the viewer in the appropriate emotional channel. These arguments show that it makes sense to investigate computational models based on multimodal features to predict the emotions provoked by movies.

For Wiley, movie emotions are desired: in watching a movie, the viewer wants to feel frequent, dense and almost wall-to-wall emotions [25]. This may result for example from a need to escape boredom or to forget the troubles of the day. Movie emotions are also quite intense: the viewers are dealing with more emotions than in a comparable period of time of a typical day. Movie emotions can be increased with social aspects (laughs of co-viewers). It is also easier to admit and talk about movie emotions because they are just fantasy so nobody can be blamed for having them.

2.2.3 Types of emotional processes in response to multimedia

There are three types of emotional processes in response to multimedia: emotion induction, emotion contagion, and empathic sympathy [26].

The induced emotions are the emotions that viewers feel in response to a multimedia content with respect to their goals and values. For example, in the dataset presented in Chapter 4, the animation movie "Big Buck Bunny" features an evil squirrel tearing up a butterfly. The situation is likely to elicit negative emotions to the viewers (disgust, anger,...), although the imaginary squirrel is enjoying the situation. The negative response from the viewers, *i.e.* the induced emotion, is due to their perception of the context according to their goals and values biased by the identification built with one or several characters of the movie.

With the emotional contagion process, the viewer is affected by the expressed emotion from a multimedia content without understanding in detail how the emotional expression of the multimedia content may have been developed. This process has to be distinguished from emotion perception, which refers to the perception of emotions expressed by the multimedia content without evoking affective responses in the viewers. Last but not least, empathic sympathy occurs when the viewers are not affected by the situation or event directly, but follow the appraisal steps leading to the emotion experienced by the characters in the multimedia content.

Finally, emotions, as defined above, have to be distinguished from other affective phenomena such as feelings or moods. A feeling is a subjective experience of an emotional episode whereas moods are diffuse affect states generally of low intensity, may last over hours or even days, and is often not clearly linked to an event or specific appraisals [13].

2.3 Representations

Diverse representations for emotions have been proposed in the literature. They are derived from the theories introduced in Section 2.1.

2.3.1 Categorical

The categorical emotions approach is very natural since it goes back to the origin of language and the emergence of words and expressions representing clearly separable states. Many discrete categorizations of emotions have been proposed, such as the six basic universal emotions proposed by Ekman in [17], already introduced in Section 2.1.1, or the eight primary emotions defined by Plutchik [27]. Plutchik suggested eight emotions, namely anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. Plutchik theorized that these basic emotions are biologically primitive and have evolved in order to increase the reproductive fitness of the animal. This categorical representation faces a granularity issue since the number of emotion classes is too small in comparison with the diversity of emotion perceived by film viewers. In case the number of classes is increased, ambiguities due to language difficulties or personal interpretation appear.

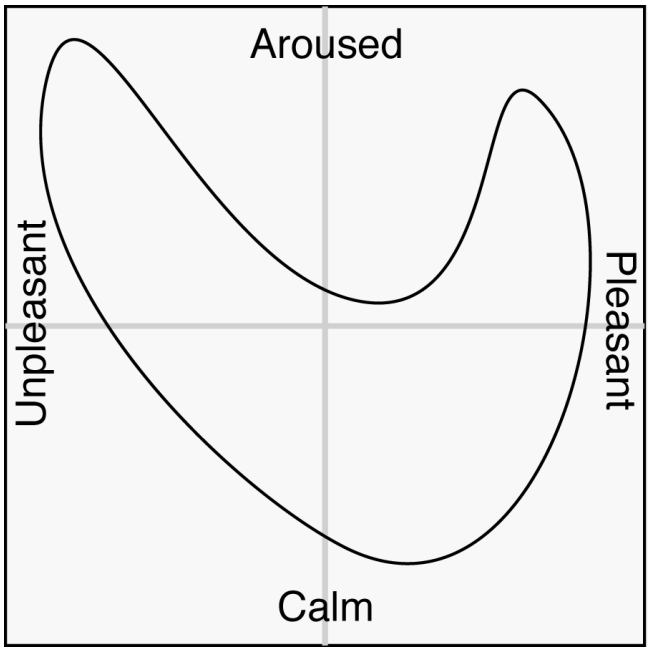


Figure 2.3 – *Illustration of the parabolic 2D affect space (originally published in [3])*

2.3.2 Dimensional

Dimensional approaches have also been proposed to model the emotions as points in a continuous n-dimensional space. The most famous one is the valence-arousal-dominance space, also known as pleasure-arousal-dominance (PAD) space introduced by Russell and Mehrabian [28] and extensively used in researches dealing with affective understanding. In this space, each subjective feeling can be described by its position in a three-dimensional space formed by the dimensions of valence, arousal, and dominance. Valence ranges from negative (*e.g.*, sad, disappointed) to positive (*e.g.*, joyous, elated), whereas arousal can range from inactive (*e.g.*, tired, pensive) to active (*e.g.*, alarmed, angry), and dominance ranges from dominated (*e.g.*, bored, sad) to in control (*e.g.*, excited, delighted). Given the difficulty of consistently identifying a third dimension (such as dominance, tension or potency) which differs from arousal, many studies, including this work, limit themselves to the valence and arousal (VA) dimensions. Indeed, especially when dealing with emotions induced by videos, valence and arousal account for most of the independent variance [29, 30]. Moreover, psychophysiological experiments have revealed that only certain areas of this two-dimensional space are relevant [3] and that emotions induced by media can be mapped onto a parabolic space created by the arousal and valence axes (see Figure 2.3).

However, this common two-dimensional space is questioned by Fontaine *et al.* who demonstrated that two dimensions are not sufficient to satisfactorily represent emotions [31]. They showed that using at least four dimensions is more appropriate to represent the diversity of emotions (valence, arousal, dominance, and predictability) but that the optimal number of dimensions to be included in a model depends on the purpose of the model.

Summary

This chapter introduced the main concepts used throughout this thesis. Several psychological models of emotion, and more specifically of aesthetic emotion, were first presented. Then, the types of emotional processes in response to multimedia were described and in particular the induced emotions. Finally, discrete and dimensional representations of emotions were introduced.

Due to the objectives of this thesis defined in Chapter 1, several concepts defined in this chapter are particularly important. First, we focus in this thesis on the very specific type of emotions called aesthetic emotions, defined in Section 2.2.1. More precisely, since we are interested in the emotions that are felt by the viewers, we focus on a specific type of emotional processes in response to multimedia content called emotion induction defined in Section 2.2.3. To model the induced emotions, we will use in this thesis the universal 2-dimensional VA representation described in Section 2.3.2. Finally, for the design of our computational model, we keep from these psychological models that the evaluation of an emotion is an iterative process, and that aesthetic emotions are influenced by low-level concepts (*e.g.*, complexity, contrast or symmetry of the stimulus) as well as higher-level concepts (e.g., style, content).

Computational Models and Datasets for Affective Video Content Analysis

3

Contents

Affective video content analysis aims at automatically predicting the emotions elicited by videos. Work on affective video analysis can be categorized into two subgroups: continuous affective video content analysis, which estimates an affective score for consecutive portions (*e.g.* each frame or group of frames) of a video, and discrete affective video content analysis, which assigns an affective score to a video. Some work represents emotions in the 2D valence-arousal space or in the 3D valence-arousal-dominance space, while other work represents emotions using discrete categories. Furthermore, the models are sometimes dedicated to specific video categories, *i.e.* music videos or a particular movie genre.

There are also studies on emotion assessment using physiological signals beyond audiovisual features. However, this topic is out of the scope of this thesis, although we demonstrate correlations between perceived emotions and physiological signals in Chapter 7.

3.1 Affective video content analysis: the perspectives

The video investigated with respect to its affective content can be either the stimulus eliciting emotions, it is called "affective movie content analysis", or a tool to investigate the emotions expressed by agents ("video emotion recognition"). In this thesis we will focus on affective movie content analysis but both affective video content analysis fields are related and will be presented in this chapter, with emphasis on previous affective movie content analysis work.

Video emotion recognition work aims at automatically estimating the emotion expressed by an agent from a video recording, in relation to an emotion induced by a stimulus. Sometimes, these emotions are not real but played by an actor [32]. The main goal of video emotion recognition models is to make possible affective interaction between human beings and computers. Such models are inherently different from affective movie content analysis models. However, the temporal modeling of emotions of such models may be useful for designing new continuous affective movie content analysis frameworks. This is why video emotion recognition work is discussed in Section 3.3.2.

Affective movie content analysis work focuses on the video which is the stimulus of the emotion to be investigated. The emotion to be investigated, related to the induced emotion defined in Chapter 2, is defined by the perspective of the models. There are three perspectives for affective movie content analysis work, each related to specific emotion detection: **intended**, **induced** and **expected** emotion. The *intended* emotion is the emotion the film maker wants to induce to the viewers. The *induced* emotion is the emotion that viewers feel in response to the movie. The *expected* emotion is the expected value of experienced (*i.e.* induced) emotion in a population.

Due to the exciting new possibilities offered by such affective computing techniques, they can be naturally applied to help standard multimedia systems [33]. Thus, affective movie content analysis work has a large number of applications, including mood based personalized content recommendation [34] or video indexing [35], and efficient movie visualization and browsing [36]. Beyond the analysis of existing video material, affective computing techniques can also be used to generate new content, *e.g.*, movie summarization [37], or personalized soundtrack recommendation to make user-generated videos more attractive [38]. Affective techniques have even been used to enhance the user engagement with advertising content by optimizing the way ads are inserted inside videos [39].

3.2 Affective multimedia databases

Creation of an affective database is a necessary step in affective computing studies. While there are many databases composed of facial expression videos for emotion recognition, there are not many databases of video clips annotated according to the emotions they induce in viewers.

Philippot [40], as well as Gross and Levenson [41], were the first to propose small sets of film excerpts assumed to elicit specific emotions in the laboratory. To achieve this goal, they selected specific excerpts most likely

Table 3.1 – *Downloadable video datasets annotated using labels considering induced emotion*

Name	Size	Emotional labels
HUMAINE	50 clips from 5 seconds to 3 minutes long	Wide range of labels at a global level (emotion-related states, context labels, key events, emotion words, *etc.*) and frame-by-frame level (intensity, arousal, valence, dominance, predictability, *etc.*)
FilmStim	70 film excerpts from 1 to 7 minutes long	24 classification criteria: subjective arousal, positive and negative affect, a positive and negative affect scores derived from the Differential Emotions Scale, six emotion discreteness scores and 15 mixed feelings scores
DEAP	120 one-minute music videos	Ratings from an online self-assessment on arousal, valence and dominance and physiological recordings with face video for a subset of 40 music videos
MAHNOB-HCI	20 film excerpts from 35 to 117 seconds long	Emotional keyword, arousal, valence, dominance and predictability combined with facial videos, EEG, audio, gaze and peripheral physiological recordings
EMDB	52 non-auditory film clips of 40 seconds long	Global ratings for the induced arousal, valence, dominance dimensions
VIOLENT SCENES DATASET	25 full-length movies	Annotations include the list of the movie segments containing physical violence according to two different definitions and also include 10 high-level concepts for the visual and audio modalities (presence of blood, fights, gunshots, screams, *etc.*)
LIRIS-ACCEDE	9,800 excerpts from 8 to 12 seconds long and 30 full movies	Discrete rankings and ratings for arousal and valence dimensions, continuous arousal and valence self-assessments, and continuous physiological recordings

to elicit strong emotions, which thus do not represent the full range of emotions that movies can potentially elicit. Even if increased efforts have recently been made to standardize film clip databases, there are no multimedia databases annotated along induced emotional axes dealing with the full spectrum of emotions in movies that are large enough to be used in machine learning and that do not suffer from copyright infringement.

The HUMAINE database [42] created by Douglas-Cowie *et al.* consists of a subset of three naturalistic and six induced reaction databases. The purpose of the database is to illustrate key principles of affective computing instead of applying it to machine learning. It is made up of 50 clips: naturalistic and induced data ranging from 5 seconds to 3 minutes. These have been annotated according to a wide range of labels detailed in Table 3.1.

Introduced by Schaefer *et al.* in [43], the FilmStim database consists of 70 film excerpts intended to elicit emotional states in experimental psychology experiments. 10 films are selected per emotional category (*i.e.* anger, sadness, fear, disgust, amusement, tenderness and neutral state) and cut into clips ranging from 1 to 7 minutes. 364 participants rated each film clip, and ranking scores were computed for 24 classification criteria displayed in Table 3.1. Even if it is one of the biggest databases of videos annotated along induced emotional labels, videos are labeled globally. Yet, emotions are a relatively fast phenomenon lasting a few seconds from onset to end [44]. This is why a unique global label is not sufficient to build ground truth data for induced emotion models.

The DEAP database is another publicly available database that has been created recently by Koelstra *et al.* [45]. It is composed of 120 one-minute long excerpts of music videos. Each one was rated by at least 14 volunteers from an online self-assessment based on induced arousal, valence and dominance. Physiological signals were recorded from participants, while they rated a subset of 40 of the above music videos in terms of arousal, valence, like/dislike, dominance and familiarity levels. Music videos protected by copyright are not available alongside the annotations. Instead, the YouTube links are given, but some of them are no longer available on YouTube, sometimes due to copyright claims. This shows the need for a database that does not depend on third parties to share its material legally.

The same year, Soleymani *et al.* released MAHNOB-HCI [46] which is a multimodal database composed of 20 short emotional excerpts extracted from commercially produced movies and video websites. These stimuli were selected in order to elicit 5 emotions (disgust, amusement, joy, fear and sadness). Participants watching these fragments were asked to annotate their own emotive state on a scale in terms of arousal and valence. Facial videos, electroencephalograms (EEG), audio, gaze and peripheral physiological recordings were also recorded for all 30 participants.

Carvalho *et al.* built in [47] the emotional movie database (EMDB) made up of 52 non-auditory film clips. Film clips are extracted from commercial films and last 40 seconds. They have been selected to cover the entire affective space. 113 participants rated each film clip in terms of induced valence, arousal and dominance on a 9-point scale. Non-auditory clips were used to enhance the scope for future experimental manipula-

tions. However, this clearly modifies how viewers perceive the video clips. Furthermore, multimodal processing is not possible in this case.

Still more recently, the Violent Scene Dataset was made available by Demarty *et al.* [48]. This is a collection of ground truth annotations based on extraction of violent events in movies, together with high level audio and video concepts. This dataset has been used since 2011 in the MediaEval multimedia benchmarking affect task "Violent Scenes Detection". Violent scene detection and prediction of induced emotions are clearly related since they are both part of the affective content analysis field. Violent scenes are most likely to be highly arousing and elicit negative emotions. Due to copyright issues, the 25 annotated movies cannot be delivered alongside the annotations. However, the links to the DVDs used for the annotation on the Amazon web site are provided.

Last but not least, it is worth mentioning the MIT dataset dedicated to animated GIFs [49]. Such kinds of short video footage are becoming increasingly popular by means of social networks. They are so widely adopted that the MIT team is currently and seriously working on predicting perceived emotions from such media support.

All these datasets, summarized in Table 3.1, either have different emotional labels or are not representative of the whole range of emotions in movies. Most of them only give global annotated emotions while they should be typically time dependent. Thus, a huge database of videos annotated using induced emotional labels potentially suitable for research is a requirement of the affective computing community.

3.3 Continuous affective video content analysis

This section presents previous work on continuous affective video content analysis, including movie content analysis and video emotion recognition. To predict or classify emotions, previous work either directly combines linearly audiovisual (AV) features extracted from the data, or either uses machine learning models. The temporal information can be included in the machine learning models, for example using Long Short-Term Memory neural networks, or by simply applying a temporal smoothing to the predicted values. A summary is given in Table 3.2.

3.3.1 Continuous valence and arousal movie content analysis

Hanjalic and Xu pioneered in [50] the analysis of affective movie content by directly mapping video features onto the valence-arousal space to create continuous representations. Based on film theorists work, they selected low-level features that are known to be related to arousal or valence such as motion intensity, shots lengths, and audio features (loudness, speech rate, rhythm, ...). They manually designed the functions modeling arousal and valence for consecutive frames based on the selected features and used a Kaiser window to temporally smooth the resulting curves. However, they only offered a qualitative evaluation of their model.

Soleymani *et al.* introduced a Bayesian framework for video affective representation [51] using audiovisual features and textual features

Table 3.2 – *Summary of previous work on continuous affective movie content analysis and video emotion recognition*

Authors	Method	Output	Ground truth	Annotators	Result
Hanjalic and Xu [50]	No classification: AV features are linearly combined	Continuous induced arousal and valence functions modeled at a frame level	Unknown number of movie scenes and soccer television broadcasts	None	Qualitative evaluation only
Soleymani *et al.* [51]	Bayesian framework using AV and textual features, and relying on temporal priors	Videos categorized into three induced emotional classes	21 full-length popular movies	1	Accuracy: 64%, F1 measure: 63%
Malandrakis *et al.* [52]	Two HMMs using AV features at frame level	Time series of 7 categories interpolated into continuous intended arousal or valence curves	30-min video clips from 12 movies	7	Correlation for arousal: 0.54, and valence: 0.23
Nicolaou *et al.* [53]	LSTM-RNN-based framework using facial expression, shoulder gesture, and audio cues	Continuous recognition of emotions expressed by actors in terms of arousal and valence	10 hours of footage from the SAL database capturing the interaction between a human and an operator	4	Correlation for arousal: 0.642, and valence: 0.796
Kahou *et al.* [54]	Combination of multiple deep neural networks for different data modalities	Emotions expressed by the main actor in a video among 7 emotional classes	Short video clips extracted from movies, provided for the 2013 Emotion Recognition in the Wild Challenge	2	Accuracy: 41.03%

extracted from subtitles but also taking into account contextual information (*e.g.* user's personal profile, gender or age). However, their ground truth being annotated by a single participant only, they did not study the relevance of such contextual information and assumed the model to be personalized for this participant. The arousal information of each shot is obtained by computing linear weights by means of a Relevance Vector Machine (RVM) using low-level features published in previous work. Arousal estimation is then used as an arousal indicator feature and merged with other content-based features for discrete scene affective classification thanks to a Bayesian framework. Thus, their framework is a trade-off between continuous and discrete affective video content analysis. The Bayesian framework relies on two priors: the movie genre prior and the temporal dimension prior consisting of the probability transition between emotions in consecutive scenes. However, movies scenes are categorized into three emotional classes, which is too restrictive. Furthermore, they only provided a qualitative evaluation of the continuous arousal estimation but achieved an accuracy of 63.9% for the classification of the movie scenes among three emotional classes.

Malandrakis *et al.* also proposed in [52] a continuous affective video content analysis relying on audiovisual features extracted on each video frame, combined at an early stage and used by two Hidden Markov Models (HMMs). These two classifiers are trained independently to model simultaneously the intended arousal and valence. However, HMMs predict discrete labels. Arousal and valence are thus discretized into seven categories and the model only allows transitions between adjacent categories. They finally output time series of seven categories interpolated into a continuous-valued curve via spline interpolation, but the continuous curves are thus approximations and cannot recover the precision lost by discretizing the affective space. Their discrete and continuous curves are compared using the leave-one-movie-out approach to the ground truth collected on 30-min video clips from 12 movies. The smoothed predicted curves achieved an average correlation of 0.54 for arousal and 0.23 for valence.

3.3.2 Video emotion recognition

As stated in Section 3.1, video emotion recognition models are inherently different from affective movie content analysis models. While affective movie content analysis models aim at estimating the induced emotion from the intrinsic properties of a movie, video emotion recognition models aim at automatically estimating the emotion expressed by an agent from a video recording, in relation to an emotion induced by a stimulus. Emotion recognition models thus typically rely on facial characteristics such as action units (AU) [55]. Since felt and expressed emotions are linked, the temporal modeling of emotions of emotion recognition models may be of interest for designing new continuous affective movie content analysis frameworks.

Nicolaou *et al.* introduced in [53] a framework for continuous prediction of spontaneous affect in the VA space based on facial expression, shoulder gesture, and audio cues. They compared the performance of

the bidirectional Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) and Support Vector Machines for Regression (SVR) for continuous spontaneous affect prediction and proposed an output-associative prediction framework. The output-associative framework is a modified bidirectional LSTM-RNN taking into account the correlation between the predicted valence and arousal dimensions: it depends on the entire sequence of intermediate output predictions of both dimensions to perform the prediction. They showed that the bidirectional LSTM-RNN outperforms the SVR and that the output-associative prediction framework significantly improves prediction performance. Inspired by such promising results, LSTM-RNNs are used in Chapter 9 to continuously predict the emotions induced by videos. However, to train and evaluate their models, Nicolaou *et al.* used recordings made in a lab setting, using a uniform background and constant lighting conditions. Their model is thus highly sensitive to the recording conditions such as illumination and occlusions.

More recently, Kahou *et al.* designed a framework to assign one of seven acted-out emotions to very short video clips (1 to 2 seconds long) extracted from Hollywood movies [54]. It was the winning submission in the 2013 emotion recognition in the wild challenge. Unlike [53], videos depict acted-out emotions under realistic conditions (large degree of variation in attributes such as pose and illumination). Their framework combines a Convolutional Neural Network (CNN) focusing on capturing visual information in detected faces, a Deep Belief Network for the representation of the audio stream, a K-Means based model for extracting visual features around the mouth region, and a relational autoencoder to take into account the spatio-temporal aspects of videos. To efficiently train the CNN, they downloaded two large alternative image databases of facial expressions for the seven emotion categories. The CNN is thus highly generalizable and avoids overfitting issues. They assessed several methods for the combination of cues from the modalities and the best result was obtained with a random search over simple weighted averages. Thus, the final result is a concatenation of models based on a single modality. However, due to the characteristics of the data provided for the challenge, emotional relationships between consecutive video segments are not investigated. Using this temporal information, as Nicolaou *et al.* did through the use of LSTM-RNNs [53], may help improving the prediction performance.

3.4 Discrete affective movie content analysis

This section focuses on previous work on discrete affective video content analysis, including valence and arousal estimation but also violence detection. A summary is given in Table 3.3.

3.4.1 Discrete valence and arousal movie content analysis

Discrete affective video content analysis has been more frequently investigated than continuous affective video content analysis over the last decade.

Kang [56] was the first to propose a model where classifiers are adopted for affective analysis. He suggested detecting affective states in

Table 3.3 – *Summary of previous work on discrete affective movie content analysis and violence detection*

Authors	Method	Output	Ground truth	Annotators	Result
Kang [56]	HMMs relying on visual features only	Discrete labels among 3 classes (fear, sadness, joy)	Scenes extracted from six 30min videos	10	Accuracy: ≈79%
Wang and Cheong [57]	Two SVMs using audio cues or AV features	Video scenes classified using 7 emotional classes	2,040 scenes from 36 full-length popular movies	3	Accuracy: 74.69%
Sun and Yu [58]	4 HMMs using AV features	Labels among 4 classes (anger, fear, sadness, joy)	10 popular movies labeled at different levels	30	Precision: ≈68%, Recall: ≈79%
Xu *et al.* [59]	5 HMMs using AV cues	5 affective discrete classes	Videos from 24 movies	?	Accuracy: 80.7%
Soleymani *et al.* [60]	Personalized RVM using AV or physiological cues	1 global arousal and valence score per video	64 movie scenes from 8 popular movies	8	MSEs for each participant
Zhang *et al.* [35]	2 SVRs using both VA features and user profile	One valence and arousal score per music video	552 representative music videos	10 & 27	Performances for two applications
Irie *et al.* [61]	Latent topic driving model using VA features	Probabilities for 9 emotion categories for each scene	206 scenes from 24 movie titles available as DVDs	16	Agreement: 85.5%
Acar *et al.* [62]	Two CNNs and one SVM using AV features	1 of the 4 quadrants of the VA space for each video	Music videos from the DEAP dataset	32	Accuracy: 52.63%
Penet *et al.* [63]	2 Bayesian networks using temporal AV features	Violence probability for each video shot	MediaEval 2011 Affect Task corpus	7	False alarms and missed curves
Eyben *et al.* [64]	SVMs using temporal AV features	Confidence score to classify a video shot as violent	MediaEval 2012 Affect Task corpus	7	MAP@100: 0.398

movies including “sadness", “joy" and “fear" from low-level features using HMMs. The topology of the HMM is designed such that the only possible transitions between consecutive affective states are the ones between the neutral state and the other affective states. However, this topology is very restrictive and not realistic.

Wang and Cheong introduced features inspired from psychology and film-making rules [57]. One Support Vector Machine (SVM) is dedicated to audio cues to obtain high-level audio information at scene level. Each video segment is then classified with a second SVM to obtain probabilistic membership vectors for seven discrete emotional states. Their training data are made up of 36 full-length popular Hollywood movies divided into 2,040 scenes labeled with one or two emotional states by only three annotators. Due to the limited number of annotators, ambiguities arise and make it necessary to assign some videos with two labels. Furthermore, they do not consider the relations between consecutive scenes.

In the work of Sun and Yu [58], movie units are first represented in different granularities using an excitement curve based on the arousal curve introduced in [50]. Then, four HMMs are trained independently using features extracted on these granularities to recognize one of the four emotional states among “joy", “anger", “sadness" and “fear". Each HMM has the same topology and is composed of two states: a neutral state, and a state representing the emotion assigned to the HMM. Thus, each HMM only computes for a given observation sequence, the state transition probabilities between the neutral state and one of the four emotional states. As in [56], this topology is restrictive and not realistic. Their ground truth consists of 10 movies labeled at different levels by 30 annotators. Xu *et al.* used a similar approach in [59] sharing the same disadvantages. However, to compute the emotion intensity level they used fuzzy clustering instead of linearly combining audiovisual features, which is closer to human perception. Then, five HMMs are trained using emotion intensity and low-level features to model five emotional classes with different levels of valence. They evaluated the efficiency of their method for several movie genres, where the highest accuracy was obtained for action movies.

Soleymani *et al.* [60] compared in the VA space the values obtained automatically from either physiological responses or from audiovisual features. They showed significant correlations between multimedia features, physiological features and spectators' self-assessments for both valence and arousal. Affective scores are estimated by a linear combination of content-based features and compared to the estimation of affective scores using a linear combination of physiological features only. They showed that the performance of both models is the same and thus that none has a significant advantage over the other. A dataset composed of 64 movie scenes extracted from 8 Hollywood movies was created to assess the performance of the model. To generate the results, 42 movies scenes were randomly selected for the training set, and the remaining ones were used in the test set. Movie scenes extracted from the same movie can thus be part of the training set and also of the test set, questioning the reliability of the results.

Zhang *et al.* developed in [35] a personalized affective analysis for music videos composed of SVR-based arousal and valence models using

both multimedia features and user profiles. In fact, two nonlinear SVRs are learned for each user, taking into account the underlying relationships between user's affective descriptions and the extracted features. However, SVRs are not able to model the temporal transition characteristics of emotions. Their dataset of 552 music videos is used to train and update the models based on user feedback.

Irie *et al.* [61] proposed an approach based on latent Dirichlet allocation considering the temporal transition characteristics of emotions. Emotion-category-specific audiovisual features are extracted and transformed into affective audio-visual words using k-mean clustering. These higher level features are then used to classify movie scenes using a latent topic driving model. The model considers the probability of emotional changes between consecutive scenes based on the Plutchik's wheel [27]. However, these probabilities have not been estimated from a real dataset but empirically defined by the authors. The rate of agreement of their model equals 85.5%. The good results obtained by their framework may be due to their evaluation protocol. Their data, composed of 206 scenes from 24 movie titles available as DVDs, were randomly selected to form the training and test sets. Consequently, most films appear both in the training and the test sets as in [60], which biases the results.

More recently, Acar *et al.* proposed the use of CNNs in order to learn mid-level representations for the affective classification of music video clips [62]. Two CNNs are learned to output mid-level representation of 5-second music video clips: one uses as input audio features (MFCC) and the other one uses as input one color channel, *i.e.*, red, green or blue color channel, of the resized frame in the middle of the video segment. The mid-level representations are then each used in a dedicated SVM, and their predictions are fed to a final multi-class audiovisual SVM to output the category of the video clip (one of the four quadrants of the valence-arousal space). Their framework achieves an accuracy of 52.63% on the DEAP dataset [45]. As stated in Section 3.5, since their framework extracts basic features (MFCC and separated color channels), it lacks the ability of CNN to use raw inputs to automatically learn mid-level representations.

3.4.2 Violence detection

Emotion detection is closely related to violence detection. Both work presented in this section aim at detecting whether a video shot contains violence, defined as a physical action that results in human injury or pain.

Penet *et al.* introduced a framework using both multimodal and temporal information for violence detection [63]. For the temporal integration, their model uses the features extracted from the investigated video excerpts but also contextual features extracted from the five previous and next video shots. However, it not clear why they preferred fixed-length time windows computing features for 10 consecutive video shots. The audio features (audio energy, audio zero crossing rate, ...) , and video features (shot duration, number of flashes, ...) are separately used in two independent Bayesian networks. The two probabilities given by the networks are finally fused using a late fusion approach.

One year later, to classify video excerpts as violent or non-violent, Ey-

ben *et al.* fused by simple score averaging the predictions made by acoustic and visual linear kernel SVMs [64]. To capture temporal dynamics, numerous statistics are computed for the features over windows of fixed size. The authors also provided a detailed analysis of the features extracted from the audio and video modalities. They showed that some features are particularly relevant for violence detection: color and optical flow for the video modality, and spectral distribution descriptors and peak-based functional extraction for the audio channel.

While using short fixed-length time windows may be relevant for violence detection, estimating induced emotions requires considering longer-term dependencies [65].

3.5 Issues with the existing work

The main approach for estimating the affective content of videos is to use machine learning to build models such as HMMs [52, 58, 59], SVMs or SVRs [35, 57, 64], RVMs [51, 66], and more recently CNNs [54, 62, 67], trained on a given dataset.

Most of these models use as input a predefined set of handcrafted audiovisual features extracted from videos (see Tables 3.2 and 3.3). However, building complex handcrafted features requires strong domain knowledge, since the choice of the features is highly problem-dependent. Obtaining a satisfying feature extraction is thus hard to come by. On the other hand, CNNs are a type of deep models that can use the raw inputs directly, thus automating the process of feature construction. Nevertheless, CNNs require a relatively large number of samples to be trained compared to standard approaches, and there do not exist any public affective dataset large enough to train reliable CNNs for affective movie content analysis. To work around this problem, Kahou *et al.* trained a CNN for the analysis of facial expressions within video frames using large static image databases of facial expressions [54]. However, this model efficient to classify the emotions expressed by the primary human subject in videos may not be as efficient to estimate the emotions induced by videos which may not contain any face. Acar *et al.* chose to learn CNNs using MFCC feature vectors and color values in the RGB space extracted from one music video segment for affective music video classification [62]. By reducing the size of the input layer through the use of low-level features, they can build lighter CNNs that can be trained using smaller datasets, but they lack at the same time the ability of CNNs to use raw inputs to automatically learn mid-level representations.

Another issue is that a lot of existing work does not take into account the fact that, as stated in Chapter 2, an emotional episode is a recursive process. This is not the case for HMM-based [56, 58, 59] and RNN-based [53] affective frameworks. Indeed, HMMs are statistical models of sequential data, inherently able to take into consideration consecutive emotional changes through hidden state transitions. However, they are composed of a specific number of discrete states and thus cannot be used to directly infer dimensional scores. Malandrakis *et al.* converted discrete affective curves obtained with HMMs into continuous curves using a Savitzky-Golay filter [52], but the continuous curves are thus approximations and

cannot recover the precision lost by discretizing the affective space. RNNs-based affective frameworks are also able to take into account the temporal transitions between consecutive emotions. Combined with LSTM cells, they can learn efficiently long-term dependencies, as shown by Nicolaou *et al.* for video emotion recognition [53].

The last issue presented in this section concerns benchmarking previous work. Due to the constraints on databases presented in Section 3.2, most state of the art work about affective movie content analysis uses a private dataset of a very limited size and content diversity, designed according to their goals and needs (see Tables 3.2 and 3.3). Thus, it makes fair comparisons and results reproducibility impossible, preventing achievement of major strides in the field. For example, some work represents emotions in the 2D VA space or in the 3D valence-arousal-dominance space [35, 8], while other work represents emotions using discrete categories [57]. Furthermore, the models are sometimes dedicated to specific video categories, *i.e.*, music videos [35, 62] or a particular movie genre [68]. Horvat *et al.* showed in a survey [69] that, for researchers in the affective science field, current emotionally annotated databases lack at least some stimuli inducing a particular emotion. Participants additionally indicated that they would greatly benefit from large emotionally annotated databases composed of video clips. Soleymani *et al.* also expressed this major need and defined in [70] the specifications to be considered to allow standardized evaluation and to bypass the size and scope of related limitations of existing databases used to train and evaluate computational models in the field of affective content analysis.

Thus, the needs to build a comprehensive affective dataset include: representative videos, self-assessments annotated along a universal representation, a large number of video samples to make possible its use for machine learning processes, no copyright issues. An efficient computational model should be trained and tested using such a dataset, should not use handcrafted features, and should model the emotional relationships between consecutive video segments.

Summary

This chapter presented previous work on continuous and discrete affective video content analysis, including video emotion recognition and violence detection that are related to the induced affect estimation from audiovisual features. It appears from previous work that the main approach to analyze the affective content of videos is to use machine learning to build, using a dataset, models such as HMMs, SVRs, and CNNs.

Many studies such as [69] or [71] deplore the lack of a standard affective video database which, combined with the lack of standard evaluation protocols, decreases the efficiency of the affective research community [70]. Indeed, benchmarking and reproducibility both make it easier to know how computational models perform with respect to the state of the art, and to focus on promising research avenues. This is why we introduce LIRIS-ACCEDE and define reproducible protocols in the following chapters.

Part II

Affective Dataset Development

"Photography is truth. And cinema is truth twenty-four times per second."

— Jean-Luc Godard, *Le Petit Soldat*

LIRIS-ACCEDE: A Video Database for Affective Content Analysis

4

Contents

A large and publicly available dataset of quality video excerpts with high content diversities, along with ground truth affective annotations is needed by the research community to overcome the limitations of the existing affective video datasets and foster research in affective video content analysis. This second part of this thesis is thus dedicated to the creation of a dataset, built to be as universal as possible, annotated with complementary ground truths, each collected using a specific experimental protocol designed to take into account the specificities of the modality to be collected. In this way, the second part of this thesis is composed of four chapters, each describing an experiment adding a new modality to the publicly available dataset released in this thesis: the LIRIS-ACCEDE dataset.

In this chapter, ground truth affective annotations are collected from a wide variety of raters through crowdsourcing. The proposed dataset,

Table 4.1 – *Composition of LIRIS-ACCEDE*

Type	Data	Emotional labels
Discrete annotations	9,800 excerpts from 8 to 12 seconds long extracted from 160 movies shared under Creative Commons licenses	Crowdsourced induced arousal and valence rankings, estimated arousal and valence ratings, and violence ratings
Continuous annotations	30 full movies selected from the 160 movies used in the discrete part of the dataset	Continuous induced arousal and valence self-assessments, and continuous physiological recordings

namely LIRIS-ACCEDE, contains 9,800 video excerpts shared under Creative Commons licenses, making it possible to release the database without copyright issues. The dataset was first introduced in [11] and then fully described in [12].

4.1 Specifications of LIRIS-ACCEDE

How can a large and reliable dataset be built that could serve the community as a reliable benchmark? Crowdsourcing is often the recommended solution for creating a large dataset representing a condition. This makes it possible to reach a large number of remunerated annotators, while also guaranteeing reliability of annotators' answers via specific mechanisms. In this chapter, we will show that ranking approaches are more suited than rating approaches in crowdsourced experiments since it is more difficult to ensure that the affective scale is used consistently in crowdsourced experiments. This is why in this chapter, we take advantage of crowdsourcing to reach a large number of annotators to collect affective ranks for short video segments. These crowdsourced affective ranks are converted into affective scores in Chapter 5 thanks to a complementary experiment in a more controlled laboratory environment. Both affective scores and affective ranks can be used in discrete affective video content analysis work [1].

Continuous affective movie content analysis work has not been overlooked. Continuous affective self-assessments and physiological measurements have been collected in a laboratory environment from participants watching long movies, respectively in Chapters 6 and 7. Contrarily to the discrete annotations, these continuous annotations make possible to create models taking into account the fact that previous movies scenes may reasonably influence the emotions induced by future ones. The composition of the LIRIS-ACCEDE dataset is summarized in Table 4.1.

LIRIS-ACCEDE uses the widely employed 2D valence-arousal space to collect the affective self-assessments. However, as the database is freely shared, everyone is free to add new modalities, thus enhancing the range

1. Discrete violence annotations have also been released publicly.

of possible applications. Furthermore, this database is very large and diversified, unlike most of the databases presented in the previous sections. As a consequence, we think it could be general enough to be used as a reference in the future.

4.2 Discrete database description

The discrete part of LIRIS-ACCEDE is made up of 9,800 excerpts extracted from 160 feature films and short films. It is the largest video database currently in existence annotated by a broad and representative population using induced emotional labels.

4.2.1 Movies used in LIRIS-ACCEDE

One of the main requirements of LIRIS-ACCEDE was that it should be freely available to the research community. That is why the 160 movies used for creating the database are shared under Creative Commons licenses. Creative Commons is a non-profit corporation providing standardized free copyright licenses to mark a creative work with the freedom the creator wants it to convey. The CC BY license known as “Attribution" is the most accommodating license since users can reuse the original creation as long as they credit the creator. Three modules adding more restrictive conditions can be combined. The SA module (ShareAlike) requires that works based on other works shared using this module, have to be licensed under identical terms. The NC module (NonCommercial) prevents original works from being reused for commercial purposes. Last but not least, the ND module (No Derivative Works) prohibits altering, transforming, or building upon original works. To create the database, we have used only movies shared under a Creative Commons license that do not contain the ND module, because our goal was to modify the selected movies by extracting several excerpts from them. Thus, using videos shared under Creative Commons licenses makes it possible to share the database publicly without copyright issues.

Most of the 160 movies used for creating LIRIS-ACCEDE come from the video platform VODO. This references best free-to-share feature films and short films that have been submitted on the website and makes them easily available to millions of people. It is important to notice that free-to-share films do not mean User Generated Contents with low expertise levels. Movies referenced on VODO have been created by filmmakers with excellent technical expertise. Many films in the database have been screened during film festivals including, but not limited to, “RIP! A remix manifesto" directed by Brett Gaylor (Special Jury Prize at the “Festival du Nouveau Cinéma in Montreal"), “Emperor" directed by Juliane Blockand (winner of the Feature Category at the Portable Film Festival) and “Pioneer One" produced by Josh Bernhard and Bracey Smith (winner of the Best Drama Pilot at the New York Television Festival). The “Home" documentary directed by Yann Arthus-Bertrand included in the database is a special case since it is a big budget movie distributed by 20th Century Fox that has no copyright.

In brief, 40 high quality feature films and 120 short films shared under Creative Commons licenses have been collected to create the 9,800 excerpts making up LIRIS-ACCEDE. The total time of all 160 films is 73 hours, 41 minutes and 7 seconds. A list of 9 representative movie genres describes the movies: Comedy, Animation, Action, Adventure, Thriller, Documentary, Romance, Drama and Horror. By displaying the normalized distribution of movies by genre in LIRIS-ACCEDE compared to the normalized distribution of movies by genre referenced on IMDB [2] and on ScreenRush [3], it can be observed that distributions appear to be similar. Thus, movies used in LIRIS-ACCEDE are representative of today's movies. Languages are mainly English with a small set of French, German, Icelandic, Hindi, Italian, Norwegian, Spanish, Swedish and Turkish films, subtitled in English. Note that 14 movies are silent movies.

4.2.2 Characteristics of LIRIS-ACCEDE

The database is made up of 9,800 excerpts extracted from the 160 selected movies listed in Appendix A.

1,000 excerpts have been manually segmented because they were part of the pilot test to ensure the reliability of the annotations. Subsequently, the other excerpts have been automatically segmented using a robust cut and fade in/out detection, implemented based on the algorithms described in [72]. Because all the segmented excerpts start or end with a cut or a fade, it is very likely that each segment be perceived by users as semantically coherent.

The 9,800 segmented video clips last between 8 and 12 seconds, and the total time of all 9,800 excerpts is 26 hours, 57 minutes and 8 seconds. Even if the temporal resolution, or granularity, of emotions is still under debate, most of psychologists agree that they are part of a complex but very rapid process [73]. They are phenomena with onsets and ends over seconds [44]. Indeed, the length of extracted segments in LIRIS-ACCEDE is large enough to obtain consistent excerpts, making it possible for the viewer to feel emotions. For example Gross and Levenson successfully elicited emotions in the laboratory using short excerpts lasting a few seconds [41]. Moreover, Metallinou and Narayanan have shown in [74] that global ratings of perceived emotion for movies lasting a few minutes are not simple averages over time, but rather are more influenced by highly arousing events with low valence. By using short excerpts, we greatly minimize the probability that annotations are a weighted average of consecutive emotions felt during successive events.

Despite the short duration of excerpts, most are composed of several video-editing features (*e.g.*, shot cuts, dissolves, digital zooming). This is essential since many previous studies, including [50] and [75], have shown that the arousal dimension was correlated to editing features such as the shot cut rate or the presence of dissolves. Only 1,760 excerpts do not include any scene cut or fade in/out. On average excerpts are composed of 2.8 video-editing features (this statistic does not count the editing features on the boundaries).

2. http://www.imdb.com/
3. http://www.screenrush.co.uk/

More generally, we achieved a great variety of excerpts reflecting the variety of selected movies. The excerpts contain scenes of violence, sexuality, murders, but also more common scenes such as landscapes, interviews and many positive scenes of daily life. This variety is confirmed thanks to another experiment we conducted using crowdsourcing where workers had to annotate the context of the videos according to four categories (indoor, urban, nature and other). 6,441 excerpts (65.7%) have been categorized as Indoor scenes, 1,060 as Nature and 2,008 as Urban. The remaining ones correspond to 291 excerpts labeled as Other (screen captures, texts,...). LIRIS-ACCEDE is currently the only video database annotated along induced emotions that includes such a large range of contexts.

4.3 Discrete data annotation

4.3.1 Experimental design

The annotation process aims at sorting the 9,800 excerpts independently along the induced valence and arousal axes. Crowdsourcing is an appropriate choice for achieving this goal requiring a huge amount of annotations, and has proved to be useful in various annotation studies (*e.g.* [76, 77, 78]). To annotate LIRIS-ACCEDE data, video excerpts were presented to annotators, also known as workers, on CrowdFlower. [4]

Rating-by-comparison experiments, *i.e.*, ranking approaches, are more suited than rating approaches in experiments conducted on crowdsourcing platforms. Plausibly, asking for pairwise comparisons seems less complex than asking for an absolute value. Indeed, ratings require that annotators understand the range of an emotional scale, which is a sizable cognitive load [79], and it is quite difficult to ensure that the scale is used consistently. Russell and Gray [80] showed that raters using rating scales tend to only use a small subset of the range, while Ovadia [81] pointed out that inter-annotators ratings, *i.e.*, ratings from different annotators, and even intra-annotator ratings, *i.e.*, ratings from the same annotator, may not be consistent. By choosing pairwise comparisons instead of ratings, the consistency of the annotations is improved, as annotators tend to agree more when describing emotions in relative terms than in absolute terms [74]. Pairwise comparisons are also more appropriate detectors of user states, discarding the subjectivity of rating scales and implicit effects linked to the order of annotations [82]. Yang and Chen also showed in [79] that pairwise comparisons enhance the reliability of the ground truth compared to rating approaches, and simplify emotion annotation. This simplification also makes tasks more attractive and interesting to annotators. From an involved annotator's point of view, because the amount of money they earn is proportional to the quality of their answers and the amount of time they spend on the task, the simpler a task is, the more they are disposed to annotate additional comparisons.

4. While we conducted the experiments (summer 2013), CrowdFlower was distributing tasks over 50 labor channel partners, including Amazon Mechanical Turk and TrialPay. Since late 2013, the number of its labor channel partners has been considerably reduced. For example, CrowdFlower does not offer task distribution on Mechanical Turk anymore and it is no longer possible to choose on which labor channel the tasks are distributed.

Accordingly, the choice of a rating-by-comparison experiment to annotate LIRIS-ACCEDE stands out. For each pair of video excerpts presented to workers on CrowdFlower, annotators had to select the one which conveyed most strongly the given emotion in terms of valence or arousal. The advantage of forced choice pairwise comparisons is that annotators must come to a decision. Forced choice pairwise comparisons enhance the reliability of experiments compared to other protocols such as displaying a single stimulus and a categorical rating scale [83] and encourage more thorough processing of response options [84].

If all possible comparisons had been generated and annotated by three crowdworkers, the experiments would have cost US$2,880,906 each. Thus, it was essential to choose an algorithm to select carefully and efficiently the comparisons judged by the annotators. The quicksort algorithm [85] was used to generate the comparisons and rank the video excerpts according to the annotations gathered from CrowdFlower. This is one of the most efficient sorting algorithms. Indeed, the average computational complexity of this algorithm is $O(n \log n)$, where n is the number of data to sort. In the worst case, complexity is $O(n^2)$, but this performance is extremely rare and in practice the quicksort is often faster than other $O(n \log n)$ algorithms [86]. As the cost of the sorting operation is proportional to the number of comparisons, the quicksort seems the best choice for reducing costs to sort the whole database compared with other sorting algorithms. In practice, the quicksort algorithm allows costs to be reduced to approximately US$10,000 for ranking of the whole dataset along one axis. The principle of the quicksort algorithm is to choose an element, called a pivot, to which all other elements are compared. Thus, two subgroups of unsorted elements are created, one with a higher value than the pivot and the other with a lower value. Each subgroup is then sorted recursively in the same way until every element of each group is sorted.

The subgroups generated by the quicksort algorithm depend on the annotations gathered for a particular task. Consequently, the pivot and the comparisons vary from one axis to another. That is why the annotation process of LIRIS-ACCEDE was divided into two experiments: one for annotation of valence and another for annotation of arousal. The experimental protocol was virtually the same for each axis and is described in Section 4.3.2.

4.3.2 Experimental setup

The annotation of the database along the arousal axis was performed three months after the annotation along the valence axis. Meanwhile, a new interface for displaying tasks to workers had been released on CrowdFlower. This explains why there are few changes in both protocols to adapt the experimental setup to the new interface.

Given a pair of video excerpts, annotators had to select the one that conveys “the most positive emotion" (for valence) or “the calmest emotion" (for arousal). The words “valence" and “arousal" were not used since they might be misunderstood by the annotators. They were asked to focus on the emotion they felt when watching the video clips, *i.e.* the induced emotion, and not on that of the characters. As the arousal axis was more

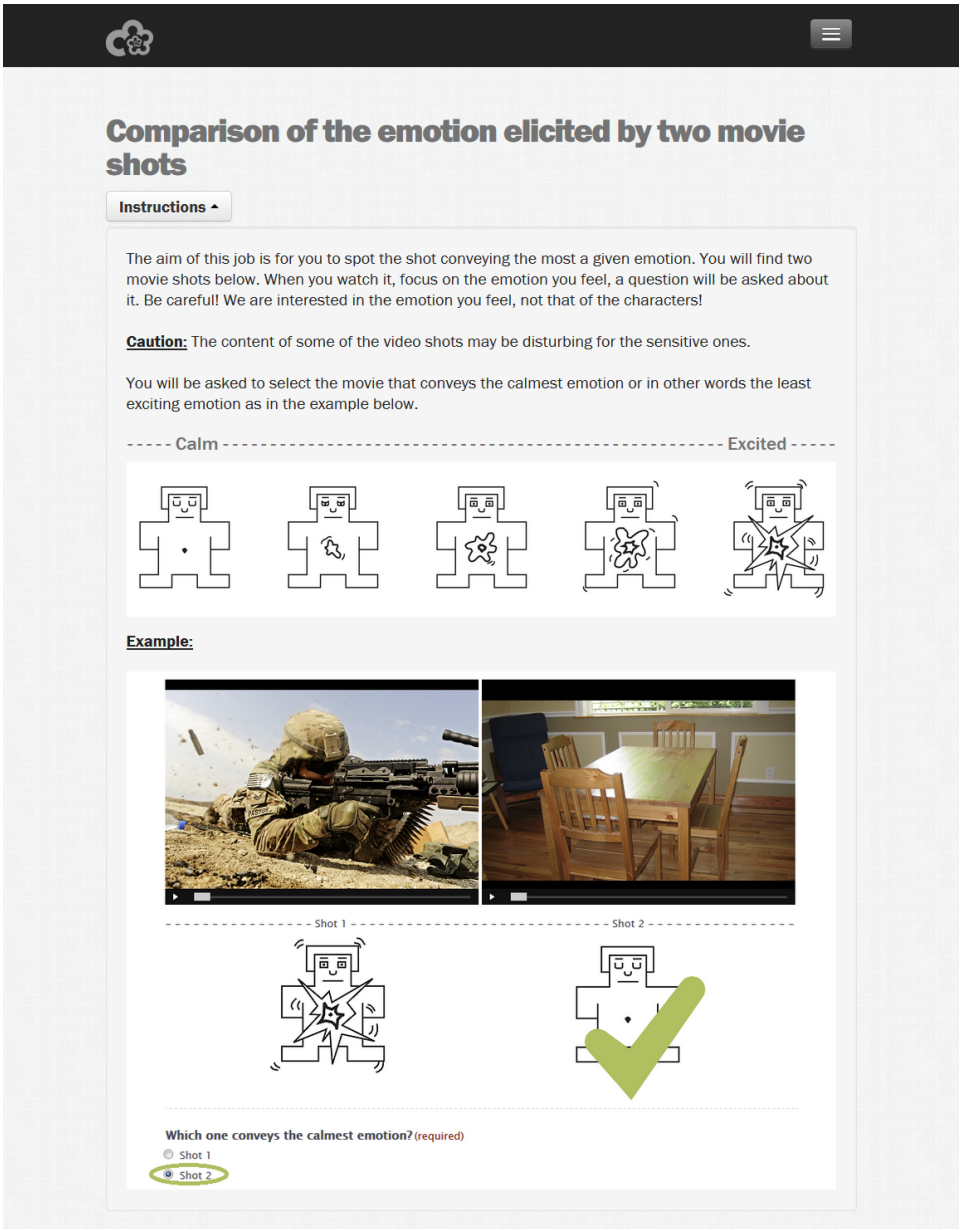


Figure 4.1 – *Interface displayed to workers for annotation of the arousal axis.*

challenging to annotate, the arousal axis of the Self-Assessment Manikin and an example were displayed at the beginning of the task to make sure that annotators had understood the task properly. The Self-Assessment Manikin [87] is a powerful pictorial system used in experiments to represent emotional valence, arousal and dominance axes. Its non-verbal design makes it easy to use regardless of age, educational or cultural background. The interface displayed to workers for annotation of LIRIS-ACCEDE along the arousal axis is shown in Figure 4.1.

Video clips were displayed with a size of 280×390 pixels for annotation of the valence dimension and with a size of 189×336 pixels for the arousal dimension, to comply with the width of the new interface and use a more common aspect ratio. These clips were displayed using an embedded video player, meaning that workers were free to play each video clip as many times as they wanted. Workers were paid US$0.05 for answering

five comparisons but could exit the task at any time. Despite the low reward for completing tasks, feedback on specialized crowdsourcing forums was very positive. Workers pointed out that the tasks were very easy, fun and enjoyable. Here are a few of their comments: “That’s awesome!", “I did that last time, want to do that again, very easy :)".

To ensure the accuracy of annotations, 100 unnoticeable test questions, also called “gold units", were created for each axis and randomly inserted throughout the tasks. This made it possible to test and track annotators’ performance by regularly testing them to ensure that they take the video clips comparisons seriously. The gold units correspond to unambiguous pairs of easily comparable video clips. If a wrong answer was given, a small paragraph was displayed explaining the reason why the answer was the other one. Workers were able to question the reason and send a message to explain their point of view. This system made it possible to forgive them when their protest was well-founded and to modify accordingly several gold units that were too subjective. However, if a worker gives too many wrong answers to gold units, none of his answers are considered, he receives no remuneration and his trust level on CrowdFlower drops. Thus, annotators are well aware that they must not answer the questions at random. For annotation of the arousal axis that took place 3 months after the annotation of the valence axis, a new advanced tool called “Quiz Mode" was available on CrowdFlower: annotators first have to answer six test questions and achieve an accuracy threshold of 70% in order to pass the quiz and work on the job. This ensures that only higher performing annotators are allowed to work on the tasks. Test questions were also randomly inserted to test annotators that passed the quiz on an on-going basis as they worked through the job.

In concrete terms, the quicksort algorithm was used in both annotation experiments to generate the comparisons. First, an initial video excerpt was randomly chosen to be the first pivot. All the other clips from the database were compared to this excerpt meaning that 9,799 pairwise comparisons were generated for the first iteration. Each pair was displayed to workers until three annotations were gathered. We found this was a good compromise between the cost and the accuracy of the experiment. Once three annotations per comparison were made for all the comparisons, all the annotations were collected. In each comparison, the pivot was considered as inducing the most positive emotion for valence or the calmest emotion for arousal if at least two annotators selected the pivot during the annotation process. The final rank of the pivot was thus computed. Assuming that the pivot does not induce the lowest or the highest valence or arousal, this process splits the database into two subgroups. For the second iteration, one pivot was selected in each subgroup and the two pivots were compared to the other video clips inside their subgroup, generating 9,797 new comparisons. For the next iteration, four pivots were selected and so on. The process was repeated until a rank was assigned to all the 9,800 video excerpts. Finally, each video excerpt is accompanied by two discrete values ranging from 0 to 9,799 representing its arousal and valence ranks.

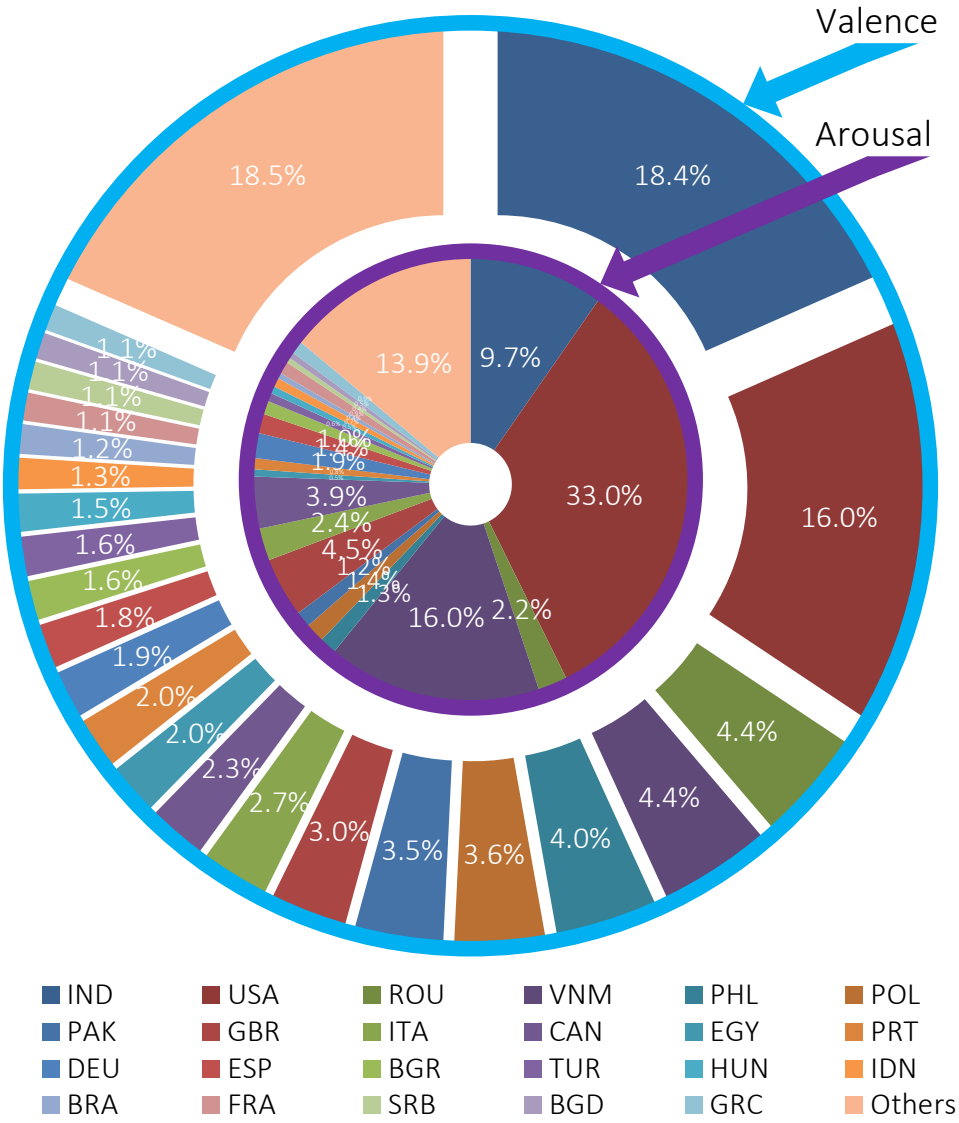


Figure 4.2 – *Countries of the annotators for both the valence (external circle) and arousal (internal circle) annotation experiments. Countries accounting for less than 1% of the total in both experiments are classified as "Others".*

4.3.3 Annotation statistics

For annotation of the valence axis, more than 582,000 annotations for about 187,000 comparisons were gathered from 1,517 trusted annotators from various countries. Annotators from 89 countries participated in the experiment, reflecting a huge diversity in cultural background. The majority of workers originated from India (18%), USA (16%), Romania (4%) and Vietnam (4%). A more detailed distribution of countries is displayed in Figure 4.2. Over 90% of data come from 530 of these annotators. The 1,517 trusted annotators showed an accuracy of 94.2% on test questions, whereas this accuracy was about 42.3% for untrusted annotators.

More iterations were needed to fully rank the database along the arousal axis. More than 665,000 annotations for around 221,000 unique comparisons were gathered from 2,442 trusted annotators also from 89 countries. As displayed in Figure 4.2, the countries of annotators are also

diversified but different since most of the workers are American (33%), Vietnamese (16%), Indian (10%) and British (5%). As a point of comparison, this time over 90% of data come from 830 annotators. The accuracies on test questions for trusted and untrusted annotators were approximately the same as for those annotating the database along the valence axis. However, the number of untrusted annotators was slightly lower than for the first experiment thanks to the Quiz Mode.

When creating crowdsourcing tasks, ethical concerns have to be considered and the anonymity of the crowdworkers must be preserved. It is worth mentioning that, in our experiments, crowdworker privacy has been protected since the annotations that led to the ranking of the excerpts are not published. Only the final ranks for valence and arousal are released.

Combination of valence and arousal annotations shows convincing results. Dietz and Lang have shown in [3] that arousal and valence are correlated and that certain areas of this space are more relevant than others. Figure 4.3 shows the two-dimensional quantized histogram of ranks computed from annotations in the 2D valence-arousal space. Each cell indicates the number of video clips with a ranking for valence and arousal between the values represented on both axes. For example the top-left cell shows the number of excerpts with a ranking between 0 and 700 for valence and between 9,100 and 9,800 for arousal. Similarly to other studies such as [3] and [52], Figure 4.3 shows that there are relatively few stimuli eliciting responses annotated as low arousal and negative valence and that there are also less excerpts eliciting high arousal and neutral valence. Note that, the values displayed in Figure 4.3 are the relative positions of excerpts in the valence-arousal space and not their absolute position.

4.3.4 Inter-annotator reliability

Inter-annotator reliability is an indication of how independent annotators participate in an experiment and reach the same conclusion despite the subjectivity of the task. It is essential to evaluate the consistency of the annotations to detect whether the scale is defective or whether the annotators need to be re-trained. Several measures of inter-annotator agreement are used in the literature such as percent agreement, Fleiss' kappa [88] and Krippendorff's alpha [89]. Percent agreement is widely used and intuitive but overestimates inter-annotator reliability since it does not take into account the agreement expected by chance. Most of the annotators who answered randomly have been discarded using gold data, which is why this measure will also be considered in Table 4.2. Fleiss' kappa and Krippendorff's alpha both take into account observed disagreement and expected disagreement but are sensitive to trait prevalence: they consider that annotators have a priori knowledge of the quantity of cases that should be distributed in each category [90] (*e.g.* "Excerpt 1" or "Excerpt 2" conveys most strongly the given emotion). The result is that, especially using binary answers which is the case here, if a value is very rare, reliability is low even if there are few mistakes in the annotations. In our annotation process this is a problem because the rarity of a category (the shot that conveyed most strongly a given emotion) greatly depends on the choice of pivots in the quicksort algorithm. For example, if a pivot with

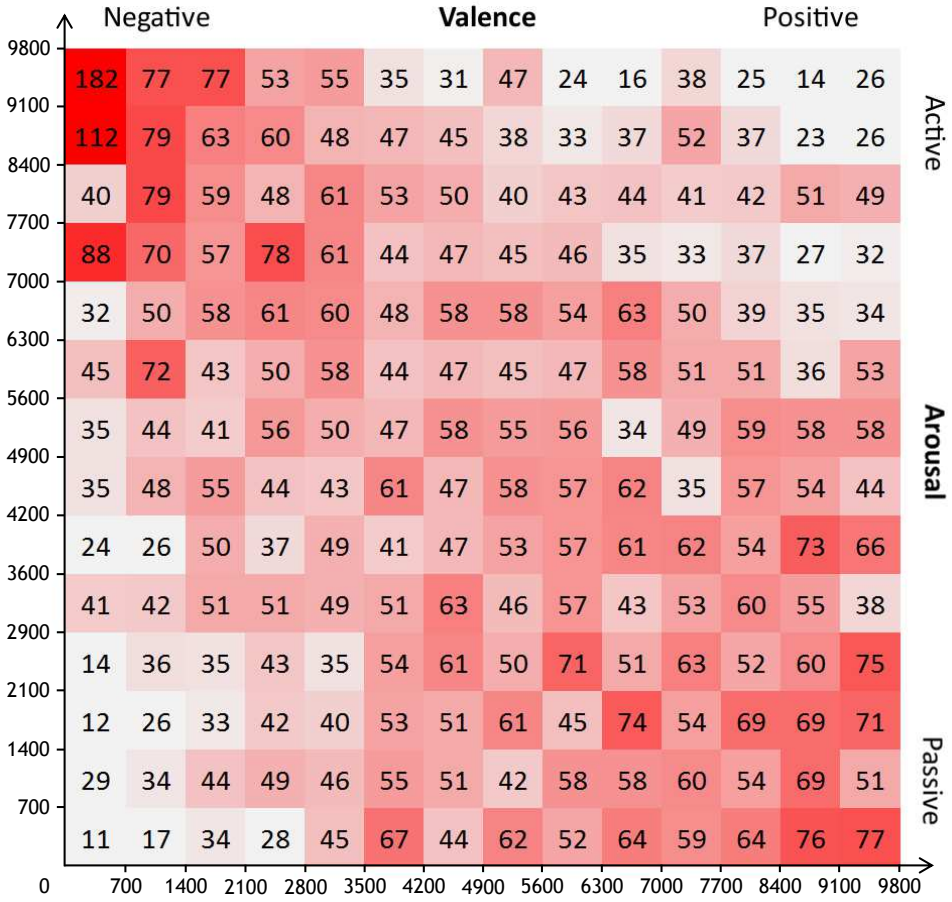


Figure 4.3 – *Joint quantized histogram of ranks for the 9,800 excerpts in the valence-arousal space. For example, the bottom-left cell shows the number of video clips with a valence and an arousal rank between 0 and 700.*

Table 4.2 – *Inter-annotator reliability*

Measure	Arousal	Valence
Percent agreement	0.862	0.835
Fleiss' κ	0.190	0.179
Krippendorff's α	0.191	0.180
Randolph's κ_{free}	0.452	0.375

a high valence is selected, most annotators will answer that the pivot (always displayed as "Excerpt 2") has the highest valence. This will result in a low reliability using Fleiss' kappa and Krippendorff's alpha measures. Randolph's multirater kappa free [90] is not subject to prevalence because it does not depend on how many values are in each category. All these reliability coefficients are displayed in Table 4.2 to ensure a point of comparison. Appendix B lists the formulas used to compute these reliability coefficients.

Both kappa values need a fixed number of annotators per comparison to be computed. However, comparisons can be annotated by different annotators. For this reason, all comparisons that have been annotated by more than three people are discarded to compute both kappa values, corresponding to 7,459 units discarded for valence and 1,539 for arousal. Krippendorff's alpha is more flexible and allows missing data (comparisons can be annotated by any number of workers), thus no comparisons are discarded to compute this measure. The inter-annotator reliabilities for these subsamples are displayed in Table 4.2. Their values can range from 0 to 1 for percent agreement and from -1 to 1 for the other measures. For Fleiss' kappa, Krippendorff's alpha and Randolph's kappa, a value below 0 indicates that disagreements are systematic and exceed what can be expected by chance, a value equal to 0 indicates the absence of reliability, and a value higher than 1 indicates an agreement between annotators (1 for perfect reliability). In Table 4.2, all values are positive, which means that agreement is slightly better than what would have been expected by chance and is similar to other emotion annotation studies such as [52] or [78]. The percent agreement indicates that annotators agreed on 83.5% and 86.2% of comparisons. Randolph's kappa, which is robust against prevalence, yields the best reliability value compared to Fleiss' kappa and Krippendorff's alpha. This is not surprising since it is the only measure that is not influenced by the proportion of answers in each category, as the value of the arousal experiment is higher than the value of the valence experiment. For Randolph's kappa measure, Landis and Koch [91] suggest that a score of 0.375 indicates a fair agreement and that a score of 0.452 corresponds to a moderate agreement. Thus, these results show that annotators have fully understood the tasks and achieved good agreement despite the subjectivity of both annotation experiments.

4.4 Testing protocols

The goal of this section is to introduce several protocols to assess the performance of computational models using LIRIS-ACCEDE in different ways. These reproducible protocols will allow fair comparisons between future discrete models and the baseline described in Chapter 8.

Protocol A: Predefined subgroups — In this protocol, the training and test sets have been manually defined to make sure that they each include 4,900 excerpts from 80 films. In this way, the excerpts extracted from the same film are only in one of the sets, either the training set or the test set. Insofar as possible, we tried to distribute movies equally in the sets according to their genres. We also defined a validation set, should it be needed in future studies, by dividing the training set into two subgroups, each made up of 2,450 excerpts extracted from 40 films. The list of excerpts in each set is available alongside the database and the annotations.

Protocol B: Leave-One-Movie-Out — This protocol is a standard protocol used in numerous studies in affective analysis. It consists in selecting the excerpts of one movie for testing while using the rest for training. This process is repeated for the 160 movies in the database.

Protocol C: Same genre — It could also be interesting to focus on specific genres to study the efficiency of models and the effect of features depending on the movie genre. The protocol is the leave-one-movie-out protocol for movies that share the same genre.

Protocol D: Same movie — The purpose of this last protocol is to gain insight into the regularity of the movie in terms of affective impact. Indeed, by learning on samples from the first half of a movie and testing on the remaining excerpts, the results can provide information on how well the first part of a movie is able to model and to be generalized to the induced valence and arousal of the whole movie.

4.5 Discussion

One of the main limitations of the proposed database lies in the fact that the video clips have been ranked relatively to each other. Thus, the rankings provide no information on distribution of the database in the 2D valence-arousal space. In other words, it is uncertain whether the extreme cases with the lowest or highest ranks elicit extreme emotions. Furthermore, these ranks are relative to this particular database, which prevents comparison with other video clips annotated with absolute valence and arousal scores. To address this limitation, we have carried out a complementary experiment described in Chapter 5.

Several other unknown factors can potentially affect the ratings and would require further research.

First, crowdworkers were asked to focus on what they felt in response to the video excerpts. Contact with the crowdworkers was quite limited. As such, it was not possible to ascertain that annotators were annotating the induced emotion and not the perceived emotion or even the emotion they thought they should feel, since it is possible to make judgments on

the basis of conventional characteristics without experiencing any emotion [92]. If some crowdworkers did not distinguish between felt and perceived emotions, noisiness could potentially be introduced in our data as Zenter *et al.* showed that ratings of perceived emotion differ significantly from ratings of felt emotion [93]. The distinction between ratings of perceived or felt emotion is outside the scope of this thesis. Thus, in this work, we do not try to distinguish ratings of felt emotion from ratings of perceived emotion.

Second, there was no way to make sure that crowdworkers really turned on the volume to judge the videos. While creating the gold data, sound was taken into consideration. Thus, we assume that most workers passing the gold data turned the volume on. Furthermore in Chapter 5, the correlation between affective ratings collected in a controlled environment where the sound was turned on and crowdsourced rankings is significantly high. As a consequence, we hypothesize that most crowdworkers turned the volume on to rate the pairwise comparisons.

Third, the crowdworkers made the annotations in various uncontrolled environments under different conditions. However, elicitation of an emotion is a subtle process depending on a large number of factors (e.g. listener, performance or contextual features) [94]. Despite this, inter-annotator reliability indicates that an overall agreement was achieved among crowdworkers and that annotations tend to be stable. Moreover, these results have been compared in Chapter 5 to ratings gathered in controlled conditions in order to validate the annotations made in uncontrolled conditions and to detect potential outliers. The correlation between affective ratings and crowdsourced rankings is significantly high, thus cross-validating the overall database for future uses in research work. These affective ratings also make it possible to enhance the range of applications for automatic approaches capable of predicting the affective impact. Indeed, it will be easier to create new evaluation protocols, such as separating data to create two or more meaningful categories to evaluate the efficiency of classifiers for which precise affective ratings are not necessary.

Summary

This chapter has addressed the lack of large video databases for affective video analysis, as current existing databases are limited in size and not representative of today's movies. We proposed LIRIS-ACCEDE, a large video database freely shared to be used by the research community. The database is made up of 9,800 excerpts lasting from 8 to 12 seconds, extracted from 160 diversified movies. All the 160 movies are shared under Creative Commons licenses, thus allowing the database to be shared publicly without copyright issues. It is available at: http://liris-accede.ec-lyon.fr/.

All the excerpts have been ranked along the induced valence and arousal axes by means of two experiments conducted on a crowdsourcing platform. Both experiments were highly attractive. A large number of annotators performed each experiment, making it possible to collect large volumes of affective responses from a wide diversity of annotators

and from a large spectrum of contexts. With this experimental design, high inter-annotator reliabilities were achieved considering the subjectivity of the experiments. We also introduced standard protocols using the database in an attempt to perform standardized and reproducible evaluations to fairly compare future work within the field of affective computing. Four protocols were proposed corresponding to different goals and needs.

From Crowdsourced Rankings To Affective Ratings

5

Contents

The previous chapter introduced LIRIS-ACCEDE, a dataset in which 9,800 video excerpts have been annotated with pairwise comparisons using crowdsourcing along the induced dimensions of the 2D valence-arousal emotional space. All the video clips being ranked along both dimensions, the rankings provide no information about the distances between them. However, the pairwise annotations made in various uncontrolled environments have not been validated yet. Furthermore, these ranks are relative to this particular database which prevents the comparison with other video clips annotated with absolute valence and arousal values.

This is why the goal of this chapter is to cross-validate and to enrich the LIRIS-ACCEDE database by providing absolute video ratings in addition to video rankings that are already available. The new absolute video ratings are generated thanks to a regression analysis, allowing to map the ranked database into the 2D valence-arousal affective space. Gaussian Processes for Regression were preferred over other existing regression techniques since they can model the noisiness from measurements

and thus take into account the subjectivity of emotions. The proposed regression analysis is performed using rating values collected on a subset of the database.

5.1 Controlled rating experiment

To cross-validate the annotations gathered from various uncontrolled environments using crowdsourcing and to provide absolute video ratings for the whole dataset, another experiment has been created to collect ratings for a subset of the database in a controlled environment.

5.1.1 Selecting stimuli from the LIRIS-ACCEDE dataset

Eliciting emotional reactions from test participants in laboratory experiments is quite tricky, that is why it is crucial to select most effective stimuli. Therefore, Krippendorff's alpha measure has been computed to select a subset of the dataset to be used in the user study [1]. It ensures that the highest reliable film clips in eliciting induced emotions are selected. Krippendorff's alpha reliability coefficient is a generalization of several known reliability measures [89]. It applies to any number of observers, any number of categories, any metric, incomplete or missing data, and large or small sample sizes not requiring a minimum. The reliability of the excerpt $i \in \{0, \ldots, 9799\}$ for arousal or valence is defined as:

$$\alpha^i = 1 - \frac{D_0^i}{D_e^i} \tag{5.1}$$

where D_0^i is the observed disagreement among values assigned to pairwise comparisons in which one of the two compared excerpts is the excerpt i and D_e^i is the expected disagreement when the annotations are attributable to chance:

$$D_0^i = \frac{1}{n_i} \sum_c \sum_k o_{ck}^i \cdot \delta_{ck}^2 \tag{5.2}$$

$$D_e^i = \frac{1}{n_i \left(n_i - 1\right)} \sum_c \sum_k n_c^i \cdot n_k^i \cdot \delta_{ck}^2 \tag{5.3}$$

with c and k the categories of annotations' values, n_i the number of pairable annotations, *i.e.* $n_i = \sum_c \sum_k o_{ck}^i$, and n_c^i, n_k^i the number of pairable annotations with value c and k respectively, *i.e.* $n_c^i = \sum_k o_{ck}^i$ and $n_k^i = \sum_c o_{ck}^i$. The coincidence matrix o_{ck}^i is defined as:

$$o_{ck}^i = \sum_{u \in U^i} \frac{\text{Number of } c \text{ - } k \text{ pairs in comparison } u}{m_u - 1} \tag{5.4}$$

with U^i the subset of the pairwise comparisons for which one of the two compared excerpts is the excerpt i and m_u the number of annotations for comparison u.

1. In Chapter 4, we mentioned that the Krippendorff's alpha measure was subject to prevalence and thus it was difficult to analyze the value computed by the Krippendorff's alpha. However, it is still a reliable measure to compare the reliabilities of the excerpts, in particular it does not discard any data since it allows missing data.

Actually, the LIRIS-ACCEDE dataset has been annotated using forced-choice pairwise comparisons [11] so that each video excerpt is accompanied by two discrete values ranging from 0 to 9,799 representing its arousal and valence ranks. In concrete terms, a comparison between two video clips was displayed to workers until three annotations were gathered, *i.e.* $m_u = 3, \forall u \in U^i, \forall i \in \{0, \dots, 9799\}$. The crowdworkers had to select the excerpt that conveyed the most the given emotion in terms of valence or arousal. Thus, c and k values in eq. (5.4) represent the excerpt selected by a crowdworker and can be equal to "excerpt 1" or "excerpt 2". The δ^2 coefficient for nominal data is defined as:

$$\delta^2_{ck} = \begin{cases} 0 & \text{if } c = k \\ 1 & \text{if } c \neq k \end{cases} \tag{5.5}$$

In other words, for each excerpt, eq. (5.1) is used to compute its Krippendorff's alpha coefficient for valence using the crowdsourced annotations of valence and its Krippendorff's alpha coefficient for arousal using the crowdsourced annotations of arousal. These values are used to select 20 excerpts per axis (valence and arousal) that are regularly distributed in order to get enough excerpts to represent the whole dataset in the 2D valence-arousal space while being relatively few to create an experiment of acceptable duration. For each axis, the 20 excerpts that form a perfect regular distribution are the ones so that their rank equals to $\frac{9800}{19} \times n$ with $n \in \{0, \dots, 19\}$. These ranks are called the optimum ranks. Thus, for each axis and each optimum rank, we select the excerpt i with $\alpha^i \geq 0.6$ as close as possible to the optimum rank. This process ensures that the 40 selected film clips have different levels of valence and arousal and thus are representative of the full dataset. They are also representative of the agreement among the crowdworkers since just half of the video clips are highly reliable in eliciting valence or arousal, *i.e.* the 20 video clips that are highly reliable in eliciting valence may not be highly reliable in eliciting arousal and vice versa.

5.1.2 Experimental protocol

One of the objectives of this user study is to provide ratings of arousal and valence for each of the 40 film clips selected in the previous section.

28 volunteers participated in the experiment (8 females and 20 males), aged between 20 and 52 ($mean = 34,93 \pm 8,99$). All the participants are working at Technicolor as researchers, PhD candidates or trainees. Of these individuals, 23 are French and the others are Bangladeshi, Chinese, Ethiopian, Romanian or Vietnamese. 8 out of the 28 participants participated in the experiment in the morning. They were asked to read a set of instructions informing them of the protocol of the experiment and the meaning of the two different scales used for self-assessments (see Fig. 5.1(a)). Following the procedure of Philippot [40], participants were instructed to report what they actually felt while watching the video excerpts, rather than what they thought they should feel. They were also asked to focus on what they felt at the time they watched the film clips, rather than their general mood of the day. Moreover, they were told that they were free to withdraw from the test at any time. Five test video

clips from the LIRIS-ACCEDE dataset, but different from the 40 videos selected for this experiment in Section 5.1.1, were shown to the participants to make them understand the type of stimuli they could see during the test. An experimenter was also present at the beginning to answer any questions.

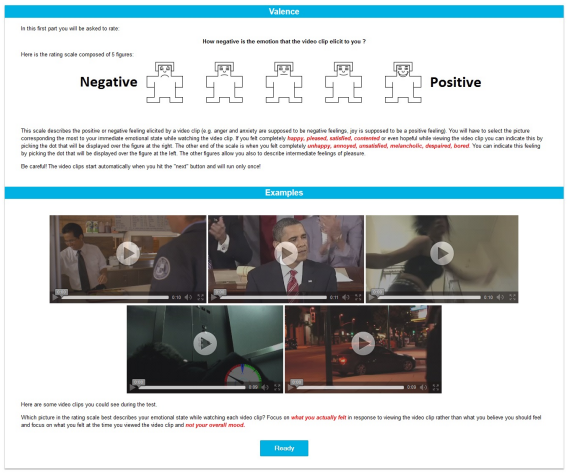
In addition to the 40 film clips selected in the previous section, 6 videos from these film clips were repeated twice in order to measure the intra-rater reliability. Consequently, 46 film clips (but 40 unique videos) were shown to the participants. The videos were presented in a dark room on a 22-inch screen (1,920 ×1,200, 60 Hz) in S-RGB mode and all film clips were displayed with a resolution height of 780px, the width depending on the ratio of each video. Participants were seated approximately 1 meter from the screen. The stereo Sennheiser PXC 360 BT headphone was used and the volume was set at a comfortable level.

The scale to report the ratings was the Self-Assessment Manikin (SAM) [87]. Affective ratings were made on the discrete 5-point scale versions for valence and arousal. Instructions were adapted from Lang *et al.* [30] (see Fig. 5.1(a)). The experiment took place in two rounds. During the first round, participants performed a self-assessment of their level of valence directly after viewing each film clip. All the videos were presented in a random order and the participant was asked to rate immediately "How negative is the emotion that the video clip elicits to you?" (see Fig. 5.1(b)). For the second round, participants performed a self-assessment of their level of arousal according to "How calm is the emotion that the video clip elicits to you?" (see Fig. 5.1(c)), all the videos being also presented in a random order. The video excerpts were run only once but participants had unlimited time to rate the videos. The next video started immediately once the participant hit the "OK" button. The vocabulary used in this test to describe the valence and arousal is the same than the one used to annotate the whole dataset in Chapter 4. Valence has been intentionally annotated before arousal because it is intuitively easier to assess and thus more encouraging and motivating.

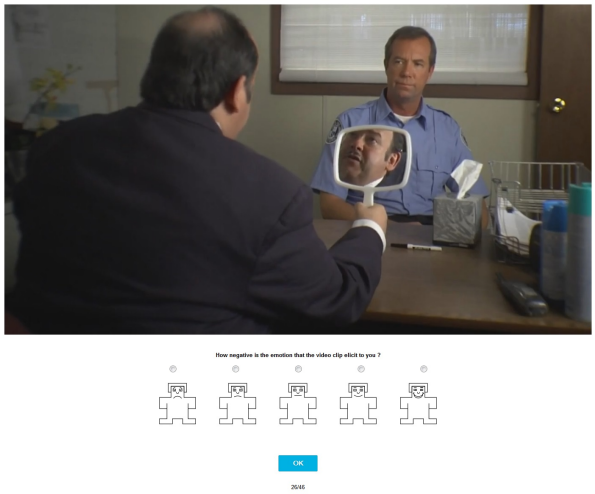
5.1.3 Analysis of the Annotations

Fig. 5.2 shows the distribution of the ratings of valence and arousal, suggesting that negative film clips were rated as more arousing than positive ones. This correlation is not surprising since Lang *et al.* showed that only specific areas of the 2D valence-arousal space are relevant [95]. The distribution displayed in Fig. 5.2 is also similar to those depicted in previous works dealing with the affective impact of multimedia content [50, 30, 52] , except that the distributions illustrated in these works show much more data eliciting positive and arousing emotions (see Section 5.2.2 for further discussion).

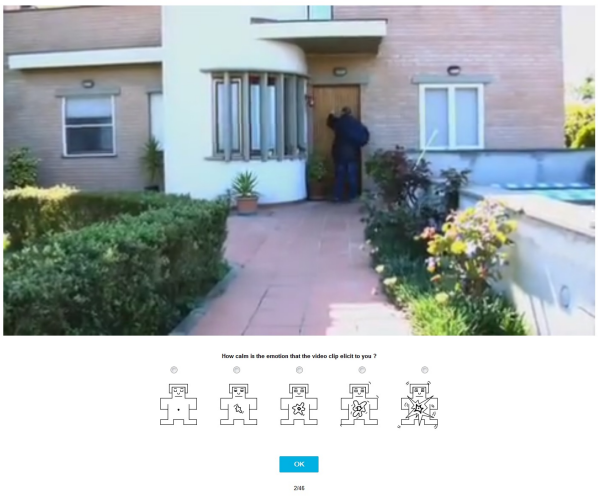
Globally, the mean standard deviation of the ratings is higher for arousal ($SD = 0.771$) than for valence ($SD = 0.631$), indicating that participants agreed more when assessing valence. It is confirmed by the Krippendorff's alpha, measuring the inter-annotator agreement, which is higher for the self-assessment of the level of valence ($\alpha = 0.282$) than for the self-assessment of their level of arousal ($\alpha = 0.225$). Both values are



(a) Instructions given before the self-assessment for valence



(b) Round 1: Valence



(c) Round 2: Arousal

Figure 5.1 – *Screenshots of the interface used for the experiment.*

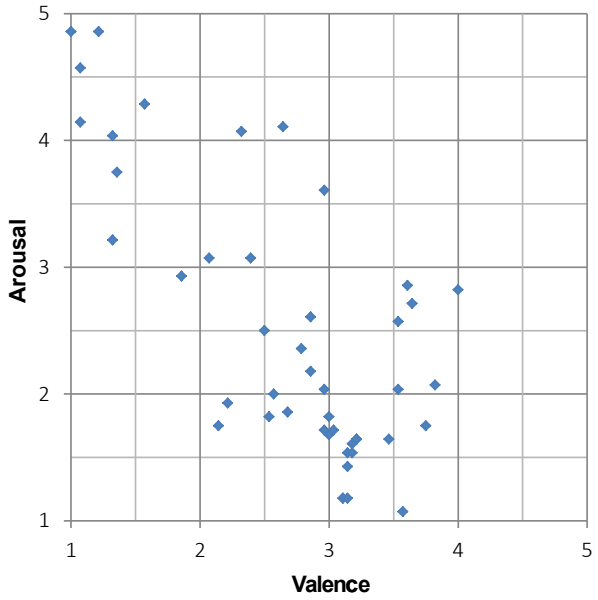


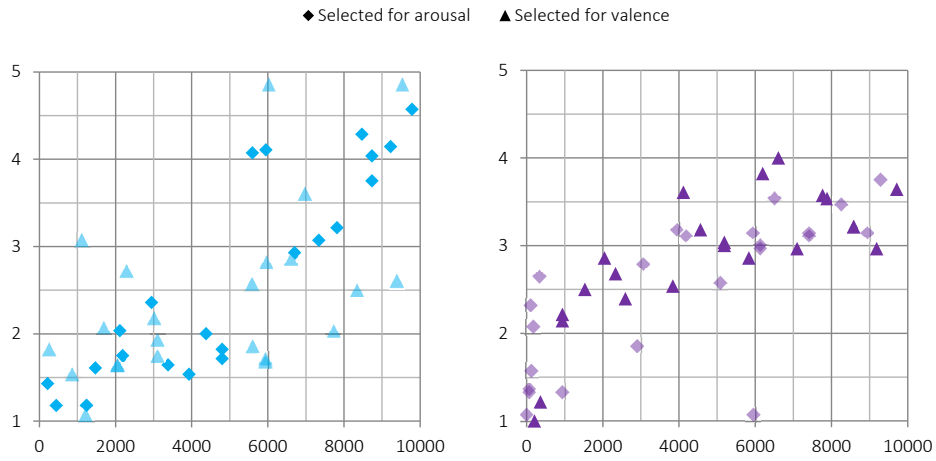
Figure 5.2 – *Distribution of the 46 film clips in the affective space (mean values for valence and arousal).*

positive which indicates that there is an agreement between annotators despite the subjectivity of the experiment and are comparable to other studies dealing with affective computing [52, 78]. A two-factor (Women, Men) ANOVA failed to reveal significant gender differences. It is interesting to mention that another two-factor (AM, PM) ANOVA revealed that participants who started the experiment in the afternoon tend to report greater levels of arousal ($F = 30.1, p = 1.79 \times 10^{-6}$). This observation is consistent with the findings of Soleymani *et al.* indicating that average arousal ratings in response to videos increase with time of day [70].

The intra-rater reliability can also be computed thanks to the six film clips that have been annotated twice by each annotator. The mean-square error (MSE) of the ratings of the duplicated film clips is very low for valence ($MSE = 0.002$) as well as for arousal ($MSE = 0.021$) meaning that the repeatability of the experiment is high for a short period of time and consequently that annotators understood the scales and did not answer at random. To qualify these high intra-rater reliabilities, it is worth considering that due to the short duration of the experiment, some participants could have remembered the score given to the first occurrence of the video clip, thus reducing the impact of this criterion. Nevertheless, it is consistent to use these ratings to validate the crowdsourced annotations of the LIRIS-ACCEDE dataset.

5.2 Cross-validation & bias of LIRIS-ACCEDE

The results from the controlled rating experiment presented in this chapter allow to cross-validate the dataset that has been previously ranked in the affective space thanks to numerous crowdsourced pairwise annotations gathered in Chapter 4 from various uncontrolled environments.



(a) Arousal (b) Valence

Figure 5.3 – *Correlation between rankings (horizontal axis) and ratings (vertical axis) for both arousal and valence for the 46 films clips. A distinction is made between the 23 film clips selected for arousal, i.e., the excerpts that are highly reliable in eliciting arousal based on their Krippendorff's alpha computed using the crowdsourced annotations of arousal, and the 23 others selected for valence that are highly reliable in eliciting valence.*

They also allow to better understand the distribution and the bias of the dataset in the affective space.

5.2.1 Cross-validation

A t-test revealed that the Spearman's rank correlation coefficient (SRCC) between the rankings of the 46 film clips in the LIRIS-ACCEDE dataset and the ratings collected in this experiment exhibits a statistically highly significant correlation for both arousal ($SRCC = 0.751, t(44) = 7.635, p < 1 \times 10^{-8}$) and valence ($SRCC = 0.795, t(44) = 8.801, p < 1 \times 10^{-10}$). It indicates that the annotations gathered in an uncontrolled environment using crowdsourcing are highly correlated with the ratings gathered in a controlled environment. Fig. 5.3 shows that the excerpts selected for a specific axis are even more correlated with this axis than the other excerpts. Consequently, this new experiment in a controlled environment validates the annotations gathered using crowdsourcing that lead to the ranking of the LIRIS-ACCEDE dataset.

5.2.2 Discussion

The results from the experiment also exhibit a bias in the dataset. Indeed, the distribution of the ratings for valence (see Fig. 5.2) shows that there are no film clips inducing high valence, which could be due to several factors.

First, it has been shown in previous work [96, 97] that positive evaluations were more subjective than negative ones. As a consequence, people agree more when they rate negative emotions than positive emotions. The plot of valence self-assessments corroborates the tendency that video clips that make people feel negative emotions elicit more consistent ratings than those that make people feel positive emotions (see Fig. 5.4). Since

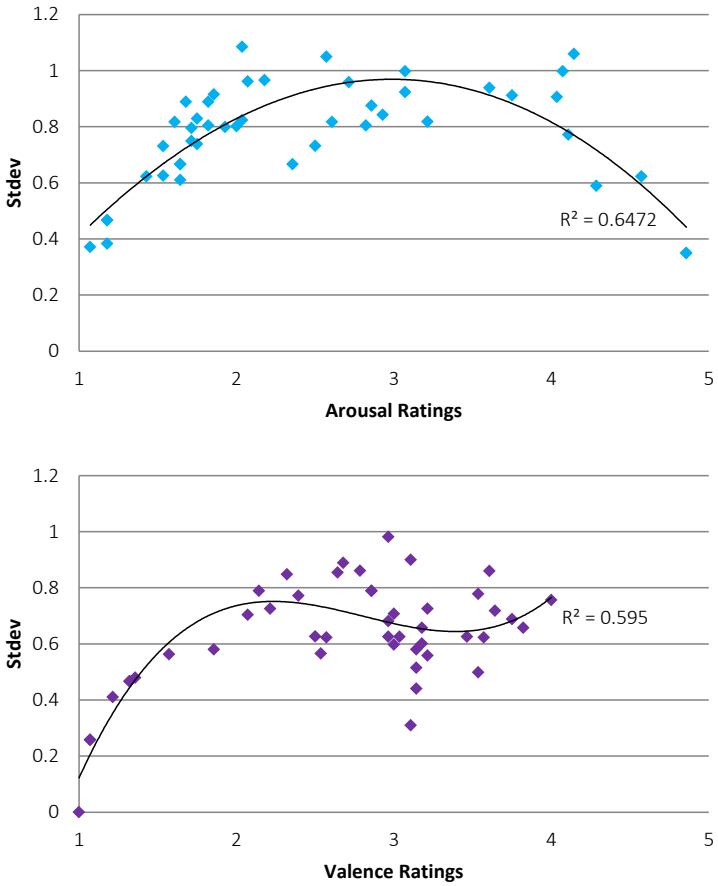


Figure 5.4 – *Standard deviations for both arousal and valence ratings for the 46 films clips and the associated best third-degree polynomial fitting curves. The coefficient of determination of the trend-lines is also indicated.*

the ground truth is obtained by averaging subjects' ratings, the larger the standard deviation, the smoother the final value to a neutral value. In contrast to valence, Fig. 5.4 shows that the standard deviations of the ratings of the video clips eliciting extreme arousal (calm or excited) are lower than neutral ones.

Second, this bias may also be due to the fact that no movie induces high valence. However, 15% of the excerpts (1,477 film clips) in LIRIS-ACCEDE have been extracted from 25 comedy films. Major genres represented in the dataset are drama (28%), action/adventure films (16%), comedies (15%) and documentaries (14%) which are representative of current most popular movie genres. Contrarily to other affective video databases [45, 43], the excerpts in the dataset have been automatically segmented and thus have not been preselected in order to cover the whole affective space. But it seems highly unlikely that no excerpt or at least no scene in the selected movies induces high valence. Finally, another explanation of this bias is that it may be more challenging to induce very positive emotions in a short time than negative emotions. Indeed, the length of excerpts in the LIRIS-ACCEDE dataset varies from 8 to 12 seconds which may not be sufficient to elicit very positive emotions.

The rating experiment also reveals that the dataset suffers from another

bias to the extent that there are less film clips with positive valence inducing high arousal, making the dataset asymmetrical. This bias can be found in other databases such as the EMDB dataset introduced by Carvalho *et al.* [47] that claimed that it is related to the existence of stronger response and attentional allocation to negative stimuli [98]. However, Lang *et al.* showed that the sexual stimuli included in the IAPS dataset elicited the most arousing and positive emotional reactions [95]. Because of ethical concerns, such sexual content is not included in the publicly available LIRIS-ACCEDE dataset, which might also partially explain the lack of highly arousing and positive content in the dataset.

5.3 Regression analysis

The goal of this section is to use the rankings and ratings available for the 40 video clips annotated in Section 5.1 to perform a regression analysis between the rankings and the ratings to convert the relative rankings into absolute scores.

Among all existing regression models, we used the Gaussian Processes for Regression as they can model the noisiness from measurements and thus take into account the subjectivity of emotions.

Two different Gaussian Process Regression Models are learned in this part, one for the valence axis and a second one for arousal. From the rank given as input (ranging from 0 to 9,799), the goal of the models is to predict its affective rating for the dedicated axis (ranging from 1 to 5). To learn the models, we will use the crowdsourced ranks and the corresponding affective ratings gathered in Section 5.1. The variance of the annotations gathered in this controlled rating experiment will be used to provide guidance to learn the models. Thus, these variances will be needed only during the learning step and will no longer be necessary to predict new affective ratings.

Knowing the rank x for valence or arousal of a video clip in the database, the goal of the Gaussian Processes regression models is to estimate the score $g(x)$ of the video clip in the affective space.

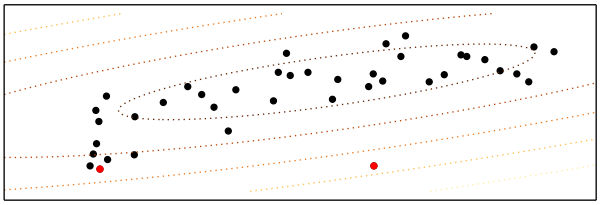
Knowing the rank x for valence or arousal of a video clip in the database, the goal of the Gaussian Processes regression models is to estimate the score $g(x)$ of the video clip in the affective space. Rasmussen and Williams [99] define a Gaussian Process (GP) as a collection of random variables, any GP finite number of which have a joint Gaussian distribution. The predictions from a GP model take the form of a full predictive distribution:

$$g(x) = f(x) + \mathbf{h}(x)^T\beta, \text{ with } f(x) \sim GP(0, k(x, x')) \tag{5.6}$$

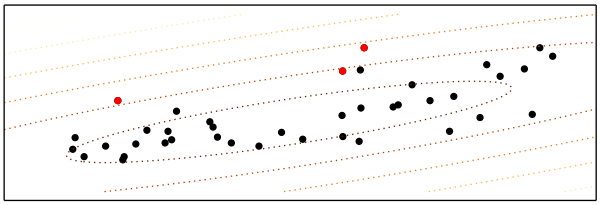
where $f(x)$ is a zero mean GP, $\mathbf{h}(x)$ are a set of fixed basis functions, and β are additional parameters. For valence we used linear basis functions whereas quadratic basis functions were selected for arousal.

We used the squared exponential kernel for which, during interpolation at new values, distant observations will have negligible effect:

$$k(x, x') = \sigma_f^2 \times exp\left(\frac{-(x - x')^2}{2l^2}\right) + \sigma_n^2\delta(x, x') \tag{5.7}$$



(a) Valence



(b) Arousal

Figure 5.5 – *Mahalanobis distances between the 40 video clips and the estimated center of mass, with respect to the estimated covariance in the ranking/rating space. Red points are the video clips considered as outliers.*

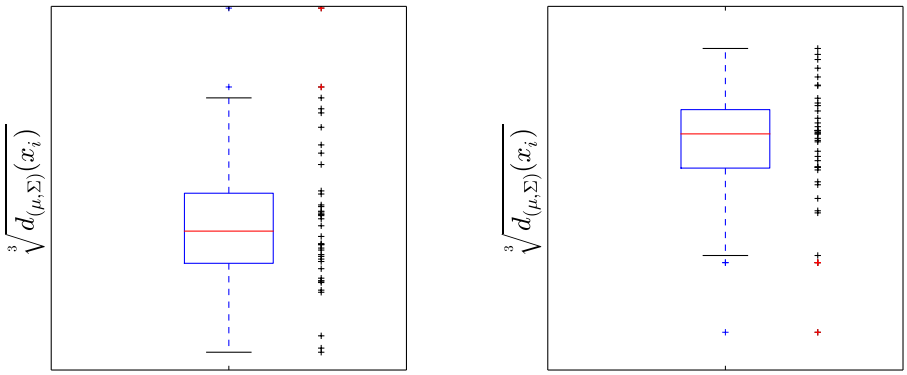
where the length-scale l and the signal variance σ_f are hyperparameters, σ_n is the noise variance and $\delta(x, x')$ is the Kronecker delta. All the parameters are estimated using the maximum likelihood principle. In this work, σ_n values are not hyperparameters since they represent the known variance of annotations gathered in the controlled rating experiment described in section 5.1. They are added to the diagonal of the assumed training covariance. As a consequence, the GP is also able to model the subjectivity of emotions from this experiment.

5.4 Outlier detection

To perform a regression analysis on clean data, the first step is to detect outliers.

The Minimum Covariance Determinant (MCD) estimator introduced by Rousseeuw in [100] is a highly robust estimator for estimating the center and scatter of a high dimensional data set without being influenced by outliers. Assuming that the inlier data are Gaussian distributed, it consists in finding a subset of observations whose empirical covariance has the smallest determinant. The MCD estimate of location μ is then the average of the "pure" observations in the selected subset and the MCD estimate of scatter is their covariance matrix Σ. In this work, we used the fast MCD algorithm implemented in [101] to estimate the covariance of the 40 video clips defined in section 5.1, described in the 2D ranking/rating space by their rank and rating score.

Once the center and covariance matrix have been estimated, the Mahalanobis distance of centered observations can be computed. It provides a relative measure of an observation from the center of mass taking into account the correlation between those points. The Mahalanobis distance



(a) Valence (b) Arousal

Figure 5.6 – *Box plot of the Mahalanobis distances for valence and arousal. The whiskers show the lowest and highest values still within the 1,5 IQR. Red points are the video clips considered as outliers.*

of a video clip x_i is defined as:

$$d_{(\mu,\Sigma)}(x_i)^2 = (x_i - \mu)\Sigma^{-1}(x_i - \mu) \tag{5.8}$$

where μ and Σ are the estimated center of mass and covariance of the underlying Gaussian distribution. Figure 5.5 shows the shape of the Mahalanobis distances for the valence and arousal data sets.

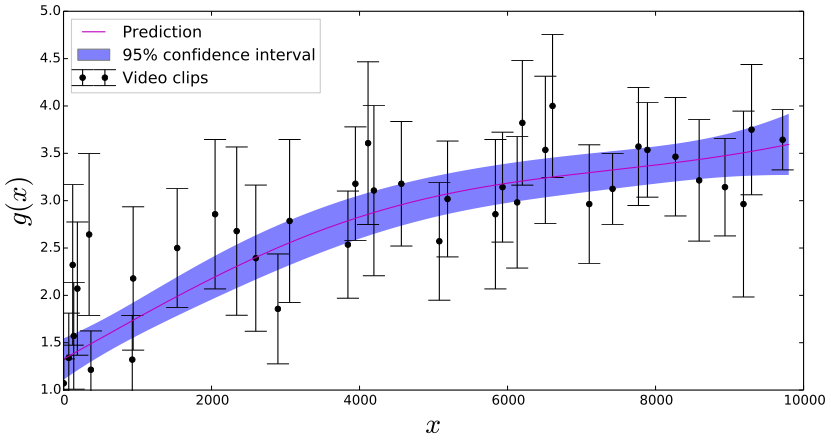
By considering the covariance of the data and the scales of the different variables, the Mahalanobis distance is useful for detecting outliers in such cases. As a rule of thumb, a video clip x_i is considered as an outlier if $d_{(\mu,\Sigma)}(x_i) < Q1 - 1.5 \times IQR$ or if $d_{(\mu,\Sigma)}(x_i) > Q3 + 1.5 \times IQR$ with $Q1$ and $Q3$ the first and third quartiles and IQR the Inter-quartile Range. The boxplots showing the outliers detected for valence and arousal during this process are illustrated in Figure 5.6.

In our experiments, two video clips are categorized as outliers for valence and three video clips for arousal. Thus, these video clips are removed from the data sets in order to perform a regression analysis only on "clean" data sets. As a consequence, 38 video clips are used to perform the regression analysis for valence while for arousal the data set is composed of 37 video clips.

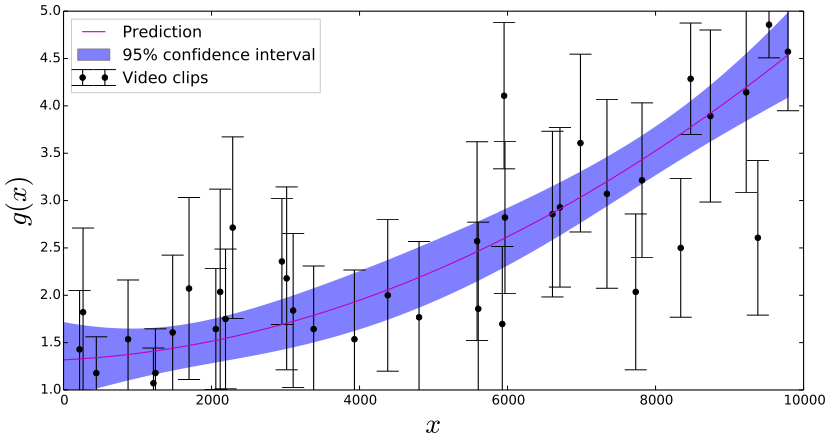
5.5 Results

To measure the prediction power of the learned regression models depicted in Figure 5.7, we calculated in addition to the well-known conventional squared correlation coefficient R^2, the predictive leave-one-out squared correlation coefficient Q^2_{loo} defined as:

$$Q^2_{loo} = 1 - \frac{\sum_{i=1}^{N} \left(y_i^{pred(N-1)} - y_i\right)^2}{\sum_{i=1}^{N} \left(y_i - y_{mean}^{N-1,i}\right)^2} \tag{5.9}$$



(a) Valence



(b) Arousal

Figure 5.7 – *Gaussian Process Models learned for valence and arousal converting ranks (horizontal axis) into ratings (vertical axis). Black bars show the variance of the annotations.*

with y_i the true rating value of the video clip i and $y_i^{pred(N-1)}$ the prediction of the model learned with the initial training set from which the video clip i was removed. Note that the arithmetic mean used in equation (5.9), $y_{mean}^{N-1,i}$, is different for each test set and calculated for the observed values comprised in the training set.

R^2 measures the goodness of fit of a model while Q_{loo}^2 computed using the leave-one-out cross-validation technique measures the model prediction power. Both values for valence and arousal are shown in Table 5.1.

These results are remarkably high considering that we are modeling crowdsourced ranks and affective ratings that are both subject to the subjectivity of human emotions. Thus, our proposed regression models successfully learned to fit input observations. Furthermore, Q_{loo}^2 values show that the models are also able to provide valid predictions for new observations.

Table 5.1 – *Performance of the Gaussian Process Models learned predicting valence and arousal.*

Measure	Valence	Arousal
R^2	0.657	0.632
Q^2_{loo}	0.621	0.586

Summary

This chapter addressed the validation of the LIRIS-ACCEDE affective video dataset and showed that it is possible to estimate absolute values in the emotional space using affective ranks while taking into account the subjectivity of emotions.

First, we have proposed an experimental protocol consisting in collecting ratings for a subset of the dataset using the Self-Assessment Manikin (SAM) scales in a controlled setup. This subset consists of 40 excerpts that have been carefully selected based on their reliability to induce emotions during the crowdsourced experiment. The results have shown that the correlation between affective ratings and crowdsourced rankings is significantly high thus validating the overall dataset for future uses in research works.

Based on these results, we have been able to enrich the LIRIS-ACCEDE database by providing in addition to video rankings that were already available, video ratings thanks to a regression analysis that allows mapping all the 9,800 video clips included in the dataset into the 2D valence-arousal affective space. The Gaussian process regression models, taking into account the variance of the annotation of the absolute scores, achieved a good performance, confirming our intuition that absolute scores in the affective space can be estimated using relative ranks.

Extension to Time-Continuous Annotations

6

Contents

LIRIS-ACCEDE, the dataset presented in Chapter 4, proposes 9,800 excerpts extracted from 160 movies. However, these 9,800 excerpts have been annotated independently, limiting their use for learning models for longer movies where previous scenes may reasonably influence the emotion inference of future ones.

Thus, we set up a new experiment where annotations are collected on long movies, making possible the learning of more psychologically relevant computational models.

6.1 Movie Selection

The aim of this new experiment is to collect continuous annotations on whole movies. To select the movies to be annotated, we simply looked at the movies included in the LIRIS-ACCEDE dataset [1] since they all share the desirable property to be shared under Creative Commons licenses and can thus be freely used and distributed without copyright issues as long as the original creator is credited. The total length of the selected movies was the only constraint. It had to be smaller than eight hours to create an experiment of acceptable duration.

1. An exhaustive list of the movies included in the LIRIS-ACCEDE dataset as well as their credits and license information is listed in Appendix A

The selection process ended with the choice of 30 movies so that their genre, content, language and duration are diverse enough to be representative of the original LIRIS-ACCEDE dataset. The selected videos are between 117 and 4,566 seconds long ($mean = 884.2sec \pm 766.7sec\ SD$). The total length of the 30 selected movies is 7 hours, 22 minutes and 5 seconds. The list of the 30 movies included in this experiment is detailed in Table 6.1.

6.2 Experimental Design

The annotation process aims at continuously collecting the self-assessments of arousal and valence that viewers feel while watching the movies.

6.2.1 Annotation tool

To collect continuous annotations, we have used a modified version of the GTrace program originally developed by Cowie *et al.* [102]. GTrace has been specifically created to collect annotations of emotional attributes over time. However, we considered that the design of the original GTrace interface during the annotation process is not optimal: the video to be rated is small, the annotation scale is far from it, and other elements may disrupt the annotator's task. That is why we modified the interface of GTrace in order to be less disruptive and distract annotators' attention from the movie as little as possible.

First, we redesigned the user-interface so that the layout is more intuitive for the annotator. During the annotation process, the software is now in full screen and its background is black. The video is bigger, thus more visible, and the rating scale is placed below the video (Figure 6.1(b)).

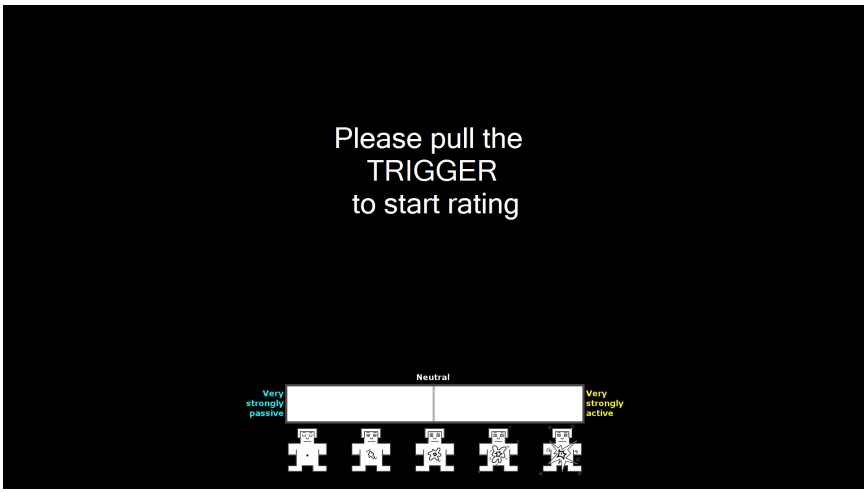
Second, we used the possibility offered by GTrace to create new scales. We designed new rating scales for both arousal (Figure 6.1(a)) and valence (Figure 6.1(b)). Under both scale, the corresponding SAM scale is displayed [87].

Third, instead of using a mouse, the annotator used a joystick to move the cursor which is much more intuitive. To link the joystick to GTrace, we used a software that simulates the movement of the mouse cursor when the joystick is used.

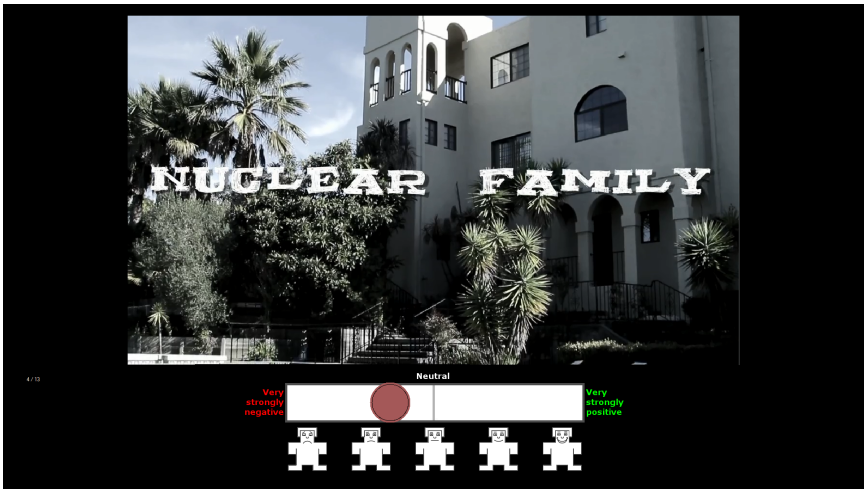
6.2.2 Protocol

In the experimental protocol described below, each movie is watched by an annotator only once. Indeed, the novelty criterion that influences the appraisal process for an emotional experience should be taken into consideration [103].

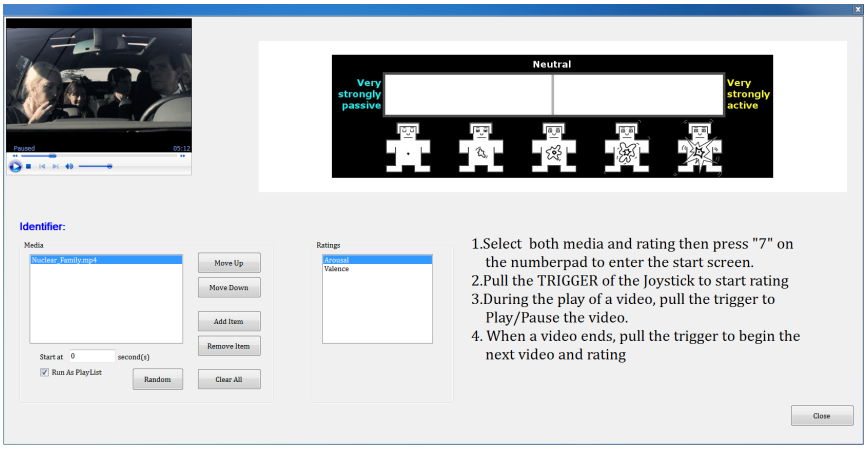
Annotations were collected from ten French paid participants (seven female and three male) ranging in age from 18 to 27 years ($mean = 21.9 \pm 2.5\ SD$). Participants had different educational backgrounds, from undergraduate students to recently graduated master students. The experiment was divided into 4 sessions, each took place on a different half-day. The movies were organized into 4 sets (Table 6.1). Before the first session,



(a) Screenshot before the annotation along the arousal axis



(b) Screenshot during the annotation along the valence axis



(c) Modified GTrace menu

Figure 6.1 – *Screenshots of the modified GTrace annotation tool. Nuclear Family is shared under a Creative Commons Attribution-NonCommercial 3.0 Unported United States License at http://dominicmercurio.com/nuclearfamily/.*

Table 6.1 – *List of the 30 movies on which continuous annotations have been collected*

Sets	Duration	Movies
A	01:50:14	Damaged Kung Fu, Tears of Steel, Big Buck Bunny, Riding The Rails, Norm, You Again, On time, Chatter, Cloudland & After The Rain
B	01:50:03	Barely Legal Stories, Spaceman, Sintel, Between Viewings, Nuclear Family, Islands, The Room of Franz Kafka & Parafundit
C	01:50:36	Full Service, Attitude Matters, Elephant's Dream, First Bite, Lesson Learned, The Secret Number & Superhero
D	01:51:12	Payload, Decay, Origami, Wanted & To Claire From Sonny

participants were informed about the purpose of the experiment. They had to sign a consent form and fill a questionnaire (Appendix C). Participants were trained to use the interface thanks to three short videos they had to annotate before starting the annotation of the whole first session. The participants were also introduced to the meaning of the valence and arousal scales.

Participants were asked to annotate the movies included in the first two sessions along the induced valence axis and the movies in the last two sessions along the induced arousal axis. This process ensures that each movie is watched by an annotator only once. The order of the sets with respect to the four sessions was different for all the annotators. For example, the first participant annotated the movies from sets A and B along the induced valence axis and the movies from sets C and D along the induced arousal axis whereas the second participant annotated the movies from sets B and C along the induced valence axis and the movies from sets D and A along the induced arousal axis. Furthermore, the videos inside each session were played randomly. After watching a movie, the participant had to manually pull the trigger of the joystick in order to play the next movie.

Finally, each movie is annotated by five annotators along the induced valence axis and five other annotators along the induced arousal axis.

6.3 Post-processing

Defining a reliable ground truth from continuous self-assessments from various annotators is a critical aspect since the ground truth is used to train and evaluate emotion prediction systems. Two aspects are particularly important: there are annotator-specific delays amongst the annotations and the aggregation of the multiple annotators' self-assessments must take into account the variability of the annotations [74].

Several techniques have been investigated in the literature to deal with the synchronisation of various individual ratings. In this work, we com-

bine and adapt the approaches proposed by Mariooryad and Busso [104] and by Nicolaou *et al.* [53] to deal with both the annotation delays and variability.

First, the self-assessments recorded at a rate of 100 values per second are down-sampled by averaging the annotations over windows of 10 seconds with 1 second overlap (*i.e.* 1 value per second). This process removes most of the noise mostly due to unintended moves of the joystick. Furthermore, due to the granularity of emotions, one value per second is enough for representing the emotions induced by movies [105, 74].

Then, each self-assessment is shifted so that the τ-sec-shifted annotations maximize the inter-rater agreement between the τ-sec-shifted self-assessment and the non-shifted self-assessments from the other raters. The inter-rater agreement is measured using the Randolph's multirater kappa free [90]. Similarly to Mariooryad and Busso [104], the investigated delay values τ range from 0 to 10 sec. However, in practice, τ ranged from 0 to 6 sec and the largest values (5 or 6 sec) were rarely encountered ($mean = 1.47 \pm 1.53\ SD$). As suggested by Landis and Koch [91], the average Randolph's multirater kappa free shows a moderate agreement for the shifted arousal self-assessments ($\kappa = 0.511 \pm 0.082\ SD$), as well as for the shifted valence self-assessments ($\kappa = 0.515 \pm 0.086\ SD$).

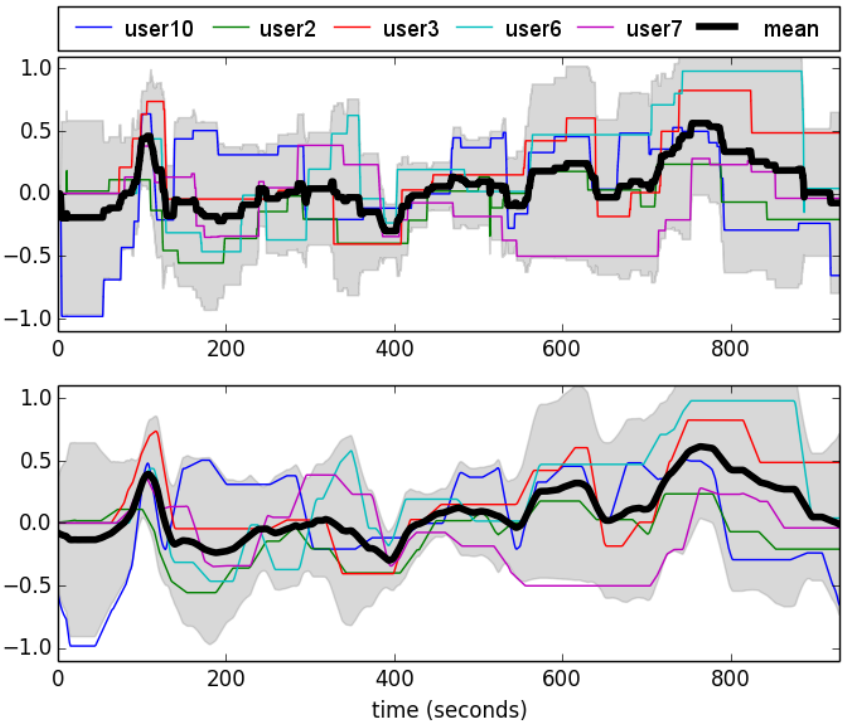
Finally, to aggregate the different ratings we use an approach similar to the one proposed in [53]. The inter-coder correlation is used to obtain a measure of how similar are one rater's self-assessments to the annotations from the other participants. The inter-coder correlation is defined as the mean of the Spearman's rank correlation coefficients (SRCC) between the annotations from the coder and each of the annotations from the rest of the coders. The SRCC has been preferred over other correlation measures since it is defined as the Pearson correlation coefficient between the ranked variables: the SRCC is computed on relative variables and thus ignores the scale interpretation from the annotators. The inter-coder correlation is used as a weight when combining the multiple annotators' annotations. The inter-coder correlation is higher in average for valence ($mean = 0.313 \pm 0.195\ SD$) than for arousal ($mean = 0.275 \pm 0.195\ SD$).

Figure 6.2 shows the raw ratings and post-processed ones for both induced arousal and valence scales for the movie Spaceman. The bold curves are the weighted average of the continuous annotations computed on the raw ratings or on the smoothed and shifted ones.

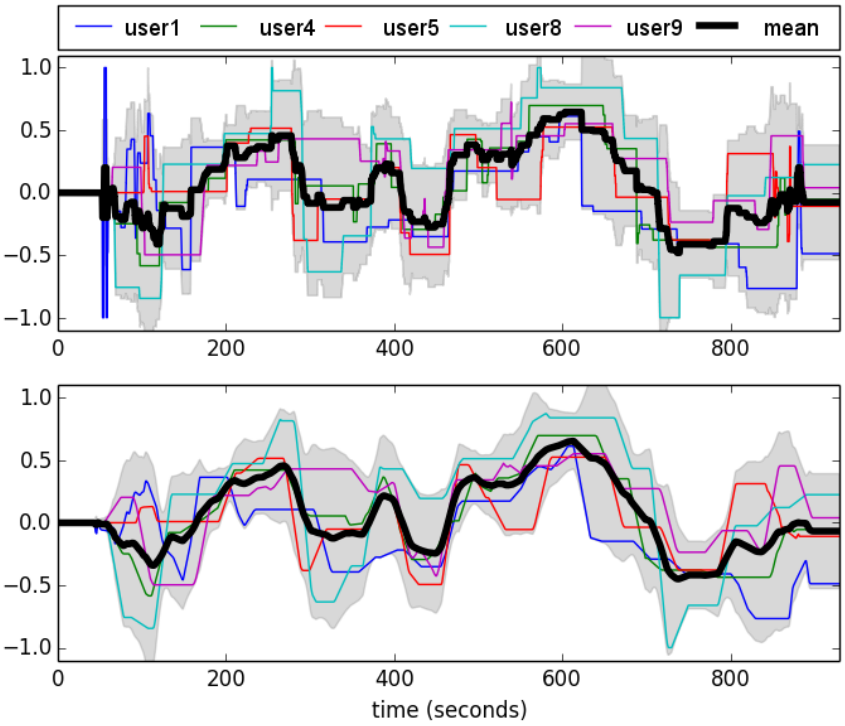
To conclude, this post-processing assigns two values at each 1-second segment of a movie: one represents the induced arousal and the other the induced valence. Both values are rescaled so that they range from 0 to 1. More precisely, 26,525 1-second segments are extracted from the 30 movies.

Summary

This chapter introduces a new dataset composed of 30 movies continuously annotated along the induced valence and arousal axes split into 26,525 1-second length video segments. The use of both joysticks and of a modified version of the GTrace annotation tool has improved the user experience of the experiment in order to be less disruptive. The annotations



(a) Raw and post-processed annotations for arousal



(b) Raw and post-processed annotations for valence

Figure 6.2 – *Annotations collected for the movie "Spaceman". Both subfigures show at the top the raw annotations and at the bottom post-processed annotations for (a) arousal and (b) valence. The shaded area represents the 95% confidence interval of the mean.*

have been post-processed to take into account the annotator-specific delays and the variability of the annotations when aggregating the multiple annotators' self-assessments.

In the second part of this thesis, the LIRIS-ACCEDE dataset will be used to compute the first baseline in Chapter 8. The continuous annotations for the 30 full-length movies presented in this chapter will be used to compute the next four baselines in Chapter 8, as well as the spatio-temporal model introduced in Chapter 9.

The Galvanic Skin Response as a Temporal Arousal Indicator

7

Contents

The recording of physiological signals is an alternative to the direct affective self-assessment of each viewer for which the continuous data collection is not realistic for large scale consumer applications. Indeed, physiological signals are associated with the felt emotions. Hence, physiological signals can be used to cross-validate affective self-assessments, and can also be used as an additional modality to help emotion detection.

In order to cross-validate the continuous annotations collected in Chapter 6, and to extend the range of applications of the LIRIS-ACCEDE dataset, we present in this chapter an experiment we conducted to collect physiological measurements, *i.e.* the Galvanic Skin Response, from participants. In this experiment, participants watched the same 30 movies used in the previous chapter to collect continuous affective self-assessments, with sensors attached to their fingers. In particular, one potential application of the use of these physiological signals for emotion detection is the automatic implicit tagging of videos, which is more robust using a physiological-based tagging than a content-based tagging only [106], but also more obtrusive since users have to wear specific sensors.

Figure 7.1 – *The Bodymedia armband used to record the GSR of participants (illustration from http://www.bodymedia.com/)*

7.1 The experiment

7.1.1 Physiological signals

Many physiological signals are known to be correlated with the emotional state, such as the EEGs, the heart rate variability, or the Galvanic Skin Response (GSR) [45, 107]. However, collecting physiological signals is often obtrusive. Some are easier to measure, less obtrusive and thus more appropriate for experiments dealing with a large number of participants. To study the emotional impact of a movie on an audience, Fleureau *et al.* measured the GSR of a total of 128 audience members [108]. GSR was chosen among other physiological measures because it can be measured in a non-obtrusive way thanks to compact sensors, and is known to be related to the level of arousal [95, 60]. For these reasons, we also measure in this chapter the GSR of participants watching a set of 30 movies from the LIRIS-ACCEDE dataset in order to offer new possibilities to the users of the dataset. The GSR, also known as the electrodermal activity, measures the variations in the electrical characteristics of the skin. These fluctuations vary with the state of the sweat glands in the skin.

To measure the GSR, we used the Bodymedia armband illustrated in Figure 7.1. We used this armband because it is user friendly, thus users rapidly understand how to wear the armband by themselves before the experiments. Users only have to place the armband on their fingers. In contact with the palm area, the armband turns on automatically and starts recording the GSR of the participants. This device allowed us to record, in addition to the GSR, the motion of the fingers and skin temperature. However these two supplementary measures are not analyzed in the following sections but may be used by future researchers using this dataset.

7.1.2 Experimental protocol

In this new experiment, we reused the 30 movies that have been continuously annotated in the previous chapter along the induced valence and

arousal dimensions. The 30 movies are between 117 and 4,566 seconds long ($mean = 884.2sec \pm 766.7sec\ SD$), and the total length is 7 hours, 22 minutes and 5 seconds. The movies are fairly distributed among four sets. The list of the 30 movies for each set is detailed in Table 6.1.

The goal of this new experiment is to record physiological signals (in particular the GSR) from spectators. Annotations were collected from 13 French paid participants (11 female and 2 male) ranging in age from 22 to 45 years ($mean = 36.1 \pm 6.7\ SD$). The experiment was divided into four sessions, each took place on a different half-day to reduce the cognitive load of the participants. Each of these sessions lasted approximately two hours, including the setup of the armbands. Before the first session, participants had to sign a consent form similar to the one in Appendix C. They were informed that the device used to record the signals was perfectly safe and that their anonymity would be protected.

The physiological signals were recorded simultaneously from the 13 participants watching the videos in a darkened air-conditioned amphitheater. They were free to select their own position in the room. Before the first session, an armband sensor was placed on the tops of two of their fingers. Instructions were given so that participants could setup the device by themselves at the beginning of the three last sessions.

7.2 Correlation with arousal self-assessments

As pointed out by Fleureau *et al.* [108], the GSR signal must be processed to remove the misleading information before being compared to the continuous arousal self-assessments.

7.2.1 Post-processing of the GSR signals

To process the GSR signals from the 13 participants and for the 30 movies, we used the algorithm described in [108] with minor changes:

1. First, for a given participant and a given movie, a low-pass filtering is applied (Finite Impulse Response filter with a 2Hz cutoff frequency) and the signal is derived.
2. As in [108], a thresholding is computed in order to remove the negative values of the derivative of the signal and to focus only on the positive segments. Indeed, positive values represent the increase of the skin conductance measured by the armband due to an increase of sweating, usually related to highly-arousing emotions [109].
3. The signal is temporally filtered and sub-sampled using a 10 seconds time window with 5 seconds overlap to obtain a time resolution of 5 seconds. The time resolution is similar to previous studies [110] and smaller than the one used in [108] to make it compliant with the shorter duration of the movies used in this experiment.
4. Finally, the signal is normalized to remove the user-dependent part related to the amplitude of the derivative of the GSR.

Using the same terminology as Fleureau *et al.* [108], the resulting signal for each user i, $1 \leq i \leq N$ (with $N = 13$, the number of participants), is called individual GSR profile and termed p_n^i.

7.2.2 Weighted mean GSR profile

Once the raw signals are post-processed, outliers can be detected more easily. There are several factors that could have affected the reliability of the signals measured with the armbands. Indeed, an experimenter checked if the armbands were correctly set up on the fingers at the beginning of each session. However, it is possible that during the sessions some participants moved their armband, although participants were instructed not to touch or interact with the device. Some signals may also be too noisy due to repeated moves of the hand on which the armband was fastened. For example, it is tempting for a participant to check his smartphone and reply to a text message if the movie is boring.

For each movie, the Spearman's rank correlation coefficient (SRCC) ρ_{ij} between the individual GSR profiles p_n^i and p_n^j of two users are computed. The agreement α_i of an annotator i for a single movie is defined as:

$$\alpha_i = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \rho_{ij} \tag{7.1}$$

A participant i is considered as an outlier if its agreement α_i is smaller than $\mu_a - \frac{1}{2}\sigma_a$, with μ_a and σ_a respectively the mean and standard deviation of the SRCCs between each of the participants and all the other participants, *i.e.*:

$$\mu_a = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \rho_{ij} \tag{7.2}$$

$$\sigma_a = \sqrt{\frac{1}{N(N-1)} \left(\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \rho_{ij}^2 \right) - \mu_a^2} \tag{7.3}$$

In average, 4 participants are discarded per movie. As mentioned by Grimm and Kroschel [111], as the remaining noise influences might be different for the different participants, it is reasonable to compute the mean GSR signal using an individual confidence score. Finally, the weighted mean GSR profile $\hat{p_n}$ for a movie is thus defined as:

$$\hat{p_n} = \frac{1}{\sum_{i \in V} c_i} \left(\sum_{i \in V} c_i p_n^i \right) \tag{7.4}$$

with V the set of the participants that are not considered as outliers, N_v the number of participants in V, and c_i the confidence score generated similarly to [111] by computing the Pearson's correlation between the individual affective profile p_n^i, and the mean GSR profile $\bar{p_n} = \frac{1}{N_v} \sum_{i \in V} p_n^i$.

7.2.3 Derived GSR and arousal peaks

Based on the intuition that the weighted mean GSR profile, generated using the derivative of the GSR measurements, should be temporally correlated with the increase or decrease of arousal, we first computed the temporal correlation between the weighted mean GSR profile and the

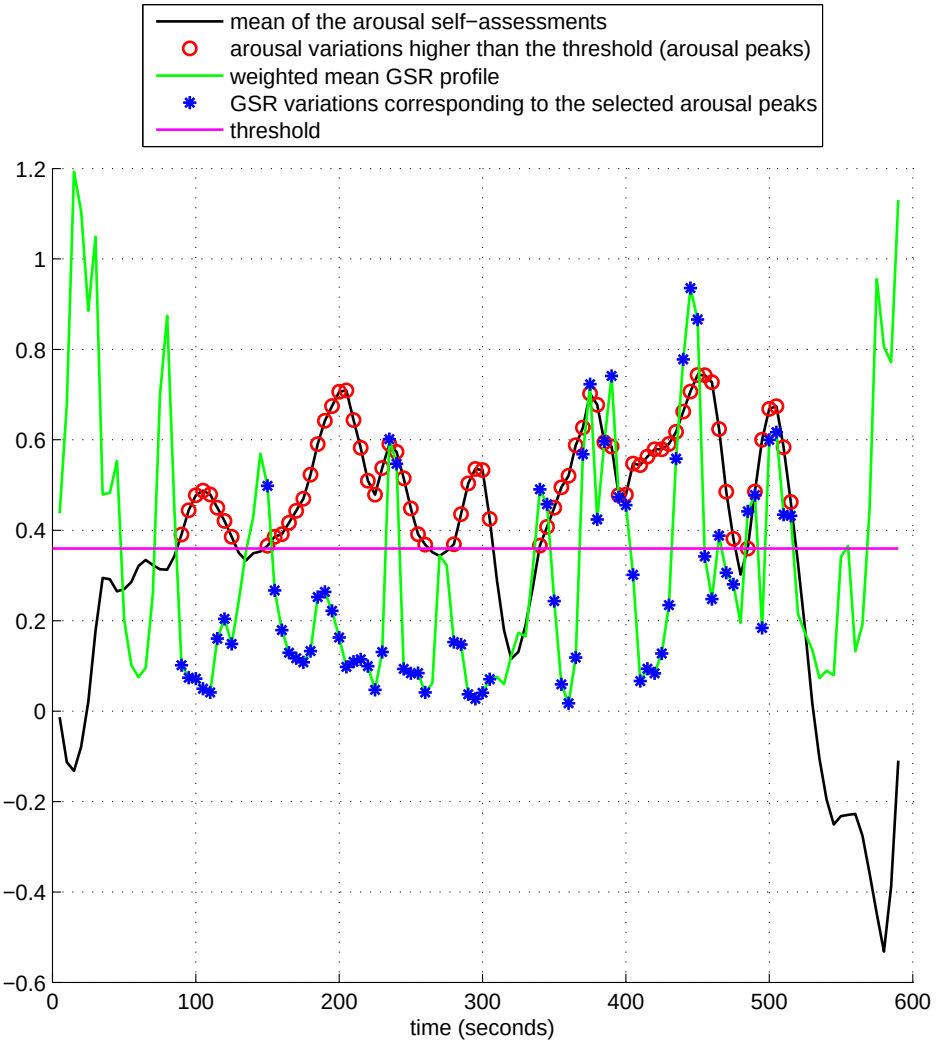


Figure 7.2 – *Example of the computation of arousal peaks from the mean arousal self-assessments for the movie Big Buck Bunny*

mean of the arousal self-assessments described in Chapter 6. However, results were not as good as expected since the mean between these Person's correlations for the 30 movies is rather low ($r = 0.070$). By introducing a threshold to compute these correlations only for the most intense parts of the movies, we show in this section that for most of the movies, the weighted mean GSR profile is in fact correlated with the arousal self-assessments during the most arousing scenes.

The most intense parts of a movie, called "arousal peaks" in the remaining of this chapter, are defined as the highest scores of the mean arousal self-assessments with respect to a threshold T. T is the percentage of the smallest values to be removed from the process. The continuous mean of the arousal self-assessments of the 30 movies has been subsampled using an overlapping time window to match the time resolution of 5 seconds of the weighted mean GSR profile $\hat{p_n}$. Figure 7.2 shows the arousal peaks (red circles) for the movie Big Buck Bunny for $T = 40\%$, *i.e.* 60% of the greatest values have been kept. The blue stars indicate

Table 7.1 – *Pearson's r and SRCC between the arousal peaks for $T = T_{max}$ and the corresponding values from the weighted mean GSR profile for the 30 movies used in the experiment. Significant correlations ($p < 0.05$) according to the t-test are indicated by stars.*

Movie	T_{max}	r	SRCC	Movie	T_{max}	r	SRCC
Damaged Kung-Fu	88%	0.364*	0.435*	**Big Buck Bunny**	66%	0.552*	0.643*
Chatter	64%	0.416*	0.537*	**Cloudland**	36%	0.032	0.081
After The Rain	86%	0.236	0.272	**Norm**	24%	0.214	0.314*
On Time	84%	0.626*	0.705*	**Islands**	50%	0.094	0.323
Tears of Steel	42%	0.389*	0.459*	**You Again**	10%	0.231*	0.188*
Barely Legal Stories	70%	0.183	0.291*	**Between Viewings**	84%	0.516*	0.466*
Riding the Rails	84%	0.542*	0.653*	**The Room of Franz Kafka**	34%	0.149	-0.240
Parafundit	88%	0.350	0.361	**Sintel**	88%	0.252*	0.284*
Lesson Learned	90%	0.754*	0.461*	**Nuclear Family**	88%	0.513*	0.357*
Attitude Matters	76%	0.262	0.490*	**Elephants Dream**	18%	0.238*	0.141*
First Bite	84%	0.349	0.333	**Full Service**	74%	0.293*	0.399*
Spaceman	0%	-0.026	0.061	**Superhero**	38%	0.446*	0.472*
The Secret Number	88%	0.359*	0.318	**To Claire From Sonny**	0%	0.183*	0.282*
Origami	22%	0.143	0.008	**Payload**	0%	0.002	0.064
Decay	0%	0.341*	0.441*	**Wanted**	66%	0.629*	0.670*

the scores of the weighted mean GSR profile corresponding to the arousal peaks for $T = 40\%$.

The Pearson's correlation between the continuous mean of the arousal self-assessments and the weighted mean GSR profile $\hat{p_n}$ for a given movie k, $1 \leq k \leq 30$, and for a given threshold T is termed r_T^k. Table 7.1 shows, for each movie, the threshold T_{max} (in %) maximizing r_T^k. The SRCC for each selected T_{max} is also indicated. 18 movies, *i.e.*, 60% of the movies, share a Pearson's r higher than 0.25, but this correlation is significant for 14 of them. In terms of SRCC, the correlation of 18 movies exceeds 0.25 with an associated p-value below 0.05. However, 5 movies in particular show low correlation values that are not significant in terms of both Pearson's r and SRCC. These movies are: Cloudland, Origami, Payload, The Room of Franz Kafka, and Spaceman. Representative screenshots for these movies are shown in Figure 7.3. This lack of correlation could be in part due either to the quality of the movie (*e.g.*, Cloudland is a concatenation of short scenes without any direct link between each other),

or to their movie style (*e.g.*, The Room of Franz Kafka is a short abstract experimental film, Origami is an animation movie with a design inspired by traditional Japanese art, and Payload and Spaceman are science fiction movies). Based on these observations, it seems reasonable to assume that two factors can make uncomfortable the observers: a non-conventional storytelling, such as for Cloudland and an artistic or abstract style, such as for the other ones. Indeed, such media may produce different physiological responses depending on the experimenters' sensitivity.

More globally, the weighted average $\hat{r}_{T_{max}}$ of the Pearson's correlations for all the movies is defined as:

$$\hat{r}_{T_{max}} = \frac{1}{\sum_{k=1}^{30} l^k_{T_{max}}} \left(\sum_{k=1}^{30} l^k_{T_{max}} r^k_{T_{max}} \right) \tag{7.5}$$

with $r^k_{T_{max}}$ the Pearson's correlation between the arousal peaks for $T = T_{max}$ and the corresponding values from the weighted mean GSR profile of the k^{th} movie, and l_k the number of values in the arousal peaks of the k^{th} movie for $T = T_{max}$.

The average weighted Pearson's correlation $\hat{r}_{T_{max}}$ equals 0.264 and the average weighted SRCC, generated in the same way using the SRCC values for each movie, is equal to 0.336. These correlations confirm that arousal and GSR are correlated, but foremost validate the reliability of both the arousal self-assessments and GSR measurements.

7.3 Discussion

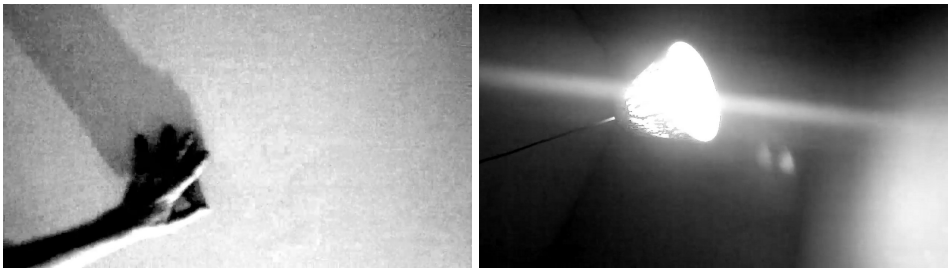
In the previous section, a correlation between the arousal self-assessments and GSR measurements was found. The strength of this conclusion lies in the cross-validation of two modalities temporally collected in two distinct experimental protocols. To the best of our knowledge, it is the first time that a temporal correlation is demonstrated between the average of arousal self-assessments and the average of post-processed GSR measurements from different participants watching videos and collected in two different experiments.

This correlation is the average of the correlations of each movie computed using the threshold T_{max} which is thus different among movies. However, it is possible to analyze the effect of this threshold on the global average weighted correlation by using the same threshold T for all the movies. Figure 7.4 shows the variation of the weighted Pearson's r with respect to the threshold T when T is identical for the 30 movies. Surprisingly, the variation is very smooth. We assumed that the correlation would have increased with the threshold T but in fact this statement is true only for $T < 46\%$. Actually, a global maximum ($r = 0.128$) is achieved for $T = 46\%$, then for $T > 60\%$ the global weighted correlation starts decreasing rapidly. It seems to indicate that the values near the most intense arousal peaks are essential to find a correlation between arousal and GSR measurements and that the smallest arousal values may not be correlated with GSR measurements. This sounds reasonable since for movie scenes inducing low arousal, the constant (or lack of) sweat is not a good indicator of the level of arousal experienced by the participant.

(a) Cloudland

(b) Origami

(c) Payload



(d) The Room of Franz Kafka

(e) Spaceman

Figure 7.3 – *Representative screenshots for the 5 movies for which the weighted mean GSR profile is not correlated with the arousal peaks. Credits and license information can be found in Appendix A.*

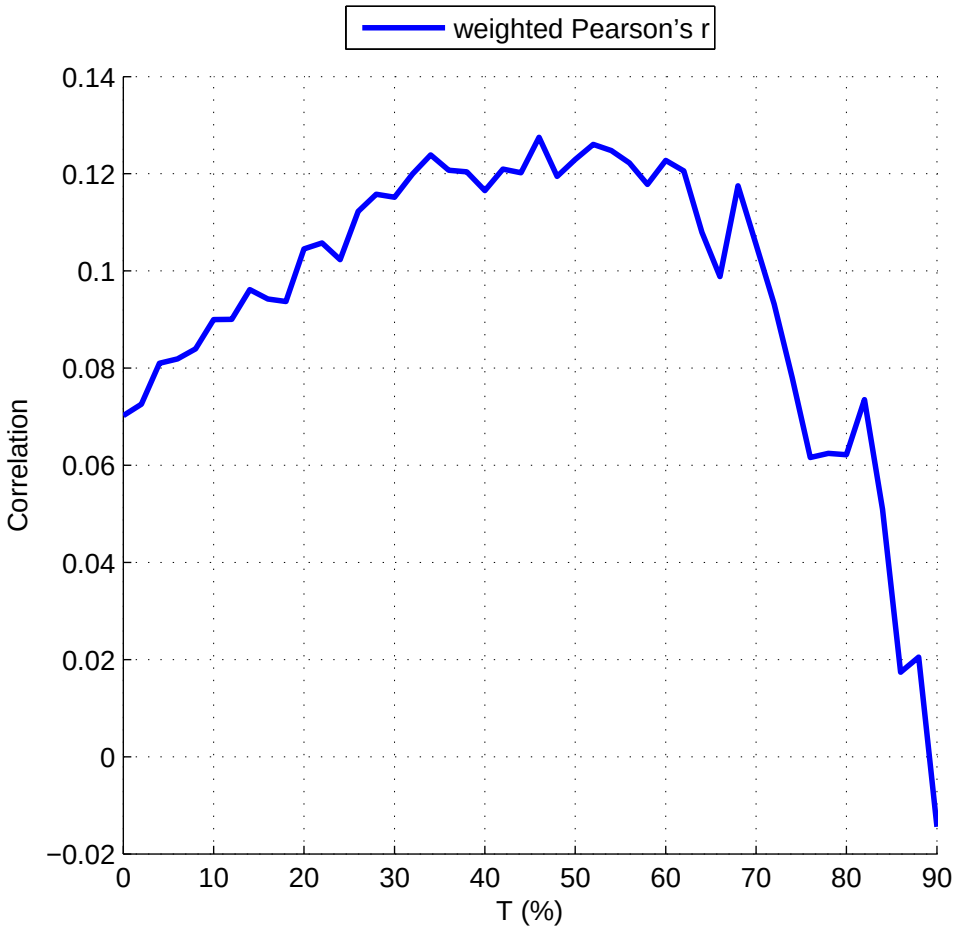


Figure 7.4 – *Evolution of the average weighted Pearson's r with respect to the selected threshold T for the 30 movies*

Summary

We presented in this chapter an experiment to collect physiological measurements, including the GSR, from 13 participants. As for all the other experiments presented in this second part of this thesis, these physiological measurements are publicly available at: http://liris-accede.ec-lyon.fr/. A correlation has been observed between the derivative of the GSR-based signal and the most intense segments of the mean of the arousal self-assessments described in Chapter 6. This confirms that arousal and GSR are correlated, and validates the reliability of both the continuous arousal self-assessments and GSR measurements.

Part III

Estimating Induced Emotions

"I am fine tuning my soul
To the universal wavelength"

— Björk Guõmundsdóttir, *Atom Dance*

Baselines

8

Contents

To enable fair comparison between future work using the 9,800 excerpts of LIRIS-ACCEDE presented in Chapter 4, in which four testing protocols were proposed, we introduce a baseline using a large set of visual and audio features for discrete affective movie content analysis. We also propose four baselines for continuous affective movie content analysis using 30 films continuously annotated presented in Chapter 6. The reproducible protocols allow fair comparisons between state of the art models described below and our different implementations in Chapter 9.

As explained in Chapter 3, existing models use a private dataset making benchmarking and results reproducibility impossible. But there are also neither baselines nor common protocols for assessing the performance of emotion prediction models. This chapter contributes to harmonize the test protocols and lays the foundations for future and fair comparisons. Thus, we aim not at maximizing absolute performance in this chapter, but rather at studying and comparing the performance of five state of the art architectures for the prediction of affective dimensions.

8.1 Baseline for discrete affective movie content analysis

In this section, a baseline is presented to estimate the crowdsourced valence and arousal affective ranks for the 9,800 excerpts of the dataset.

8.1.1 Regression framework

SVM for regression [112], also known as SVR, has demonstrated good performance in many machine learning problems and, more specifically, in affective content analysis work such as [113], [35], or more recently [34]. SVR models construct a hyperplane by mapping vectors from an input space into a high dimensional feature space such that they fall within a specified distance of the hyperplane. Since the formulation of SVM is a convex optimization problem, it guarantees that the optimal solution is found. Two independent ϵ-SVRs [114, 115] are used for this baseline to model arousal and valence separately. The Radial Basis Function (RBF) is selected as the kernel function and a grid search is run to find the C, γ and p parameters. Since the 9,800 excerpts of the LIRIS-ACCEDE database are ranked along the induced arousal and valence axis, the ground truth is made up of these raw ranks, initially ranging from 0 to 9,799, which are uniformly rescaled to a more common $[-1, 1]$ range. All features are normalized using the standard score before being used in the learning step.

8.1.2 Feature selection

A large number of features have been investigated and extracted from three modalities: audio, still image and video features.

Audio features are extracted using 40 ms windows with 20 ms overlap. Many audio features were considered: MFCC, energies, flatness, standard deviation and mean of the quadratic spline wavelet coefficients of the audio signal computed using the fast algorithm described in [116], asymmetry, zero-crossing rate, *etc.* all averaged over the signal. The audio energy contains information about the volume of the signal. The spectrum flatness measures the noisiness character of the spectrum. It is defined as being the ratio of the geometric mean and the arithmetic mean of the spectrum. The zero-crossing rate is the rate of sign-changes along the audio signal. Lastly, the asymmetry is a measurement of the symmetry of the spectrum around its mean value (centroid).

Still image features are extracted from a key frame of the excerpts. This latter is automatically selected as being the frame with the closest RGB histogram to the mean RGB histogram of the whole excerpt using the Manhattan distance. We considered many features, which have proven to be efficient in affective image analysis, as well as more uncommon ones including color harmony and aesthetic features related to the composition of the key frame. Video features contain information about the composition (number of scene cuts, fades, etc.) and motion.

We created two feature sets, one for each axis, made up of the most efficient features. Best features are selected by hierarchically merging the best performing ones as long as the mean-square error (MSE) decreases. Using this process, a set of 17 features is obtained for valence and 12 features for arousal. Rejected features were not necessarily inefficient features but features that were strongly correlated with more efficient ones. The 10 best performing features for estimating arousal and valence are summarized in Table 8.1.

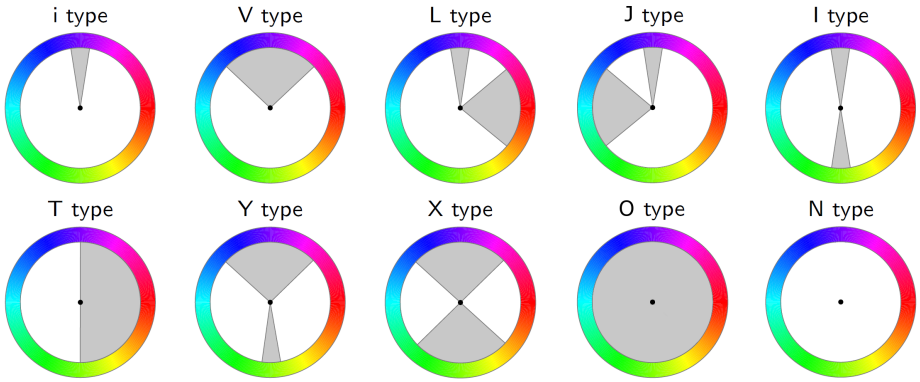


Figure 8.1 – *Harmonious templates on the hue wheel for a given angle (originally published in [4]). The complete collection of harmonious templates is obtained by rotating all templates.*

Table 8.1 – *10 best performing features for estimating arousal and valence dimensions*

Arousal	Valence
1. Global activity	1. Colorfulness [117]
2. Number of scene cuts per frame	2. Hue count [118]
3. Standard deviation of the wavelet coefficients of audio signal	3. Audio zero-crossing rate
4. Median lightness	4. Entropy complexity [119]
5. Slope of the power spectrum	5. Disparity of most salient points
6. Lighting	6. Audio asymmetry envelop
7. Colorfulness	7. Number of scene cuts per frame
8. Harmonization energy [4]	8. Depth of field
9. Length of scene cuts	9. Compositional balance [120]
10. Audio flatness envelop	10. Audio flatness

Color features performed well for detecting valence, as five features out of the 17 features were color-related. For valence, colorfulness [117] was the best performing feature followed by “hue count" [118]. Colorfulness is computed based on the distribution of the key frame pixels in the RGB color space. The other features in this set, from the third best performing feature to the least efficient feature are: audio zero-crossing rate, entropy complexity [119], disparity of most salient points (standard deviation of normalized coordinates), audio asymmetry envelope, number of scene cuts per frame, depth of field (using the blur map computed in [121]), compositional balance [120], audio flatness, orientation of the most harmonious template [4], normalized number of white frames, the color energy and color contrast [57], scene complexity (area of the bounding box that encloses the top 96.04% of edge energy [118]), number of maximum values in the saliency map and, finally, number of fades per frame.

Unsurprisingly related to arousal, motion and energy features were the best performing ones for modeling arousal. The selected features are global activity (average size of motion vectors), standard deviation of the wavelet coefficients of audio signal, the energy corresponding to the most

Table 8.2 – *Performance for Protocols A (Predefined subgroups) and B (Leave-One-Movie-Out) for the discrete emotion prediction baseline. Ground truth and estimated scores range from -1 to 1*

	Protocol A		Protocol B	
Metric	Arousal	Valence	Arousal	Valence
MSE	0.303	0.302	0.326	0.343
Pearson's r	0.308	0.310	0.242	0.221
SRCC	0.302	0.305	0.245	0.219

Table 8.3 – *Performance for Protocol C (Same genre) for the discrete emotion prediction baseline*

Genre	Arousal MSE	Valence MSE
Action	0.278	0.326
Adventure	0.389	0.363
Animation	0.336	0.335
Comedy	0.297	0.295
Documentary	0.326	0.308
Drama	0.313	0.327
Horror	0.331	0.364
Romance	0.324	0.361
Thriller	0.355	0.337

harmonious template [4], the slope of the power spectrum, median lightness, the lighting feature [120], length of scene cuts and the audio flatness envelope. As arousal and valence are correlated, it is not surprising that four features selected among the best performing ones for valence have also been selected for arousal. These features are the number of scene cuts per frame, colorfulness, the normalized number of white frames, and the orientation of the most harmonious template.

There are ten harmonious templates defined in the HSV color space (Figure 8.1) [122, 4]. For a given template at a given angle, grey areas represented in Figure 8.1 enclose hues that form a harmonious set. In other words, an image where the histogram is strictly enclosed in the grey area is considered as harmonious. The most harmonious template is the template minimizing a statistical distance between the histograms of the key frame and the harmonious template [4].

8.1.3 Regression results

The purpose of this section is to run the baseline introduced in this chapter for the standard protocols defined in Chapter 4 (Section 4.4). The MSE, Pearson's r, and Spearman's Rank Correlation Coefficient (SRCC)

Table 8.4 – *Performance for Protocol D (Same movie) for the discrete emotion prediction baseline*

Movie	Arousal MSE	Valence MSE
20 Mississippi	0.305	0.317
Dead Man Drinking	0.309	0.274
Decay	0.330	0.321
Home	0.176	0.401
Lionshare Legacy	0.443	0.273
Monolog	0.290	0.395
Sweet Hills	0.206	0.197
The Master Plan	0.303	0.344
You Again	0.089	0.098

are computed to quantify the performance of each protocol. The MSE for regression models is widely used to quantify the difference between estimated values and the true values estimated. It measures the amount by which the estimated values differ from the ground truth and assesses the quality of the regression in terms of its variation and degree of bias. The Pearson product-moment correlation coefficient (or Pearson's r) is a measure of the linear correlation between estimated and true values, while the SRCC assesses to what extent the relationship between these two variables can be described using a monotonic function.

The MSE, Pearson's r and SRCC for protocol A "Predefined subgroups" and the final averaged results for protocol B "Leave-one-movie-out" using our baseline model are shown in Table 8.2. For protocol C "Same genre", the final averaged MSE for each genre, still using the same sets of features defined in Section 8.1.2, is shown in Table 8.3. The results for protocol D "Same movie", applied to some movies of the database, are displayed in Table 8.4.

The results are promising given the huge variety of movies in the database. They indicate that regression models perform well in modeling of both induced valence and arousal, but with varying degrees of success depending on which protocol is used. Globally, MSE values are significantly smaller than MSE values computed using random sets (around 0.667 and estimated by generating large random samples made up of values between -1 and 1). As pointed out in previous chapters, it is not possible to directly compare the performance of our model to previous state of the art models. They use different test sets and, in most cases, different performance metrics and output scales. On the other hand, researchers using one of the protocols defined in Chapter 4 will be able to know how their model performs not only with respect to this baseline but also to all future work using one of these protocols.

8.2 Baselines for continuous emotion prediction

In this section, four baselines are presented to estimate the continuous valence and arousal scores for 30 movies described in Chapter 6.

In the last few years, breakthroughs in the development of Convolutional Neural Networks (CNN) have led to impressive state of the art improvements in image categorization and object detection. These breakthroughs are a consequence of the convergence of more powerful hardware, larger datasets, but also new network designs, and enhanced algorithms [5, 6]. Is it possible to benefit from these progresses for the affective movie content analysis? In this section, we benchmark four state of the art architectures for the prediction of dimensional affective scores: fine-tuned CNN, CNN learned from scratch, SVR and transfer learning.

8.2.1 Convolutional Neural Networks and Kernel Methods

As mentioned in Section 8.1.1, SVR is one of the most prevalent kernel methods in machine learning. The model learns a non-linear function by mapping the data into a high-dimensional feature space, induced by the selected kernel. As detailed in Chapter 3, SVRs have been extensively used in the affective computing field for music emotion recognition [123], as well as spontaneous emotion recognition in videos [124], and affective video content analysis [113, 34].

Beginning with LeNet-5 [125], CNNs have followed a classic structure. Indeed, they are composed of stacked convolutional layers followed by one or more fully connected layers. So far, best results on the ImageNet classification challenge have been achieved using CNN-based models [5, 6]. CNNs have been mostly used in the affective computing field for facial expression recognition [126]. Recently, Kahou *et al.* trained a CNN to recognize facial expressions in video frames [54]. Its prediction was then combined with the predictions from three other modality-specific models to finally predict the acted-out emotional category in short video clips.

The CNN approach disrupts the field of machine learning and has significantly raised the interest of the research community for deep learning frameworks. Generally applied for object recognition, its use will be naturally extended to any recognition task. The contributions using CNNs in the affective computing field will likely show up in the coming months.

8.2.2 Regression Frameworks for Emotion Prediction

In this section, we describe the four frameworks that are compared in Section 8.2.3. All the models presented in this section output a single value: the predicted valence or arousal score. Thus, they all need to be learned twice: either for predicting induced arousal scores, or for predicting induced valence scores.

Deep Learning

Two models using CNNs to directly output affective scores are investigated in this work. Both take as input the key frame of the video segment for which an arousal or valence score is predicted. The key frame

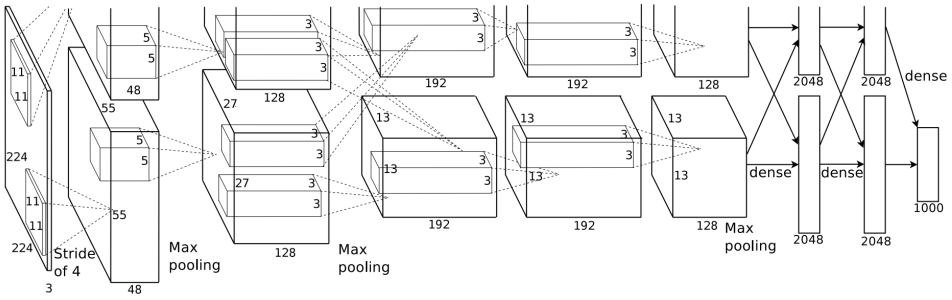


Figure 8.2 – *Illustration of the architecture of the CNN introduced by Krizhevsky et al. (originally published in [5])*

is defined as the frame with the closest RGB histogram to the mean RGB histogram of the whole excerpt using the Manhattan distance.

We used data augmentation to enlarge artificially the training set. As in [5], the model was trained using random 224×224 patches (and their horizontal reflections) extracted from the 256×256 input images. These input images were the center crop of the key frames extracted from the video segments in the training set and resized so that the original aspect ratio is preserved but their smallest dimension equals 256 pixels. The training is stopped when the Mean Square Error (MSE), measured every 500 iterations using a validation set, increases for five consecutive measurements. At validation and test time, the network makes a prediction by extracting the 224×224 center patch.

We present two approaches: one exploiting previously learned models, and the second one that attempts to learn completely a new model randomly initialized.

Fine-tuning: This first framework is based on the fine-tuning strategy. The concept of fine-tuning is to use a model pretrained on a large dataset, replace its last layers by new layers dedicated to the new task, and fine-tune the weights of the pretrained network by continuing the back-propagation. The main motivation is that the most generic features of a CNN are contained in the earlier layers and should be useful for solving many different tasks. However, later layers of a CNN become more and more specific to the task for which the CNN has been originally trained.

In this work, we fine-tune the model proposed in [5] composed of five stacked convolutional layers (some are followed by local response normalization and max-pooling), followed by three fully connected layers. This model is illustrated in Figure 8.2. To adapt this model to our task, the last layer is replaced by a fully connected layer composed of a unique neuron scaled by a sigmoid to produce the prediction score. The loss function associated to the output of the model is the Euclidean loss. Thus, the model minimizes the sum of squares of differences between the ground truth and the predicted score across training examples. All the layers of the pretrained model are fine-tuned, but the learning rate associated to the original layers is ten times smaller than the one associated with the new last neuron. Indeed, we want the pretrained layers to change very slowly, but let learn faster the new layer which is initialized from a

zero-mean Gaussian distribution with standard deviation 0.01. This is because the pretrained weights should be already relatively meaningful, and thus should not be distorted too much. We also tried to fine-tune the last three layers only, *i.e.*, the learning rate associated to the first layers is set to zero, but the prediction performance was much worse.

We trained the new fine-tuned models using the reference implementation provided by Caffe [127] using stochastic gradient descent with a batch size of 256 examples, momentum of 0.9, base learning rate of 0.0001, and weight decay of 0.0005.

Learning From Scratch: We also built and learned from scratch a CNN based on the architecture of [5] but much simpler since our training set is composed of 16,065 examples.

The model is made of two convolutional layers and three fully connected layers. As in [5], the first convolutional layer filters the $224 \times 224 \times 3$ input key frame with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolutional layer, connected to the first one, uses 256 kernels of size $5 \times 5 \times 96$. The outputs of both convolutional layers are response-normalized and pooled. The first two fully connected layers are each composed of 512 neurons and the last fully connected layer is the same as the last one added to the fine-tuned model in the previous section. The ReLU nonlinearity is applied to the output of all the layers. All the weights are initialized from a zero-mean Gaussian distribution with standard deviation 0.01. The learning parameters are also the same as those used for the fine-tuning strategy.

SVR

This model is similar to the baseline framework presented in Section 8.1: two independent ϵ-SVRs are learned to predict arousal and valence scores separately. The RBF is selected as the kernel function and a grid search is run to find the C, γ and p parameters. The SVR is fed using the early fusion scheme with the features detailed in Section 8.1.2, *i.e.*, audio, color, aesthetic, and video features. All features are normalized using the standard score.

Transfer Learning: CNN as a feature extractor

The approach is the same as in the previous section except that the 4,096 activations of the second fully connected layer called "FC7" of the original model learned in [5] are normalized using the standard score and used as features to feed the SVR using the early fusion scheme, in addition to the features detailed in Section 8.1.2. Thus, the CNN is treated as a feature extractor and is used to, hopefully, improve the performance of the SVR.

8.2.3 Performance Analysis

In this section, the performance of the four well-known state of the art architectures introduced in Section 8.2.2 is compared and discussed using the continuous annotations introduced in Chapter 6.

The Importance of Correlation

The common measure generally used to evaluate regression models is the Mean Square Error (MSE). However, the performance of the models cannot be analyzed using simply this measure. As a point of comparison, on the test set, the MSE between the ground truth (ranging from 0 to 1) for valence and random values generated between 0 and 1 equals 0.113, whereas the linear correlation (Pearson's r correlation coefficient) is close to zero. However, the ground truth is biased to the extent that a large portion of the data is neutral (*i.e.* its valence score is close to 0.5) or is distributed around the neutral score. This bias can be seen from Figure 6.2. Thus, if we create a uniform model that always outputs 0.5, its performance will be much better: its MSE is 0.029. However, the correlation between the predicted values and the ground truth will be also close to zero. The performance for the random and uniform baselines is indicated in Table 8.5. For the random distribution, we generate 100 distributions and report the average MSE and correlation.

To analyze the results and the performance of the computational models, the linear correlation has the advantages not to be affected by the range of the scores to be predicted and to measure the relationship between the predicted values and the ground truth.

Experimental Results

To learn and evaluate the various frameworks, the dataset presented in Chapter 6 and composed of 26,525 1-second segments extracted from 30 movies is distributed into a training set, a validation set and a test set. Approximately 60% of the data is assigned to the training set and 20% of the data is assigned to both the validation and test sets. More precisely, 16,065 1-second segments extracted from 15 movies are assigned to the training set, 5,310 segments from 8 movies to the validation set and finally, 5,150 segments from 7 movies to the test set. This distribution makes also sure that the genre of the movies in each set is as diverse as possible.

Table 8.5 presents the results of using CNNs (fine-tuned and learned from scratch), SVR and transfer learning for the prediction of valence and arousal dimensions based on the MSE and the Pearson's r correlation coefficient. For the four frameworks, the predicted scores as well as the ground truth for valence and arousal range from 0 to 1. Table 8.5 shows that for valence and arousal, the highest correlation is obtained by the transfer learning approach. Once again, this result reveals that CNNs provide generic mid-level image representations that can be transferred to new tasks, including the transfer from the classification of 1,000 ImageNet classes to the prediction of the valence and arousal affective scores. Transfer learning improves by 50% the performance in terms of correlation of the second best performing framework for predicting valence, and by 17%

Table 8.5 – *Prediction results for the continuous emotion prediction baselines for valence and arousal dimensions (MSE: Mean Square Error, r: Pearson correlation coefficient)*

System	Arousal		Valence	
	MSE	r	MSE	r
Random	0.109	0.0004	0.113	-0.002
Uniform	0.026	-0.016	0.029	-0.005
CNN – Fine-tuned	**0.021**	0.152	**0.027**	0.197
CNN – From scratch	0.023	0.157	0.031	0.162
SVR – Standard	0.023	0.287	0.035	0.125
SVR – Transfer learning	0.022	**0.337**	0.034	**0.296**

for arousal. However, no clear gain is obtained for MSE. For valence, the MSE is even higher than the MSE of the uniform strategy.

The fine-tuned CNN outperforms the other models in terms of MSE for both valence and arousal. The gain in terms of MSE is more important for valence. For arousal, the MSE value is close to the performance obtained by the transfer learning strategy. However, for both arousal and valence, the correlation is much lower than the performance obtained with transfer learning. Deeper analyses have shown that deep learning predictions are noisy. Post-processing the predictions using a simple temporal Gaussian smoothing greatly improves the performance of the fine-tuned CNN for both arousal and valence, outperforming the transfer learning strategy. These results are presented in the next chapter. Nevertheless, it is a promising result given that the performance of this model on the training set indicates that, despite the use of a validation set to stop learning when the performance on the validation set increased for 5 consecutive measurements, the size of the dataset is not big enough to prevent overfitting. Indeed, previous work has shown that overfitting and training set size are closely related [128]. For example, the performance of the fine-tuned model on the training set for the prediction of valence is much better ($MSE = 0.012$, $r = 0.79$). It may also explain why the performance of the CNN learned from scratch is lower than the performance of the fine-tuned CNN.

Regarding the arousal dimension, it is interesting to note that the correlation of the SVR is almost twice the correlation of the pure deep learning frameworks. This could be explained by the fact that both deep learning models lack audio and motion information, unlike the SVR framework which uses as input a combination of features extracted from both the audio signal and from statistics for consecutive frames of a video segment. However, Nicolaou *et al.*, among others, showed that the prediction of arousal is greatly enhanced by the use of audio and motion cues [53]. This is why we investigate in the next chapter the use of audio cues in CNN frameworks to produce more accurate affective predictions for videos and to take into account more than one frame to predict the induced affective score of a 1-second length video segment.

Summary

In this chapter, we introduced standard protocols using the database in an attempt to perform standardized and reproducible evaluations to fairly compare future work within the field of affective computing.

We implemented a baseline for discrete affective movie content analysis and assessed its performance using four protocols, corresponding to different goals and needs, and showing promising results. Note that all the audio and visual features used for the baseline are also released alongside the LIRIS-ACCEDE database.

We also introduced four baselines for continuous emotion prediction. We found that the fine-tuned CNN framework is a promising solution for emotion prediction. However, the limited size of the training set (16,065 samples) prevents the pure CNN-based frameworks to obtain good performances in terms of correlation. Nevertheless, intermediate layers, originally trained to perform image recognition tasks, are generic enough to provide mid-level image representations that can greatly improve the prediction of affective scores in videos.

However, these four baselines do not take into account the temporal information for continuous emotion prediction. Based on these promising results, we introduce in the next chapter a more psychologically relevant CNN-based spatio-temporal framework, modeling the effect that previous scenes may reasonably influence the emotion inference of future ones.

The Spatio-Temporal Model

9

Contents

Based on the promising results of the baselines for continuous emotion prediction introduced in the previous chapter, we introduce in this chapter a more psychologically relevant CNN-based spatio-temporal framework. First, a static model is proposed using CNNs to predict an affective score for a given short video segment based on several key frames and audio spectrograms. A multi-level data-augmentation is also introduced to improve the prediction accuracy.

As mentioned in Chapter 2, psychologists suggest that the evaluation of an emotion is an iterative process. For example, Russell states to define the Core Affect that [20]:

> "Emotional life consists of the continuous fluctuations in core affect, in pervasive perception of affective qualities, and in the frequent attribution of core affect to a single Object, all interacting with perceptual, cognitive, and behavior processes."

This process of recursive and continuous evaluations is also the core of the appraisal evaluation postulated by Scherer [1]:

> "In the case of humans, the CPM postulates that the recursive checking process repeats the sequence continuously, constantly updating the appraisal results that change rapidly with changing events and evolving [emotional] evaluation until the monitoring subsystem signals termination of or adjustment to the stimulation that originally elicited the appraisal episode."

Thus, this key aspect of emotions has to be considered in affective computational models. In our innovative approach, we propose to introduce psychological insights for the computational modeling of emotions. This is characterized by addressing the temporal dimension of the video inputs. In other words, previous induced emotions have to be taken into account and introduced recursively into the model. This is why in this work mid-level image representations are extracted from the static model to feed a bidirectional Long Short-Term Memory Recurrent Neural Network which is able to model the emotional relationships between consecutive video segments. This spatio-temporal transfer learning approach outperforms all the baselines for the prediction of the induced valence and arousal scores.

9.1 Advanced static model

Based on the promising results of the fine-tuning strategy presented in Chapter 8, we aim in this section to maximize the performance of the static model, *i.e.* the model estimating induced arousal or valence scores for 1-second movie segments.

9.1.1 Multi-level data-augmentation

Karpathy *et al.* developed several CNN architectures for video classification [129]. They found that the single-frame multi-resolution framework was the best performing architecture in terms of classification accuracy per clip. Based on this result, to improve the performance of the fine-tuned deep learning model, we also adopt in this chapter a single-frame CNN-based architecture.

However, the complex architecture of CNNs is particularly prone to overfitting due to its large number of parameters relative to the number of observations. Augmenting the size of input data is thus an appropriate technique to reduce overfitting. However, it is often time consuming to reliably collect new ground truth from annotators. This is why artificially augmenting the size of the datasets is often the preferred solution to prevent overfitting. To combat overfitting by artificially enhancing the size of the training set, a multi-level data-augmentation (MLDA) is described and variability in the training ground truth to take into account the variations of the self-assessments is introduced.

Training methodology

For the fine-tuned baseline described in Chapter 8, we will refer to as "Fine-tuned AlexNet", a single key frame was extracted and resized from each 1-second movie segment. Then, random 224×224 crops (and their

horizontal reflections) extracted from the resized key frame were used during training.

We adopt here a more aggressive MLDA strategy illustrated in Figure 9.1 inspired by the strategy used by Szegedy *et al.* who extracted crops for different resolutions [6]. First, 5 key frames are extracted for each 1-second video segment. A frame is defined as a key frame if its YUV histogram is the closest, using the Manhattan distance, to a cluster computed by the k-mean clustering algorithm (with $k = 5$). For each key frame, a first patch is extracted corresponding to the 256×256 patch maximizing the salience of the key frame. The saliency map is computed for the considered picture from the visual attention model of [130]. The saliency map provides a representation of the most visually attractive pixels. Then, the key frames are resized so that the original aspect ratio is preserved but their smallest dimension equals 1.1×256 pixels. Five 256×256 patches are extracted from the resized key frames corresponding to the top-left, top-right, center, bottom-left, and bottom-rights squares. The model is finally trained using random 224×224 crops (and their horizontal reflections) extracted from the 256×256 input patches. In this multi-level data-augmentation, $5 \times 6 = 30$ patches are used for training instead of a single one for the fine-tuned AlexNet baseline presented in the previous chapter.

We also introduce variability for the ground truth associated to the 30 patches of a segment to combat overfitting. Indeed, the dropout method has demonstrated that adding noise to the states of hidden units in a neural network helps preventing overfitting [128]. This is why in this work we investigate the addition of noise to the ground truths, in relation to the variability of the affective self-assessments used to generate the ground truth. Ground truths for the training set are no longer associated to the arousal or valence score corresponding to the 1-second video segments. Indeed, the ground truth associated to a patch is considered as a random variable generated using a Normal distribution. The mean of the distribution is the affective score (either valence or arousal) corresponding to the 1-second segment and its standard deviation equals a quarter of the standard deviation corresponding to the 1-second segment and generated from the continuous self-assessments (either valence or arousal) of the 10 participants (Figure 6.2).

Test strategy

At test time, the center 224×224 crop for the 30 patches is extracted. The final prediction is obtained by averaging the 30 scores generated by the model for each patch. A simple averaging approach was preferred over alternative approaches since Szegedy *et al.* [6] found that, using their cropping approach, the best performance was obtained with simple averaging.

Results

To analyze the performance of the test strategy, we also computed the final predictions by averaging only the predicted scores for the 5 center patches extracted from the resized key frames. We also analyzed the ef-

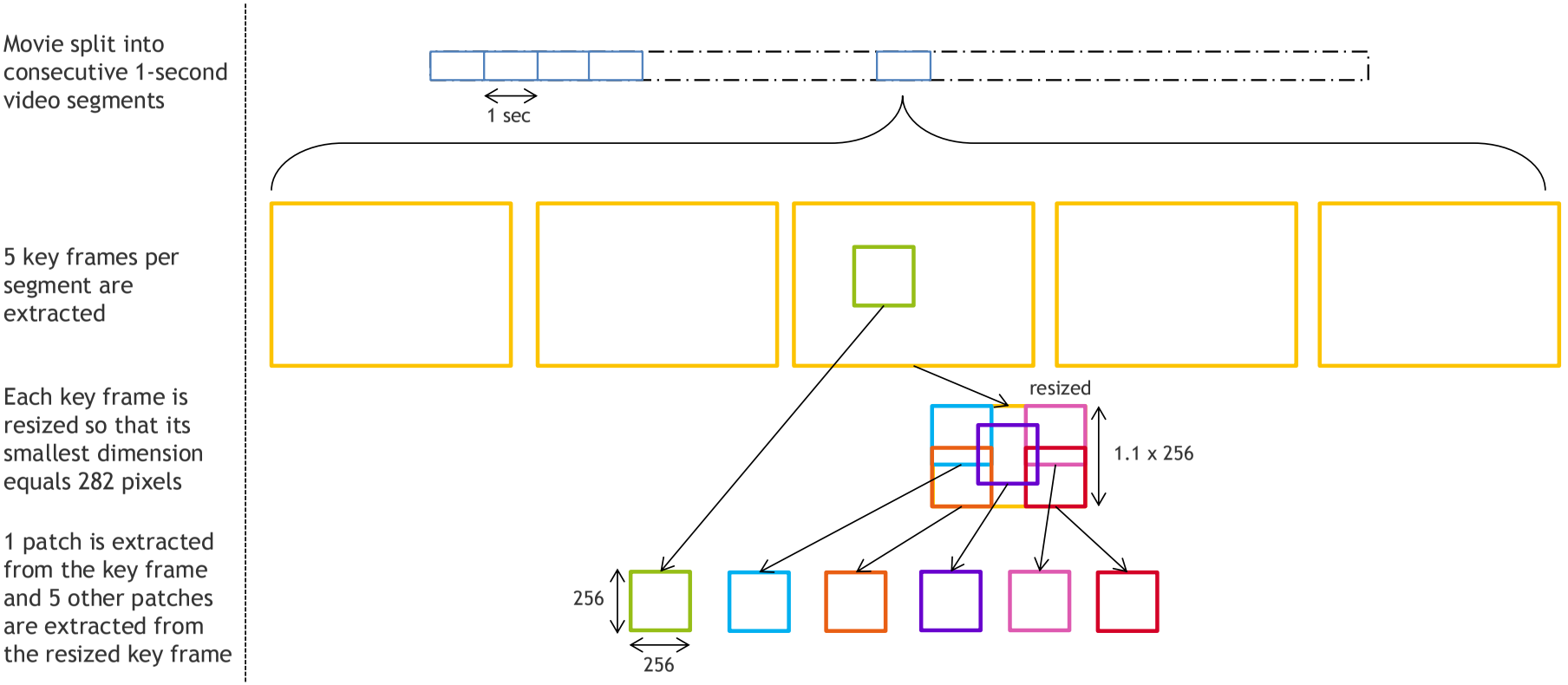


Figure 9.1 – *Multi-level data-augmentation for the learning phase of the advanced static model*

Table 9.1 – *Performance for the fine-tuned AlexNet with three multi-level data-augmentation test strategies*

System and test strategy	Arousal		Valence	
	MSE	r	MSE	r
Fine-tuned AlexNet (1 crop)	**0.021**	**0.152**	**0.027**	0.197
Fine-tuned AlexNet with MLDA (30 patches averaging)	**0.021**	0.143	**0.028**	**0.283**
Fine-tuned AlexNet with MLDA (5 center patches averaging)	0.022	0.111	0.030	0.222
Fine-tuned AlexNet with MLDA (all but salience-based patches averaging)	0.021	0.134	0.029	0.253

fect of the saliency-based patches. The results are indicated for the fine-tuned AlexNet in Table 9.1. For valence, the MLDA and 30 patches averaging strategy greatly improves the performance in terms of correlation (+43%) compared to the fine-tuned AlexNet baseline. However, the MSE is slightly higher (+4%). The salience-based patch, which is a contribution of this thesis, greatly improves the performance of the MLDA: it provides a full-resolution zoom on the most meaningful part of the key frames. For arousal, the results are more mitigated. They show that a single crop extracted from a well-chosen key frame is good enough to predict arousal scores. In fact, previous work has shown that arousal is more related to audio or temporal features than visual features [50, 59, 52]. In Section 9.1.3, we show that audio features moderately improve the prediction performance for arousal. Indeed, it is the use of the temporal information in Section 9.2 which will greatly enhance the prediction accuracy for arousal.

9.1.2 Fine-tuning GoogleNet

To surpass the performance of the fine-tuned AlexNet baseline, we fine-tune in this section the GoogleNet model introduced by Szegedy *et al.* that became in 2014 the new state of the art performance on the ImageNet dataset. The GoogleNet architecture is illustrated in Figure 9.2. GoogleNet is much deeper than AlexNet, but uses 12× fewer parameters even if the computational cost is increased by a factor of two. It uses a concatenation of nine similar "Inception" networks. An Inception network consists in 1×1, 3×3, and 5×5 convolutions stacked upon each other, with max-pooling layers to reduce the resolution. Another key aspect of GoogleNet is that all the convolutions, including those inside the Inception modules, use rectified linear activation. Given the depth of the network, two auxiliary losses are connected to intermediate layers to increase the back-propagated gradient and to prevent the vanishing gradient problem. During training, the auxiliary losses are added to the total loss of the network with a discount weight. At test time, the auxiliary losses are no longer needed and are thus removed from the network. In our fine-tuning

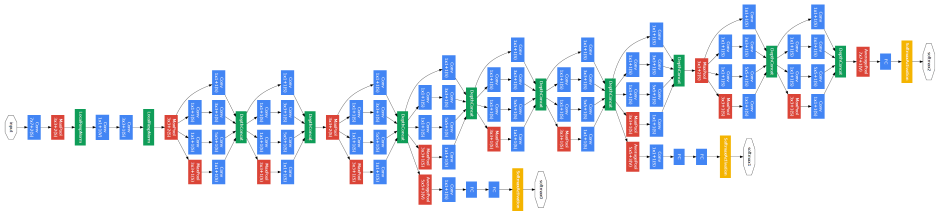


Figure 9.2 – *Illustration of the architecture of GoogleNet introduced by Szegedy et al. (originally published in [6])*

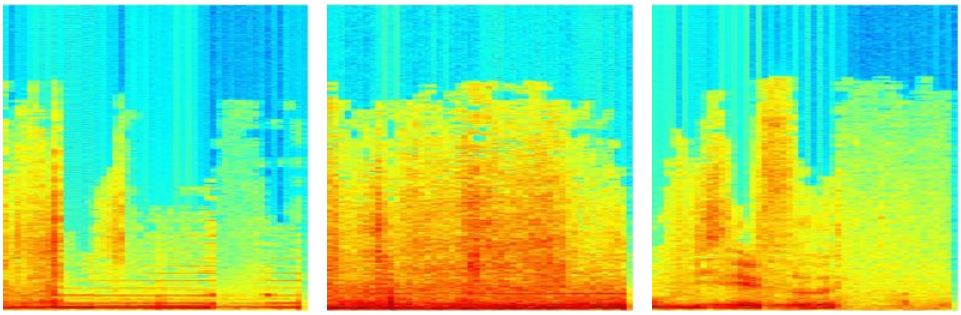


Figure 9.3 – *Examples of resized spectrograms used as input by the audio-based CNNs*

approach, the two auxiliary and the final softmax activations are replaced by a fully connected layer composed of a unique neuron scaled by a sigmoid to produce the prediction score. As for the fine-tuned AlexNet, the loss functions associated to the model are the Euclidean loss.

The performance for the fine-tuned GoogleNet with MDLA is surprisingly lower than the performance of the fine-tuned AlexNet with MDLA (Table 9.2). The GoogleNet model is more computationally expensive and requires more memory. Thus, the batch size of the GoogleNet is almost six times smaller than the batch size used for fine-tuning AlexNet (44 instead of 256), in order to fit in the GeForce GTX980 with 4GB of dedicated graphics memory. It may explain why the efficiency of the fine-tuned GoogleNet is not as good as expected.

However, the last fully connected layer provides a more compact mid-level representation (1,024 neurons) than the last fully connected layer from the fine-tuned AlexNet (4,096 neurons). This compactness may be a significant advantage to transfer mid-level representations to the temporal model in Section 9.2.

9.1.3 Audio modality

The fine-tuned AlexNet and GoogleNet presented in the previous sections only use as input the frames of the 1-second video segments. However, the audio channel is also important to predict the affective impact of movies, and in particular to predict arousal scores [59, 52].

Spectrograms

Audio spectrograms are a visual representation of the spectrum of frequencies in a sound. These visual representations have been naturally used in previous work to successfully learn CNNs for speech emotion recognition, or musical onset detection [131, 132].

Table 9.2 – *Performance for the unimodal static architectures, i.e., for the visual fine-tuned AlexNet and GoogleNet models, and for the audio-based CNN*

Model	Arousal		Valence	
	MSE	r	MSE	r
Fine-tuned AlexNet with MLDA (30 patches averaging)	0.021	0.143	0.028	**0.283**
Fine-tuned GoogleNet with MLDA (30 patches averaging)	0.022	0.136	**0.027**	0.274
Fine-tuned audio-based CNN	**0.020**	**0.145**	0.029	0.213

The left and right channels of the time-domain audio signal, extracted from a 1-second length video segment, are first converted into spectrograms using short-time Fourier transform [133]. The spectrogram has a 40 ms window size with a 20 ms overlap, windowed with a Hamming window. Both spectrograms are then summed and the resulting image is resized so that its size equals 256×256 pixels. Examples of three resized spectrograms used as inputs by the audio-based CNN presented in the next section are shown in Figure 9.3.

Audio-based CNN

Similarly to the baselines for continuous emotion prediction, several architectures are tested among: learning a lighter CNN from scratch, fine-tuning AlexNet or fine-tuning GoogleNet with spectrograms. The training strategy is also the same as the one introduced in Chapter 8, *e.g.*, no data augmentation similar to the MLDA is performed on the audio signal. Instead, random 224×224 patches (and their horizontal reflections) are extracted from the 256×256 input spectrograms to feed the audio-based CNN. Surprisingly, the best performance has been achieved by fine-tuning AlexNet for both valence and arousal. We did not expect such a result since AlexNet has been originally trained using standard pictures from the ImageNet dataset to detect objects. It seems that mid-level representations learned by the original AlexNet are general enough to embody spectrograms to predict induced emotions.

The performance of the audio-based CNN is detailed in Table 9.2. Compared to the performance of the visual-based CNNs, the audio-based CNN achieves slightly better results for the prediction of the induced arousal. For valence, the visual-based CNNs still achieve the best performance in terms of both MSE and correlation. Again, it is consistent with previous work showing that arousal is more related to audio features, while valence is more related to visual features [60, 134].

9.1.4 Multimodal results

The multimodal static framework is the combination between predictions from a visual-based CNN (*i.e.*, the fine-tuned AlexNet or GoogleNet) and from the audio-based CNN (*i.e.*, the fine-tuned AlexNet).

Table 9.3 – *Performance for the multimodal static fusions*

Fusion	Models	Arousal		Valence	
		MSE	r	MSE	r
(a)	**Fine-tuned AlexNet and audio-based CNN**	0.040	0.124	0.027	**0.349**
(b)	**Fine-tuned AlexNet and audio-based CNN**	**0.018**	**0.170**	0.028	0.291
(a)	**Fine-tuned GoogleNet and audio-based CNN**	0.040	0.129	**0.026**	**0.345**
(b)	**Fine-tuned GoogleNet and audio-based CNN**	**0.018**	0.161	0.027	0.281

Two techniques are evaluated to combine the 30 predictions from the CNN based on the visual modality and those from the CNN based on the audio modality. Either (a) the final prediction is the mean between the average of the visual predictions from the 30 patches extracted from the key frames and the prediction from the audio-based CNN, or (b) the final prediction is the mean between the visual predictions from the 30 visual patches with the prediction from the audio-based CNN. Results for both fusions are detailed in Table 9.3. Fusion (a) is more efficient to combine valence predictions, while fusion (b) gives better results for arousal. Furthermore, the performance using the fine-tuned AlexNet instead of the fine-tuned GoogleNet is not significantly higher. In fact, next section shows that the fine-tuned GoogleNet model is more appropriate to be used in a transfer learning approach.

9.2 Advanced temporal model

The multimodal static framework presented in the previous section misses crucial information to model the emotions induced by videos: the temporal information. Based on the static framework, we introduce in the following sections a temporal model to predict induced emotions for consecutive 1-second video segments.

9.2.1 LSTM-RNNs

Recurrent Neural Networks (RNNs) are powerful networks that are able to model input sequences of different lengths thanks to the idea of parameters that are shared over different parts of the network. It can be trained using the Back-Propagation Through Time (BPTT) algorithm, a generalization of the back-propagation algorithm. The main problem with the BPTT algorithm is that the gradients propagated over many stages tend most of the time to vanish exponentially with the number of time steps. Thus, the long-term dependencies tend to be hidden by the smallest fluctuations arising from the short-term dependencies [135]. Learning efficiently long-term dependencies remains an unsolved problem.

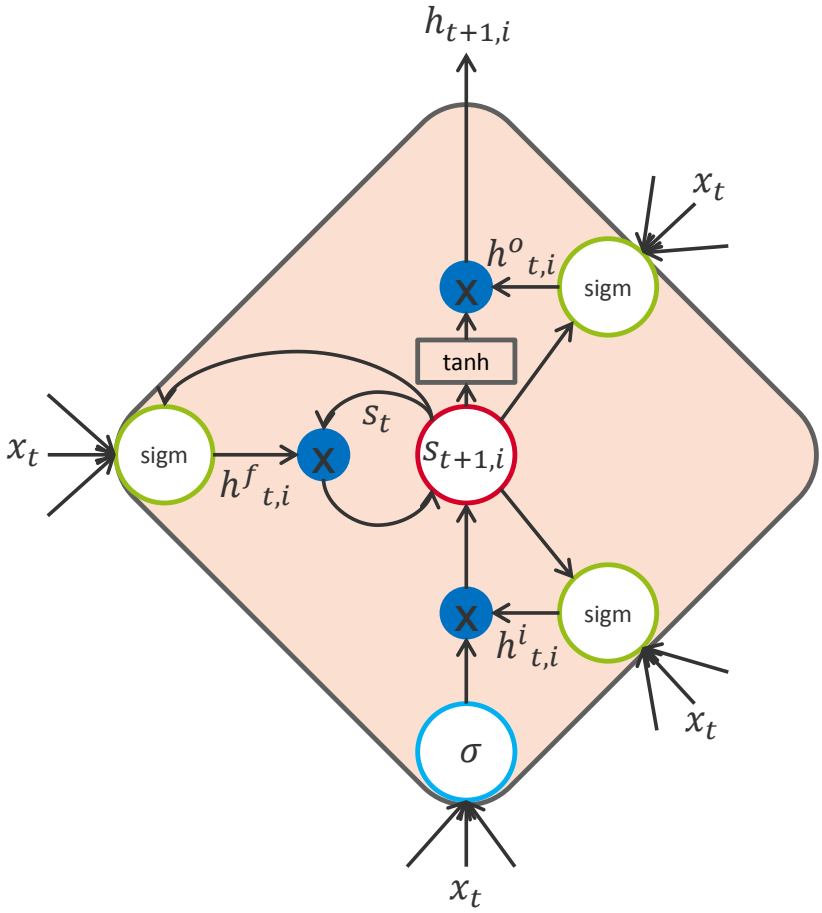


Figure 9.4 – *Schema of a LSTM unit. The red circle represents the state cell and the three green circles represent the input, forget and output gates.*

The Long-Short-Term-Memory (LSTM) units have been created in order to better capture long-term dependencies [136]. They replace the usual hidden units of neural networks and have the ability to learn when to remember and when to forget past dependencies. LSTM-RNNs have been used in several works dealing with affective computing. They have been used to predict continuous scores of spontaneous affect for multiple cues [53, 137], to predict asynchronous valence and arousal scores based on audiovisual and physiological features [134], to continuously estimate the emotions felt by participants watching videos using EEG signals and facial expressions [138], and for the task of on-line continuous-time music mood regression [139]. That is why we believe in the capacity of LSTM-RNNs to model the temporal aspect of induced emotions.

The LSTM unit used in this thesis follows the one introduced by Graves [140]. The LSTM unit is illustrated in Figure 9.4. The memory of the LSTM unit i for time step t is managed through three sigmoidal gates: the input gate $h^i_{t,i}$, the forget gate $h^f_{t,i}$, and the output gate $h^o_{t,i}$. The gates are activated by passing their input through the sigmoid function. The most important component of the LSTM unit is the state cell s_t that has a linear self-loop weighted by the forget gate. The output of the state cell is passed through

a tanh non-linearity and can be shut off with the output gate. These values are computed as follows:

$$h^f_{t,i} = sigmoid(b^f_i + \sum_j U^f_{ij} x_{t,j} + \sum_j V^f_{ij} s_{t,j} + \sum_j W^f_{ij} h_{t,j}) \tag{9.1}$$

$$s_{t+1,i} = h^f_{t,i} s_{t,i} + h^i_{t,i} \sigma(b_i + \sum_j U_{ij} x_{t,j} + \sum_j W_{ij} h_{t,j}) \tag{9.2}$$

$$h^i_{t,i} = sigmoid(b^i_i + \sum_j U^i_{ij} x_{t,j} + \sum_j V^i_{ij} s_{t,j} + \sum_j W^i_{ij} h_{t,j}) \tag{9.3}$$

$$h^o_{t,i} = sigmoid(b^o_i + \sum_j U^o_{ij} x_{t,j} + \sum_j V^o_{ij} s_{t,j} + \sum_j W^o_{ij} h_{t,j}) \tag{9.4}$$

$$h_{t+1,i} = tanh(s_{t+1,i}) h^o_{t,i} \tag{9.5}$$

where x_t is the current input vector, h_t is the current hidden layer vector, σ is the neural non-linearity (*e.g.*, sigmoid or tanh), b^f, b^i, b^o are the biases for the gates, U^f, U^i, U^o are the input weights for the gates, and V^f, V^i, V^o, W^f, W^i, W^o are the recurrent weights for the three gates.

9.2.2 Architecture

Considering the great performance of the transfer learning approach in Chapter 8, we have chosen to use this technique to build the temporal computational model. The temporal model is thus built using mid-level representations extracted from an audio-based CNN and from a visual-based CNN that are used as input by a bidirectional LSTM-RNN [141].

A bidirectional LSTM-RNN is the combination of the concepts of LSTM-RNNs detailed in the previous section, and of bidirectional RNNs [142]. A bidirectional RNN has access to all past and all future inputs through the use of two distinct input layers processing the data forward or backward. The outputs from both hidden layers are connected to the same output layer. Bidirectional LSTM-RNNs can thus access long-range dependencies in both input directions.

The general architecture of the advanced spatio-temporal model is shown in Figure 9.5. The steps to create an affective curve for a movie are:

1. The movie is first segmented into consecutive 1-second length video segments.
2. An audio spectrogram and a patch extracted from a key frame are computed from the first 1-second length segment of the movie. The selection of the k^{th} patch to be used among the 30 available visual patches allows to artificially augment the training data in Section 9.2.4.
3. Then, the spectrogram is used to feed the audio-based CNN trained in Section 9.1.3 for arousal or valence depending on the affective curve to be generated.
4. In parallel, the k^{th} visual patch feeds the visual-based CNN trained in Section 9.1.2 based on GoogleNet. This CNN framework has been

Table 9.4 – *Performance for the temporal model fed with the visual fine-tuned AlexNet or the fine-tuned GoogleNet*

Model	Arousal		Valence	
	MSE	r	MSE	r
LSTM-RNN using visual fine-tuned AlexNet activations	0.021	0.097	0.029	0.349
LSTM-RNN using visual fine-tuned GoogleNet activations	**0.021**	**0.135**	**0.025**	**0.458**

preferred over the fine-tuned AlexNet described in Section 9.1.1. Indeed, using the fine-tuned GoogleNet instead of Alexnet increases the performance, in terms of correlation, by approximately 31% for valence, and 39% for arousal (Table 9.4).

5. Mid-level representations are extracted from the activations of the neurons of both unimodal CNNs. Their dimension is reduced using Principal Component Analysis (PCA). Both modalities are then combined to form a single multimodal compact mid-level representation of the short video segment (see Section 9.2.3).
6. The multimodal compact mid-level representation feeds the bidirectional LSTM-RNN trained in Section 9.2.4 to output the affective score.
7. Steps 2 to 5 are repeated for the next 1-second length video segment until the last segment of the movie is reached.
8. Finally, the resulting affective curve, representing the induced valence or arousal, is smoothed using a Gaussian filter.

9.2.3 Combination of visual and audio modalities

For the spatio-temporal model, a feature-level fusion is adopted instead of a decision-level fusion, *i.e.*, the mid-level representations for the audio and visual modalities are concatenated before being used by the bidirectional LSTM-RNN. This decision has been made experimentally since preliminary results have shown that a feature-level fusion was more efficient. It also makes sense that both audio and visual modalities should be considered all at once to predict the emotions induced by videos.

However, the dimensions of both mid-level representations, *i.e.*, the activations of the last fully connected layers from the CNNs, are non-negligible. For the audio modality, the last fully connected layer of the CNN based on AlexNet is composed of 4,096 neurons, while for the visual modality, the last fully connected layer of the CNN based on GoogleNet is composed of 1,024 neurons. If no data reduction is performed, the dimension of the input (5,120 values) is too large compared to the size of the limited training set composed of 15 temporal sequences only, *i.e.*, 15 movies, which causes overfitting and thus poor performances for the test set. Three techniques for dimensionality reduction are mainly used: PCA, Linear Discriminant Analysis, and Independent Component Analysis. No

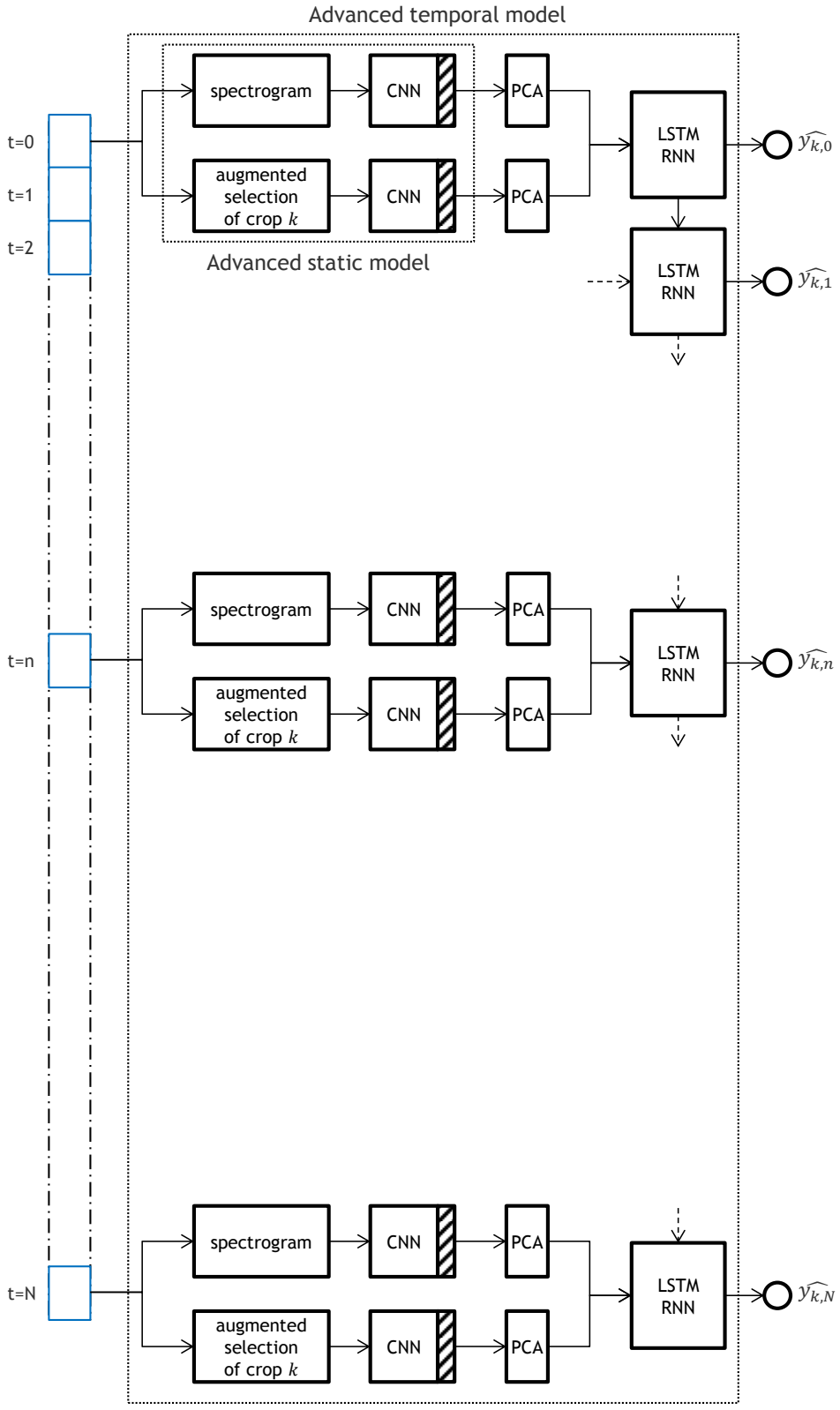


Figure 9.5 – *General architecture to compute an affective curve for a movie using the advanced spatio-temporal model*

claim can be made about which is the most efficient reduction technique [143, 144]. However, PCA seems more robust compared to the other techniques [145]. Thus, PCA is used in this work to reduce separately the dimensions of the visual and audio mid-level representations.

Two independent PCAs are learned on the training set. Audio and visual mid-level representations are first computed for the training set and standardized (*i.e.*, normalized to zero mean and unit variance). Then two PCAs are learned on the standardized values: one to reduce the dimension of the audio mid-level representation, and the second one for the visual mid-level representation. Both compact representations are then concatenated to feed the LSTM-RNN. These PCAs are applied to reduce the dimension of the mid-level representations for the validation and test sets, previously standardized using the mean and variance computed from the training set. Several reduced dimensions were investigated (dimension reduced to 128, 256, ..., 1024, *i.e.*, no data reduction is performed for the visual representations for this last value). Best performance was achieved using a dimension reduced from 1,024 to 896 for the visual mid-level representations, and from 4,096 to 128 for the audio mid-level representations, for both valence and arousal. Using these new dimensions, 98% of the variance in original data in the training set is explained for the visual mid-level representations, while this percentage equals 99% for the audio mid-level representations despite the drastic reduction in size. In fact, the audio mid-level representations are very sparse. For example, if its dimension is reduced to 512, 99,9% of the variance of the original data is still explained.

In the following sections, a compact multimodal mid-level representation of size 1,024 is used to feed the bidirectional LSTM-RNN. Figure 9.5 summarizes the main steps used to generate the compact multimodal mid-level representations.

9.2.4 Data-augmentation

As for CNNs in Section 9.1.1, LSTM-RNNs are also prone to overfitting and can benefit from a data-augmentation technique, especially since the training set is composed of only 15 temporal sequences corresponding to the consecutive compact mid-level representations for the 15 movies of the training set.

Training methodology

To artificially augment the number of temporal sequences, several sequences are generated per movie, each using the center 224×224 crop from a different 256×256 patch extracted in Section 9.1.1 to generate visual mid-level representations. As in Section 9.1.1, the ground truth associated to each 1-second video segment inside a temporal sequence is considered as a random variable using the parameters previously defined. Thus, the temporal sequences generated for a single movie show different compact-mid-level representations and slightly different ground truths.

In Section 9.1.1, six patches were extracted from five key frames. 30 temporal sequences could thus be generated per movie. However, the

temporal sequences need to be written in the NetCDF file format to be used by the tool to train LSTM-RNNs. The NetCDF file format requires that all the data to be written should fit at the same time in a single table in memory. Since 24GB of memory were available for the experiments, due to memory constraints five temporal sequences only have been generated per movie using the most representative patches, *i.e.*, the center 256×256 patch k of the five key frames, with $k \in \{1, 2, 3, 4, 5\}$ the number of the key frame from with the k^{th} patch has been extracted. In other words, for each movie, for each 1-second length video segment, the center crop of the center patch extracted from the first key frame ($k = 1$) is used to generate a compact multimodal representation. The consecutive compact multimodal representations form the first temporal sequence for the movie. Four other sequences are generated using the center crop of the center patch extracted from either the second, third, fourth, or fifth key frame. Finally, 75 temporal sequences are used to train the bidirectional LSTM-RNN for valence or arousal.

Batch learning by BPTT [146] is used to train the bidirectional LSTM-RNNs with mini batches of 20 sequences. The training is performed with the CURRENNT library [1] [147]. To prevent overfitting, Gaussian noise with standard deviation 0.1 is applied to the input sequences of the training set (it is not applied to validation and test sets). The steepest descent optimizer is used with momentum 0.9 and learning rate 1e-6. Several architectures for the bidirectional LSTM-RNN have been investigated. The best performing architecture is composed of two hidden layers with respectively 156 and 32 bidirectional LSTM cells. Each layer is fully connected. The output layer consists of a single linear summation unit to predict the induced affective score (arousal or valence).

The networks are trained for a maximum of 1,000 epochs. The network weights are initialized randomly using a normal distribution with mean 0 and standard deviation 0.1. The training is stopped if no improvement of the performance in terms of Sum of Squared Error is observed for more than 30 epochs on the validation set, evaluated at each epoch. The best network for the validation set is saved and its performance is computed for the test set.

Test strategy

To compute the performance of the advanced spatio-temporal model on the test set with respect to the continuous ground truth (for valence or arousal), five affective curves are generated using the same patches as in the data-augmentation of the training methodology. Please note that the ground truth of the validation and test set is the "raw" ground truth: it is no longer considered as composed of random variables. To combine these five affective curves, two strategies are compared. The first strategy is a simple averaging:

$$\hat{y_i} = \sum_{k=1}^{5} \hat{y_{k,i}} \tag{9.6}$$

1. http://sourceforge.net/projects/currennt/

with $\hat{y}_i$ the final affective score assigned to the i^{th} fragment of a movie, and $\hat{y_{k,i}}$ the affective score generated by the model using the k^{th} patch (represented in Figure 9.5).

The second strategy is to compute the weights of a weighted average maximizing the performance for the validation set. Then, these weights are used to combine the five curves of the test set:

$$\hat{y}_i = \sum_{k=1}^{5} w_k \hat{y_{k,i}} \tag{9.7}$$

with w_k the weight of the k^{th} affective curve.

But how the number of the key frame from which the center patch used to generate the visual-based mid-level representations could impact the performance of the affective curve? In other words, why the affective curve created using the mid-level representations generated using for example the center patch from the first key frames could be more or less reliable than the affective curve generated using the center patch from the other key frames? To try to answer to this question, we first need to go back to the key frame generation in Section 9.1.1. A frame is defined as a key frame if its YUV histogram is the closest, using the Manhattan distance, to a cluster computed by the k-mean clustering algorithm (with $k = 5$). The implementation of the k-mean algorithm biases the number of the key frame to the extent that black key frames are often labeled with the first or last labels. Thus, if a black frame exists in a 1-second length video segment, it is more likely that it will be the first or fifth key frame of the segment.

Once combined, the predicted curves for the movies are smoothed to improve the performance of the model using a Gaussian function and compared to the ground truth. Indeed, Malandrakis *et al.* have shown that smoothing continuous predicted curves can improve the performance of continuous models [52]. Again, the Gaussian parameters maximizing the performance for the validation set are computed and applied to smooth the predicted curves for the test set. The results for both combination strategies are detailed in Table 9.5 and discussed in the next section.

9.3 Experimental results

In this thesis, several computational models for continuous movie content analysis have been proposed, all learned and evaluated using a dataset composed of 30 movies annotated along the continuous induced valence and arousal axes (Chapter 6). These models are compared in this section.

9.3.1 Performance analysis

In the previous chapter, a baseline based on SVR and transfer learning has been proposed. In this chapter, a multimodal static model has first been achieved (Section 9.1.4) which predicts the induced valence or arousal for 1-second length video segments. An efficient way for such a static model to consider the consecutive video segments together is to

Table 9.5 – *Performance for the advanced temporal multimodal model compared to the static models*

Model	Arousal		Valence	
	MSE	r	MSE	r
Random	0.109	0.0004	0.113	-0.002
Uniform	0.026	-0.016	0.029	-0.005
SVR with transfer learning baseline	0.022	0.337	0.034	0.296
Multimodal static model (best fusions)	0.018	0.170	0.027	0.349
Multimodal static model with Gaussian smoothing	0.020	0.208	0.025	0.489
LSTM-RNN (simple average)	0.018	0.289	**0.024**	**0.559**
LSTM-RNN (best weights)	**0.017**	**0.361**	**0.024**	**0.559**

apply a smoothing function on the consecutive scores that have been predicted independently. For example, Hanjalic and Xu applied a Kaiser smoothing window to merge neighboring predicted local arousal and valence values [50]. Finally, a spatio-temporal model has been proposed in the previous section using bidirectional LSTM-RNN to output continuous affective curves. Table 9.5 details the performance for all these models.

First, Table 9.5 shows that a simple Gaussian smoothing significantly improves the performance of the multimodal static model for valence predictions (-7% for MSE and +40% for r). The correlation for arousal is also better but is still limited and far to be as good as the performance in terms of correlation of the SVR with transfer learning baseline for arousal. However, the multimodal static model, with and without Gaussian smoothing, outperforms the baseline for the prediction of the induced valence values.

The advanced spatio-temporal model with simple averaging as well as with the best weights approach outperforms the multimodal static model. However, with the simple averaging approach, the performance in terms of correlation is still smaller than the baseline. It is the best weights approach that makes possible for the spatio-temporal model to outperform the baseline for both valence and arousal and in terms of both MSE and correlation. The best weights selected to combine the five predicted curves using for each consecutive video segment the center patch extracted from one of the key frames shows that for valence the five curves are all equally relevant. This is why the results are the same for valence for the simple average and best weights approaches. However, for arousal the best weights show that the first and fifth predicted curves using the patches from the first or fifth key frames are less relevant. Indeed, the five weights selected to combine the five predicted curves are respectively, for the first to the last curve: 0.07, 0.22, 0.35, 0.34, and 0.02. As stated in the previous sections, black key frames are more likely to be assigned as the first or

fifth key frame. It seems from these results that they are less meaningful to predict induced arousal scores.

Even if the performance of the spatio-temporal model outperforms all the other models and is significantly better than the performance obtained by chance or from a uniform model, the performance for arousal is still limited. First, we may surmise that the architecture of the model is not optimal to predict arousal scores. Temporal information is considered once independent compact mid-level representations are generated. To enhance the prediction performance for arousal, temporal mid-level representations based on motion intensity maps for consecutive video frames could probably help. Second, since temporal information is very important to accurately predict arousal scores, we can also suppose that the temporal granularity (1Hz) is not precise enough and that using finer temporal granularities, *e.g.*, one affective score per video frame, could help improving the performances. Finally, we can assume that ground truth for arousal is not as reliable as ground truth for valence. Indeed, in Chapter 6, we mentioned that the inter-coder correlation of the self-assessments for valence ($mean = 0.313 \pm 0.195\ SD$) was higher than for arousal ($mean = 0.275 \pm 0.195\ SD$). Actually, valence is a concept easier to understand than arousal, at least for French participants. It was thus easier for them to self-assess their level of valence than their level of arousal. Consequently, the mean of the self-assessments for arousal may be less representative of the emotions actually felt by participants than for valence. To conclude, it is more likely a combination of these factors that influences the performance of the predicted arousal curves for the spatio-temporal model. The last chapter of this thesis proposes further discussion as perspectives for future research work.

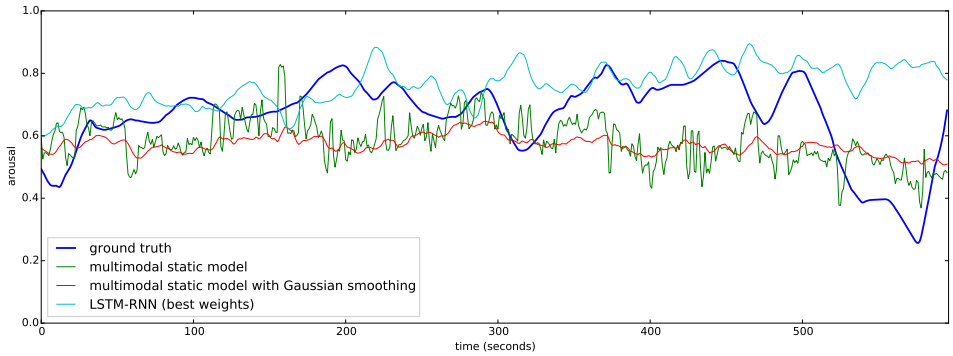
9.3.2 Affective curves

The results given in the previous section are global results that do not explicitly indicate how generated affective curves temporally fit with the ground truth. This is why in this section, a qualitative evaluation is performed for affective curves generated for three movies. Affective curves for arousal and valence, generated by the multimodal static model and by the spatio-temporal model, are shown for the movies Big Buck Bunny, Full Service, and Payload, respectively in Figures 9.6, 9.7, and 9.8.

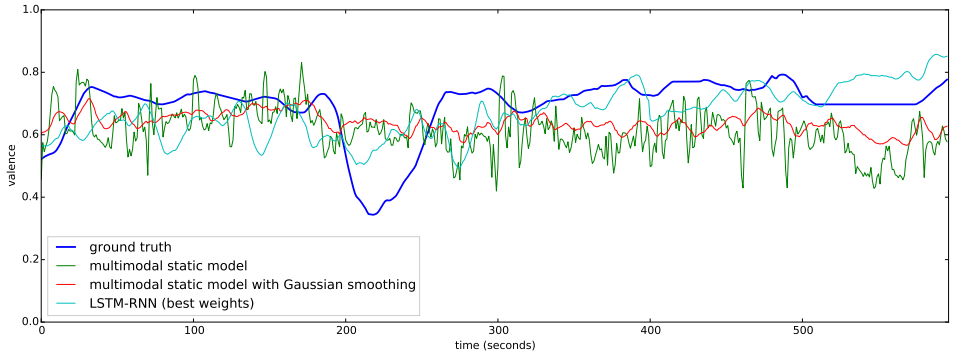
Globally, the multimodal static model seems closer to neutral values for arousal and valence while the spatio-temporal model fits more closely the trends of the ground truth. However, it seems that none of the models is able to accurately model the rapid but short increases or decreases of arousal or valence. Raw predictions from the multimodal static model are also very noisy due to the fact that it takes into account individually the consecutive video segments extracted from the movies.

Summary

In this chapter, a static model based on two unimodal fine-tuned CNNs has been presented first. The MLDA, combined with the ground truth considered as random variables, artificially augments the size of the training

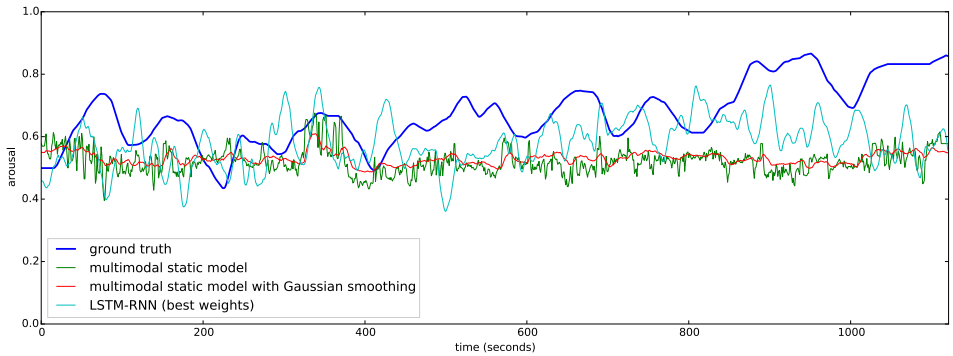


(a) Arousal

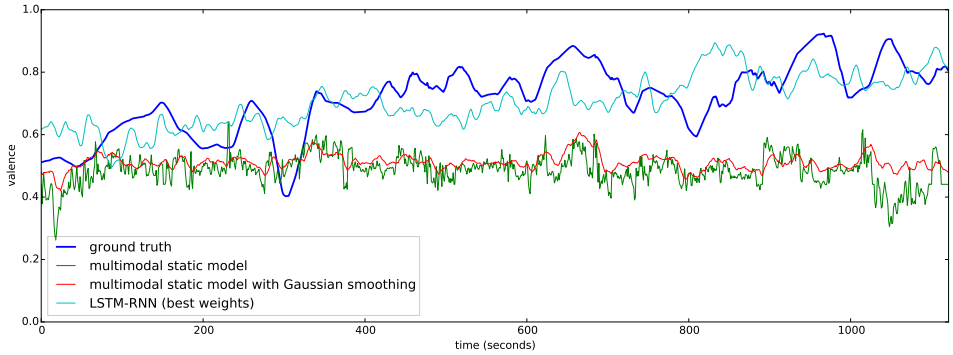


(b) Valence

Figure 9.6 – *Predicted affective curves for Big Buck Bunny*

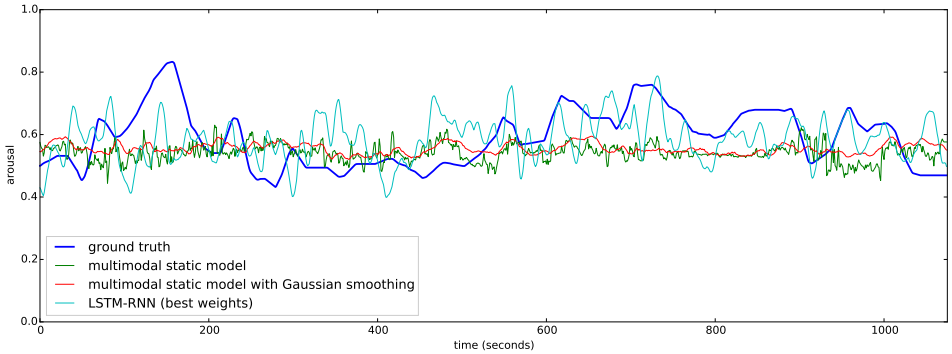


(a) Arousal

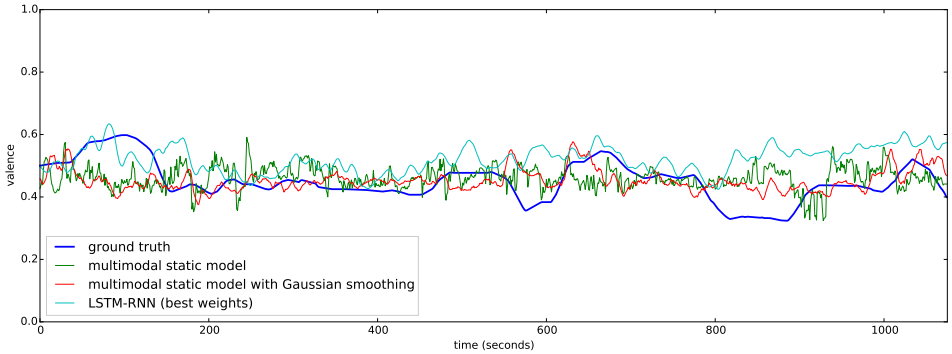


(b) Valence

Figure 9.7 – *Predicted affective curves for Full Service*



(a) Arousal



(b) Valence

Figure 9.8 – *Predicted affective curves for Payload*

set to prevent overfitting for the visual-based CNN. For the audio-based CNN, spectrograms are extracted from the audio signal. The activations of the last fully connected layers from both CNNs are used as mid-level representations. Then, the size of these mid-level representations is reduced using PCA. The compact multimodal mid-level representations are used by a bidirectional LSTM-RNN to predict continuous affective curves, taking into account the past and future long-term dependencies between the video segments. The five affective curves generated per movies are finally combined using a weighted average to compute the final affective curve.

A correlation of 0.559 is achieved for valence between the ground truth and the final affective curves for the seven movies of the test set. This correlation is equal to 0.361 for arousal. The spatio-temporal model significantly outperforms the baselines from Chapter 8.

Conclusions

The work conducted in this thesis focused on the automatic prediction of emotions induced by movies. Two main problems were addressed: the creation of a large and robust dataset that can be shared among researchers, and the automatic estimation of the emotions induced by movies using a reliable computational model.

This chapter concludes this thesis, summarizing the main contributions and results. Specific highlights are also emphasized. Beyond this work, perspectives are detailed for future affective movie content analysis work.

Achievements

In the first part of this thesis, Chapters 2 and 3 gave an overview of emotion theories, emotion representations, as well as computational models and existing datasets for affective video content analysis. A review of previous work developing computational models to infer emotions showed that they mostly suffer from one or more of these issues:

- They represent emotions using discrete categories. However, a discrete representation is subject to ambiguities and do not cover the whole range of emotions elicited by movies.
- They rely on a predefined set of handcrafted audiovisual features extracted from videos. However, building complex handcrafted features requires strong domain knowledge and is highly problem-dependent. Obtaining a satisfying feature extraction is thus not a trivial issue.
- They do not take into account the fact that an emotional episode is a recursive process.
- They use private datasets to evaluate their performance, thus making fair comparisons and results reproducibility impossible, and preventing achievement of major strides in the field.

This is why in the second part of this thesis, a public affective dataset is developed as a first key contribution shared with the community. In Chapters 4, 5, 6, and 7, the LIRIS-ACCEDE dataset is introduced. It is composed of 9,800 video segments extracted from 160 movies shared under CC licenses. 30 from these 160 movies have also been continuously annotated. The use of movies shared under CC licenses make it possible to share the database publicly without copyright issues. Details about the use of this dataset by the research community are given in the next section.

First, the 9,800 excerpts have been annotated using crowdsourcing to reach a large number of remunerated annotators. The 9,800 excerpts have

been ranked along the induced valence and arousal axes using pairwise comparisons to improve the consistency of annotations. Comparisons were generated using the quicksort algorithm to reduce costs. To rank the 9,800 excerpts along the induced valence axis, more than 582,000 annotations for about 187,000 comparisons were gathered from 1,517 trusted annotators from 89 countries, while for arousal, more than 665,000 annotations for around 221,000 unique comparisons were gathered from 2,442 trusted annotators also from 89 countries. Four reproducible protocols using these 9,800 excerpts were defined to allow fair comparisons between future computational models for discrete movie content analysis. However, the ranks provide no information on distribution of the dataset in the 2D VA space. To address this limitation, we carried out a complementary experiment to collect in a controlled environment absolute video ratings from 28 volunteers for a subset of 40 excerpts from the dataset. The significant correlation for both arousal ($SRCC = 0.751, t(44) = 7.635, p < 1 \times 10^{-8}$) and valence ($SRCC = 0.795, t(44) = 8.801, p < 1 \times 10^{-10}$) between crowdsourced ranks and the ratings collected in the controlled environment validated the annotations gathered using crowdsourcing. Using Gaussian Processes for Regression that can model the noisiness from measurements, affective scores were also estimated for all the 9,800 excerpts of the database.

Second, another experiment was performed to collect continuous ratings for 30 movies. Using joysticks and a modified version of the GTrace annotation tool [102], 10 French paid participants continuously annotated half of the movies along the induced valence axis and the other half along the induced arousal axis. Continuous ratings were post-processed in order to generate the average continuous emotions induced by the movies in terms of valence and arousal. Finally, the GSR response from 13 participants watching these movies was also recorded. A positive temporal Pearson's r correlation of 0.264 was found between the mean of the continuous arousal self-assessments and the post-processed GSR measurements. This correlation confirmed that arousal and GSR are correlated, but foremost validated the reliability of both the arousal self-assessments and GSR measurements.

Armed with a dataset, a baseline was proposed in Chapter 8 for discrete affective movie content analysis using the protocols defined for the LIRIS-ACCEDE dataset. Four baselines for continuous affective movie content analysis were also proposed in Chapter 8 based on SVRs, CNNs, or the combination of both machine learning techniques, and trained using as ground truth the continuous self-assessments for the 30 movies of LIRIS-ACCEDE. The promising performance achieved by these four baselines paved the way for the creation of an advanced spatio-temporal computational model described in Chapter 9. This model predicts affective scores for consecutive 1-second length segments extracted from a movie. Two fined-tuned CNNs generate mid-level representations. One is dedicated to the visual modality and uses patches extracted from the key frames of the video segment, while the second one is dedicated to the audio modality and uses as input an audio spectrogram generated using the audio signal of the video segment. The size of both mid-level representations is then reduced. The concatenation of both representations,

named compact multimodal mid-level representation, is then used by a bidirectional LSTM-RNN modeling long-term dependencies between the consecutive video segments from the same movies. A correlation between predicted curves and ground truths of 0.361 was obtained for arousal. For valence, the correlation was considerably higher: 0.559. The performance of the spatio-temporal model outperforms all the other baselines and is significantly better than the performance obtained by chance.

However, even if these correlations are satisfying for the automatic prediction of the emotions induced by movies, the corresponding computational model is not reliable enough to be used in commercial systems. For example, this performance for the prediction of induced emotions is far to be as good as those for emotion recognition (correlations of 0.73 ,0.74, 0.75, and 0.76 were achieved in 2014, respectively for the recognition of the session independent expressed valence, arousal, power, and expectation [148]), or less subjective detections such as object recognition (classification error rate of 6.7% was achieved in 2014 for the image classification task of the ImageNet Large Scale Visual Recognition Challenge [149]). There is still room for improvement and significant breakthroughs need to be developed before the large-scale use of such automatic affective movie content analysis models.

Highlights

As of summer 2015, the LIRIS-ACCEDE dataset has been downloaded more than 70 times from various research centers all around the world and has been already used in published papers [150, 151]. LIRIS-ACCEDE was one of the recommended dataset for the "Emotional Response to Multimedia Content" grand challenge at ACM MM 2014 [2] and it is the only dataset used for the MediaEval 2015 "Affective Impact of Movies (including Violent Scenes Detection)" task [3].

Perspectives

The automatic prediction of emotions induced by movies is a very challenging task. The framework developed in this thesis is quite general and may prove useful in a variety of video processing applications. However, leads can be followed up to solve the limitations and drawbacks of the proposed methodology.

The performance of computational models to predict induced emotion is closely tied to the difficulty of collecting large and reliable affect datasets. In particular, collecting continuous dimensional affective self-assessments is not easy and experimental protocols should be designed very carefully. The continuous annotations within the LIRIS-ACCEDE dataset have shown that it is possible to collect and release publicly such annotations. However, its size is still quite limited for current machine learning techniques and should be extended by collecting new self-assessments for new movies along new affective dimensions such as dom-

2. http://acmmm.org/2014/call_mm_grnd_chlng_sol.html
3. http://www.multimediaeval.org/mediaeval2015/affectiveimpact2015/

inance. This again shows the importance to design reproducible experimental protocols to benefit from the contributions of various research teams to build a large and representative affective dataset that can be shared among researchers.

The difficulty of collecting reliable affective ground truth is another remaining challenge. The word "ground truth" should be taken with a grain of salt in affective multimedia analysis work. Indeed, the affective self-assessments collected in the experiments presented in this thesis, but also in all the other work dealing with the emotions induced by multimedia content, are intrinsically biased to the extent that they represent the interpretation of the emotional experience that the annotator is currently feeling, which may be different from the emotion felt by the annotator. Recording the facial expression of the annotators and collecting physiological signals, such as the skin conductance, the heart rate, or even the electroencephalogram signals, could help to improve the reliability of the affective self-assessments. However, the correlation between such modalities and felt emotions is still a work in progress [152]. The continuous ground truth presented in Chapter 6 is also biased because of the post-processing steps to take into account the annotator-specific delays amongst the annotations, and because it is the result of the aggregation of the multiple annotators' self-assessments. To tackle these issues, future work should start investigating individual emotional differences to design personalized models. To summarize, the continuous "ground truth" presented in this thesis, but also the "ground truth" used in most affective multimedia analysis work, does not represent the emotions that an annotator has felt during an experiment, but rather represents the emotions that most annotators say they have experienced while watching movies during an experiment.

Many extensions and improvements can also be envisaged to improve the reliability of the continuous spatio-temporal models proposed in this thesis:

- First, psychological theories and computational models barely rely on each other. Future computational models should be designed to be closer to psychological models.
- Predicting intermediate dimensions, such as predictability or novelty, may be of interest to help estimating valence and arousal induced by movies. Psychological theories have shown that these two dimensions are important characteristics of the emotional experience [153].
- The proposed continuous spatio-temporal model is quite limited to the extent that it only uses the information provided by key frames and the audio signal. Motion, text (subtitles), or even more advanced concepts (*e.g.*, relations among movie characters [154]) should improve the reliability of the model.
- Arousal and valence are correlated. It thus makes sense to predict these affective scores using a single computational model.
- Predicting the emotions that most people feel while watching movies is extremely difficult to be solved with a universal solution. Personalized solutions should also be considered to improve the accuracy of computational models, which will be made possible

by collecting self-assessments across highly diverse viewer groups with different cultural and socio-demographical backgrounds.

This concluding discussion shows that the future tasks in this area of research are challenging and need to be addressed by close collaboration of experts in psychology, sociology, vision science and image processing.

Appendices

Appendix A: List of the Movies in LIRIS-ACCEDE

Name	Credits	Length	License
20 Mississippi	Barnett Brettler	00:58:38	20 Mississippi shared under CC BY-NC 3.0 Unported license at http://vimeo.com/20043857
21 Below	Robbie Stauder	01:31:17	21 Below shared under CC BY 3.0 Unported license at http://vimeo.com/38939998
52 Films/52 Weeks	Renee Ronceros, Samantha Simmonds & Javier Ronceros	03:01:55	52 Films/52 Weeks: a year of filmmaking shared under CC Public Domain 3.0 license at http://www.52films52weeks.com/52films52weeks/Welcome.htm
After The Rain	Hits Enterprises & Video	00:09:49	After The Rain shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/40104084
Attitude Matters	Marco Luca & Laura Aloi	00:22:52	Attitude Matters shared under CC BY 3.0 Unported license at http://vimeo.com/17778716
Barely Legal Stories	Jonathan Musset	00:16:28	Barely Legal Stories shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/30344973
Best	Nice Monster	00:07:56	Best shared under CC BY 3.0 Unported license at http://vimeo.com/44163644
Between Viewings	Raphael Biss	00:14:46	Between Viewings shared under CC BY 3.0 Unported license at http://vimeo.com/43763789
Big Buck Bunny	Sacha Goedegebure	00:09:56	Big Buck Bunny shared under CC BY 3.0 license at http://www.bigbuckbunny.org/
Boiling Point	Jack Leigh	00:14:04	Boiling Point shared under CC BY-SA 3.0 Unported license at http://vimeo.com/24169479
Burgundies Boys	Steve Galley	01:18:43	Burgundies Boys shared under CC BY-NC 3.0 Unported license at http://vodo.net/boys
California Dreaming	Bregtje van der Haak	00:49:13	California Dreaming shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/californiadreaming
Capitalism Communism: Is this love?	FILS-PRODUC-TION	00:07:27	Capitalism Communism Is this love? shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/53307142
Chatter	Leo Resnes	00:08:29	Chatter shared under CC BY-NC 3.0 Unported license at http://vodo.net/Chatter
Clickin' For Love	Pablo Pappano	01:28:39	Clickin' For Love shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/clickinforlove
Climate Cycle	Paul O'Connor	00:21:41	Climate Cycle shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/ccycle
Cloudland	LateNite Films	00:11:41	Cloudland shared under CC BY 3.0 Unported license at http://vimeo.com/17105083
Cold	Bahadir Karasu	00:21:23	Cold shared under CC BY-SA 3.0 Unported license at http://vimeo.com/41402223
Copyright is for losers	Ninotchka Art Project	00:24:39	Copyright is for losers shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/2026149
Couchsurf	Georg Boch	00:18:51	Couchsurf shared under CC BY-NC-SA 3.0 Unported License license at http://vodo.net/couchsurf
Crooked Features	Mike Peter Reed	01:25:01	Crooked Features shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/crookf
Damaged Kung-Fu	Juliane Block	00:16:54	Damaged Kung-Fu shared under CC BY-SA 3.0 license at http://www.filmannex.com/movie/damaged-kung-fu/30279
Dead Man Drinking	Rohan Harris	01:28:26	Dead Man Drinking shared under CC BY-NC-SA 3.0 Unported license at http://www.deadmandrinking.com/
Decay	CERN by physics PhD students	01:16:06	Decay shared under CC BY-SA 3.0 Unported license at http://vimeo.com/55157792
Deceived	Winkler Pictures	01:31:01	Deceived shared under CC BY 3.0 Unported license at http://vimeo.com/39057892
Dimensional Meltdown	Ofer Pedut	00:07:34	Dimensional Meltdown shared under CC BY-NC 3.0 Unported license at http://vodo.net/dimensional
Down With the King	Michael Wolcott	00:30:33	Down With the King shared under CC BY-NC 3.0 Unported license at http://vodo.net/dwtk
Elephant's Dream	Bassam Kurdali	00:10:53	Elephant's Dream shared under CC BY 3.0 license at http://orange.blender.org/
Emperor	Juliane Block & Adrian Lai	01:35:56	Emperor shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/emperor
END:CIV	Franklin Lopez	01:16:45	END:CIV shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/endciv
Fall and Love	Jordan Baker	00:12:43	Fall and Love shared under CC BY 3.0 Unported license at http://vimeo.com/52084858
First Bite	Dead Flower Productions	00:10:40	First Bite shared under CC BY 3.0 Unported license at http://vimeo.com/23980578

Continued on next page

Continued from previous page

Name	Credits	Length	License
Four Eyed Monsters	Arin Crumley & Susan Buice	01:11:57	Four Eyed Monsters shared under CC BY-SA 3.0 Unported license at http://vodo.net/foureyedmonsters
Full Service	Ian Quill	00:18:41	Full Service shared under CC BY 3.0 Unported license at http://vimeo.com/22719579
Good Boys go to Heaven and Bad Boys go to Europe	Fabrice Renucci	01:08:37	Good Boys go to Heaven and Bad Boys go to Europe shared under CC BY 3.0License license at http://vodo.net/goodboys
Here. My Explosion...	Jeffery Davis, Eleese Longino & Seth Burnham	01:14:50	Here. My Explosion... shared under CC BY-SA 3.0 Unported license at https://www.createspace.com/267787
Home	Yann Arthus-Bertrand	01:33:17	Home shared under CC BY-SA 3.0 Unported license at http://archive.org/details/Home2009
How Fear Came	Anais Caura & Bulle Tronel	00:10:06	How Fear Came shared under CC BY-NC 3.0 Unported license at http://vimeo.com/42690350
Interferencies	Debt Observatory (ODG) and Quepo	01:13:49	Interferencies shared under CC BY-NC-SA 3.0 Unported license at http://www.interferencies.cc/
Iron Sky Teaser 3: We Come In Peace	Stealth Media Group	00:01:00	Iron Sky Teaser 3 We Come In Peace shared under CC BY-SA 3.0 Unported license at http://www.ironsky.net/
Islands	Diego Contreras	00:02:53	Islands shared under CC BY 3.0 Unported license at http://vimeo.com/50512824
Je suis ce que je vois	Simon Bonneau	00:02:21	Je suis ce que je vois shared under CC BY-NC-SA 3.0 license at http://www.youtube.com/user/TheChivteam
Jiminy	Jakk in the Box	00:08:49	Jiminy shared under CC BY 3.0 Unported license at http://vimeo.com/21791346
L.U.C.K	Daniel Cooper	00:10:31	L.U.C.K shared under CC BY 3.0 Unported license at http://vimeo.com/37041387
Le Fear	Jason Croot	01:01:32	Le Fear shared under CC BY 3.0 Unported license at http://vodo.net/lefear
Lesson Learned	Fritz Joseph	00:12:58	Lesson Learned shared under CC BY-SA 3.0 Unported license at http://vimeo.com/40539260
The Lionshare	Josh Bernhard	01:08:16	The Lionshare shared under CC BY-NC 3.0 Unported license at http://vodo.net/lionshare
Lo que tu Quieras Oir	Guillermo Zapata	00:10:15	Lo que tu Quieras Oir shared under CC BY-NC-SA 3.0 Unported license at http://creativecommons.org/weblog/entry/7537
Lusty Little Heart of Mine	Martin Heuser	00:18:30	Lusty Little Heart of Mine shared under CC BY-NC 3.0 Unported license at http://vodo.net/lusty
Monolog	Eray Dinc	01:17:36	Monolog shared under CC BY 3.0 Unported license at http://vimeo.com/20235811
Nasty Old People	Hanna Skold	01:24:25	Nasty Old People shared under CC BY-NC-SA 3.0 Unported license at http://nastyoldpeople.blogspot.fr/
Norm	Elle Marsh	00:06:30	Norm shared under CC BY 3.0 Unported license at http://vimeo.com/43513380
Nuclear Family	Dominic Mercurio	00:28:20	Nuclear Family shared under CC BY-NC 3.0 Unported license at https://www.facebook.com/nuclearfamilymovie
Oceania	Harry Dehal	00:54:19	Oceania shared under CC BY-NC-SA 3.0 United StatesLicense license at http://www.hdehal.com/filmandvideo.php
Of Games and Escapes	Bevan Klassen	01:17:05	Of Games and Escapes shared under CC BY-NC 3.0 Unported license at http://vodo.net/GamesEscapes
On Time	Todd Wiseman	00:05:11	On Time shared under CC BY-NC-SA 3.0 Unported license at http://www.youtube.com/watch?v=8zI2JRLFQoE
Origami	ESMA MOVIES	00:08:20	Origami shared under CC BY 3.0 Unported license at http://vimeo.com/52560308
Parafundit	Riccardo Melato	00:13:10	Parafundit shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/50060660
Payload	Stu Willis	00:17:55	Payload shared under CC BY 3.0 Unported license at http://vimeo.com/50509389
Pennipo-tens	Heather Freeman	00:17:38	Pennipotens shared under CC BY-NC-SA 3.0 Unported license at http://pennipotens.blogspot.fr/
Pioneer One	Bracey Smith	03:43:36	Pioneer One E01-06 shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/pioneerone
Point Of Departure	Matthias Merkle	01:33:19	Point Of Departure shared under CC BY-NC-SA 3.0 Unported license at http://wiki.creativecommons.org/Point_Of_Departure
Riding the Rails	Juan Soto	00:15:00	Riding the Rails shared under CC BY 3.0 Unported license at http://vimeo.com/17465105
Riflessi	Emanuela Ponzano	00:20:52	Riflessi shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/31900322
RIP! A Remix Manifesto	Girl Talk, Lawrence Lessig, Gilberto Gil & Cory Doctorow	01:27:20	RIP! A Remix Manifesto shared under CC BY-NC 3.0 Unported license at http://creativecommons.org/tag/rip-a-remix-manifesto
Rising Tide	Philip Shotton & Dawn Furness	01:18:52	Rising Tide shared under CC BY-NC-SA 3.0 Unported License license at http://www.risingtidethemovie.com/
Roskilde The Experience	Per Tore Holmberg	00:22:15	Roskilde The Experience shared under CC BY-NC-SA 3.0 Unported license at http://www.roskilde-experience.com/
Santa Cruz Beach Boardwalk	Eugenia Loli	00:04:54	Santa Cruz Beach Boardwalk shared under CC BY 3.0 Unported license at http://vimeo.com/2886954
Scum-babies	Joseph R. Lewis	01:31:03	Scumbabies shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/scumbabies
Seven	Desmond Bowles & Dave Dornbrack	00:16:40	Seven shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/55103414
Sintel	Martin Lodewijk	00:14:48	Sintel shared under CC BY 3.0 license at www.sintel.org
Sita Sings the Blues	Nina Paley	01:21:31	Sita Sings the Blues shared under CC BY-SA 3.0 Unported license at http://www.sitasingstheblues.com/

Continued on next page

Continued from previous page

Name	Credits	Length	License
Sky in Slow motion: Fetch!	Jenifer Avila	00:03:59	Sky in Slow motion: Fetch! shared under CC BY-NC 3.0 Unported license at http://www.youtube.com/watch?v=aG6R771H1FQ
Solace	Daniel Cooper	00:10:05	Solace shared under CC BY 3.0 Unported license at http://vimeo.com/42505454
Spaceman	Jono Schaferkotter & Before North	00:15:29	Spaceman shared under CC BY-NC 3.0 Unported License license at http://vodo.net/spaceman
Steal this film	The League of Noble Peers	00:51:30	Steal this film shared under CC BY 3.0 Unported license at http://vodo.net/stf
Sugar	Andrew John Rainnie	00:22:27	Sugar shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/Sugar
Superhero	Langley McArol	00:19:01	Superhero shared under CC BY 3.0 Unported license at http://vimeo.com/23423341
Suspicious Minds	Tizaster Productions	00:08:07	Suspicious Minds shared under CC BY 3.0 Unported license at http://vimeo.com/8833583
Sweet Hills	Not Working Films	00:16:48	Sweet Hills shared under CC BY-NC 3.0 Unported license at http://vimeo.com/51445616
Tears of steel	Ian Hubert & Ton Roosendaal	00:12:14	Tears of steel shared under CC BY 3.0 Unported license at http://www.tearsofsteel.org/
The Box	BK	00:27:47	The Box shared under CC BY 3.0 Unported license at http://vimeo.com/3610952
The Cosmonaut (Trailer)	Nicolas Alcala	00:02:21	The Cosmonaut (Trailer) shared under CC BY-NC-SA 3.0 Unported (CC BY-NC-SA 3.0) license at http://www.thecosmonaut.org/
The Dabbler	Reid Gershbein	00:59:59	The Dabbler shared under CC BY-SA 3.0 Unported license at http://www.royalbaronialtheatre.com/blog/the-dabbler-film-details-2wkfilm.html
The frame of the dead	Samuel Sebastian	01:48:37	The frame of the dead shared under CC BY 3.0 Unported license at http://vodo.net/framedead
The Graduates	Ryan Gielen	01:31:08	The Graduates shared under CC BY-NC 3.0 Unported license at http://vodo.net/thegraduates
The Great Commandment	Maurice Moscovitch	01:20:12	The Great Commandment shared under Public Domain License license at http://www.youtube.com/watch?v=keOrz7n7ETo
The idea	Berthold Bartosch	00:25:21	The idea shared under Public Domain License license at http://vimeo.com/20866713
The Immortals	Matthias Merkle & Antje Borchardt	01:34:59	The Immortals shared under CC BY-NC-SA 3.0 Unported license at http://wiki.creativecommons.org/The_Immortals
The Manifesto	Mr Nobody	01:30:32	The Manifesto shared under CC BY-SA 3.0 Unported License license at http://vodo.net/Manifesto
The Mapmaker	Twisty-Headed Man Company	00:26:17	The Mapmaker shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/mapmaker
The Master Plan	Aron Campisano	01:44:20	The Master Plan shared under CC BY-NC-SA 3.0 Unported license at http://themasterplanfilm.com/
The room of Franz Kafka	Fred. L'Epee	00:04:09	The room of Franz Kafka shared under CC BY-NC-SA 3.0 Unported license at http://vimeo.com/14482569
The secret number	Colin Levy	00:15:31	The secret number shared under CC BY-NC 3.0 Unported license at http://vimeo.com/43732205
Time Expired	Nick Lawrence & Rachel Tucker	00:32:42	Time Expired shared under CC BY-NC 3.0 Unported license at http://vodo.net/timeexpired
To Claire From Sonny	Ennui Pictures	00:06:54	To Claire From Sonny shared under CC BY-NC-SA 3.0 Unported license at http://www.youtube.com/watch?v=8rKW-VRFczA
To Kill A King	Run Productions	00:21:01	To Kill A King shared under CC BY-NC 3.0 Unported license at http://vimeo.com/30847762
Torno Subito	Simone Damianiunder	01:29:07	Torno Subito shared under CC BY-NC 3.0 Unported license at http://creativecommons.org/tag/torno-subito
Valkaama	Tim Baumann	01:33:13	Valkaama shared under CC BY-SA 3.0 Unported license at http://www.valkaama.com/
Viaje a la tierra del Quebracho	TEMBE Cooperativa	00:11:53	Viaje a la tierra del Quebracho shared under CC BY-SA 3.0 Unported license at http://vodo.net/quebracho
Waldo the Dog	Kris Canonizado	02:00:39	Waldo the Dog shared under CC BY-NC-SA 3.0 Unported license at http://vodo.net/WaldotheDog
Wanted	Ezel Domanic	00:01:57	Wanted shared under CC BY 3.0 Unported license at http://www.filmannex.com/movie/wanted/30133
When Rabbits Fly	Helgi Johannsson	00:28:18	When Rabbits Fly shared under CC BY 3.0 Unported license at http://vimeo.com/58619416
You Again	Lauren Teng	00:14:30	You Again shared under CC BY 3.0 Unported license at http://vimeo.com/33454078

Appendix B: Formulas for Computing Inter-Rater Reliability

Appendix B lists all the formulas to compute the inter-rater reliability coefficients used throughout this thesis. In the following, N stands for the number of items to be annotated by the raters (also known as the number of subjects in the state of the art), n represents the number of annotations per item, k is the number of categories into which assignments are made, and n_{ij} represents the number of raters who assigned the j^{th} category to the i^{th} item.

Percent Agreement

The percent agreement P_a is defined as:

$$P_a = \frac{1}{Nn} \sum_{i=1}^{N} \max_j (n_{ij}) \tag{B.1}$$

Fleiss' κ

Fleiss' κ can be computed for any number of annotators giving categorical ratings to a fixed number of items [88]. All forms of kappa are defined as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{B.2}$$

where $\bar{P} - \bar{P}_e$ measures the degree of agreement actually achieved above chance and $1 - \bar{P}_e$ gives the maximum proportion of possible chance-adjusted agreement. Fleiss defines $\bar{P}$ and $\bar{P}_e$ as:

$$\bar{P} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{k} \left(n_{ij}^2 - Nn \right) \tag{B.3}$$

$$\bar{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \tag{B.4}$$

Fleiss' kappa is thus influenced by prevalence since it varies as a function of symmetry of marginal distributions.

Randolph's κ_{free}

Randolph's κ_{free} is subject to the same restrictions as Fleiss' κ and is based on the same equation (B.2) [90]. However, Randolph defines $\bar{P}_e$ as:

$$\bar{P}_e = \frac{1}{k} \tag{B.5}$$

Thus, Randolph's κ_{free} varies as a function of the number of categories and is thus not affected by prevalence.

Krippendorff's α

Krippendorff's α can be applied to any number of annotators, each assigning one value to one item, to incomplete data (*i.e.*, n is not fixed), and to any number and type of categories [89]. In Chapter 5, equations to compute the Krippendorff's α for a single excerpt i were detailed. The general equations are very similar and are thus not repeated here. The δ^2 coefficient for nominal categories was also defined in Chapter 5. Similarly, the δ^2 coefficient can be defined for ordinal, interval, ratio, polar, and circular categories.

Appendix C: Consent Form and Questionnaire

Personal Information

Please fill this paper before the first part of the experiment

All data will be coded so that your anonymity will be protected in any research papers and presentations that result from this work.

Name	_____	
Age	_____	
Sex	☐ Male	☐ Female
Nationality	_____	
Level of education	☐ "Collège" ☐ Bachelor ☐ Master ☐ Doctor and above	
	Other: _____	
Profession	_____	
Favorite movie genre	☐ Comedy ☐ Thriller ☐ Action ☐ Adventure	☐ Romance ☐ Drama ☐ Horror ☐ Documentary
	Other: _____	
How often do you watch movies?	☐ Never ☐ Once a month ☐ Once a week ☐ Several times a week	
	Other: _____	

Continuous annotations experiment

Consent form

1. Experiment Purpose & Procedure

The purpose of this experiment is to annotate 30 movies using a Joystick.

The experiment consists of 4 parts, during which you will be asked to watch movies and indicate the emotion you feel in response to the movie.

Before the experiment, you will be asked to complete a questionnaire.

Please note that none of the tasks is a test of your personal intelligence or ability. The objective is to collect ground-truth for research purposes.

2. Confidentiality

The following data will be recorded: age, sex, nationality, genre preference, annotations data.

All data will be coded so that your anonymity will be protected in any research papers and presentations that result from this work.

(If data is to be recorded that would identify the participant, for example photographs, audio or video, and if there is any intention to use this material in any publication or presentation, a separate release statement should be obtained after the recording has been made).

3. Record of Consent

Your signature below indicates that you have understood the information about the annotation experiment and consent to your participation. The participation is voluntary and you may refuse to answer certain questions on the questionnaire and withdraw from the study at any time with no penalty. This does not waive your legal rights. You should have received a copy of the consent form for your own record. If you have further questions related to this research, please contact the researcher.

Participant	Date	Signature
_____	_____	_____

Researcher	Date	Signature
_____	_____	_____

Bibliography

[1] K. R. Scherer, "The component process model: Architecture for a comprehensive computational model of emergent emotion," *Blueprint for affective computing: A sourcebook*, pp. 47–70, 2010. [cited at p. xiii, 10, 11, and 97]

[2] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *British Journal of Psychology*, vol. 95, pp. 489–508, Nov. 2004. [cited at p. xiii, 13, and 14]

[3] R. B. Dietz and A. Lang, "Aefective agents: Effects of agent affect on arousal, attention, liking and learning," in *Cognitive Technology Conference*, 1999. [cited at p. xiii, 16, and 44]

[4] Y. Baveye, F. Urban, C. Chamaret, V. Demoulin, and P. Hellier, "Saliency-guided consistent color harmonization," in *Computational Color Imaging Workshop*, vol. 7786, pp. 105–118, 2013. [cited at p. xiv, 87, and 88]

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012. [cited at p. xiv, 90, 91, and 92]

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014. [cited at p. xiv, 90, 99, and 102]

[7] G. M. Smith, *Film structure and the emotion system*. Cambridge University Press, 2003. [cited at p. 1]

[8] S. Arifin and P. Y. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325–1341, 2008. [cited at p. 1 and 31]

[9] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1473–1476, ACM, 2011. [cited at p. 1]

[10] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *Signal processing magazine, IEEE*, vol. 23, no. 2, pp. 90–100, 2006. [cited at p. 1]

[11] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, "A large video data base for computational models of induced emotion," in *Affective Computing and Intelligent Interaction (ACII)*, pp. 13–18, 2013. [cited at p. 3, 36, and 53]

[12] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, pp. 43–55, Jan 2015. [cited at p. 3 and 36]

[13] K. R. Scherer, "What are emotions? and how can they be measured?," *Social Science Information*, vol. 44, pp. 695–729, Dec. 2005. [cited at p. 9, 12, and 15]

[14] P. R. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion*, vol. 5, pp. 345–379, Dec. 1981. [cited at p. 9]

[15] M. Gendron and L. F. Barrett, "Reconstructing the past: A century of ideas about emotion in psychology," *Emotion review*, vol. 1, no. 4, pp. 316–339, 2009. [cited at p. 10, 11, and 12]

[16] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993. [cited at p. 10]

[17] P. Ekman, *Basic Emotions*. John Wiley & Sons, Ltd, 2005. [cited at p. 10 and 15]

[18] N. H. Frijda, *The emotions*. Cambridge University Press ; Editions de la Maison des Sciences de l'homme, 1986. [cited at p. 10]

[19] K. R. Scherer and T. Brosch, "Culture-specific appraisal biases contribute to emotion dispositions," *European Journal of Personality*, vol. 23, no. 3, pp. 265–288, 2009. [cited at p. 11]

[20] J. A. Russell, "Core affect and the psychological construction of emotion.," *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003. [cited at p. 11, 12, and 97]

[21] J. Russell and J. Snodgrass, "Emotion and the environment," *Handbook of environmental psychology*, vol. 1, pp. 245–280, 1987. [cited at p. 12]

[22] S. Marković, "Components of aesthetic experience: aesthetic fascination, aesthetic appraisal, and aesthetic emotion," *i-Perception*, vol. 3, no. 1, pp. 1–17, 2012. [cited at p. 13]

[23] L.-H. Hsu, *Visible and Expression. Study on the Intersubjective Relation between Visual Perception, Aesthetic Feeling, and Pictorial Form*. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS), June 2009. [cited at p. 13]

[24] A. R. Hochschild, "Emotion work, feeling rules, and social structure," *American Journal of Sociology*, vol. 85, no. 3, 1979. [cited at p. 14]

[25] N. Wiley, "Emotion and film theory," *Studies in Symbolic Interaction*, vol. 26, pp. 169–190, 2003. [cited at p. 14]

[26] W. Wirth and H. Schramm, "Media and emotions," *Communication research trends*, vol. 24, no. 3, pp. 3–39, 2005. [cited at p. 15]

[27] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theories of emotion*, vol. 1, 1980. [cited at p. 15 and 29]

[28] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977. [cited at p. 16]

[29] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989. [cited at p. 16]

[30] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, *International affective picture system (IAPS): Technical manual and affective ratings*. The Center for Research in Psychophysiology, University of Florida, 1999. [cited at p. 16 and 54]

[31] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, pp. 1050–1057, Dec. 2007. [cited at p. 16]

[32] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pp. 205–211, 2004. [cited at p. 20]

[33] R. W. Picard, "Affective computing: From laughter to ieee," *IEEE Transactions on Affective Computing*, vol. 1, pp. 11–17, Jan 2010. [cited at p. 20]

[34] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 636–647, 2013. [cited at p. 20, 86, and 90]

[35] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Transactions on Multimedia*, vol. 12, pp. 510–522, Oct. 2010. [cited at p. 20, 27, 28, 30, 31, and 86]

[36] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu, "Flexible presentation of videos based on affective content analysis," in *Advances in Multimedia Modeling*, pp. 368–379, 2013. [cited at p. 20]

[37] H. Katti, K. Yadati, M. Kankanhalli, and C. Tat-Seng, "Affective video summarization and story board generation using pupillary dilation and eye gaze," in *2011 IEEE International Symposium on Multimedia (ISM)*, pp. 319–326, Dec 2011. [cited at p. 20]

[38] R. R. Shah, Y. Yu, and R. Zimmermann, "Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings," in *Proceedings of the ACM International Conference on Multimedia*, pp. 607–616, 2014. [cited at p. 20]

[39] K. Yadati, H. Katti, and M. Kankanhalli, "Cavva: Computational affective video-in-video advertising," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 15–23, 2014. [cited at p. 20]

[40] P. Philippot, "Inducing and assessing differentiated emotion-feeling states in the laboratory," *Cognition & Emotion*, vol. 7, no. 2, pp. 171–193, 1993. [cited at p. 20 and 53]

[41] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, 1995. [cited at p. 20 and 38]

[42] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner,

N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction*, vol. 4738, pp. 488–500, 2007. [cited at p. 22]

[43] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, pp. 1153–1172, Nov. 2010. [cited at p. 22 and 58]

[44] J. Rottenberg, R. D. Ray, and J. J. Gross, "Emotion elicitation using films," *Handbook of emotion elicitation and assessment*, p. 9, 2007. [cited at p. 22 and 38]

[45] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: a database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, pp. 18–31, Jan. 2012. [cited at p. 22, 29, 58, and 74]

[46] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, pp. 42–55, Jan. 2012. [cited at p. 22]

[47] S. Carvalho, J. Leite, S. Galdo-Álvarez, and O. Gonçalves, "The emotional movie database (EMDB): a self-report and psychophysiological study," *Applied Psychophysiology and Biofeedback*, vol. 37, no. 4, pp. 279–294, 2012. [cited at p. 22 and 59]

[48] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "A benchmarking campaign for the multimodal detection of violent scenes in movies," in *Proceedings of the 12th International Conference on Computer Vision*, ECCV'12, pp. 416–425, 2012. [cited at p. 23]

[49] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated GIFs," in *Proceedings of the ACM International Conference on Multimedia*, MM '14, pp. 213–216, 2014. [cited at p. 23]

[50] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, pp. 143–154, Feb. 2005. [cited at p. 23, 24, 28, 38, 54, 101, and 112]

[51] M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Affective Computing and Intelligent Interaction*, pp. 1–7, Sept. 2009. [cited at p. 23, 24, and 30]

[52] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2376–2379, May 2011. [cited at p. 24, 25, 30, 44, 46, 54, 56, 101, 102, and 111]

[53] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, pp. 92–105, April 2011. [cited at p. 24, 25, 26, 30, 31, 69, 94, and 105]

[54] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pp. 543–550, 2013. [cited at p. 24, 26, 30, and 90]

[55] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001. [cited at p. 25]

[56] H.-B. Kang, "Affective content detection using HMMs," in *Proceedings of the eleventh ACM international conference on Multimedia*, MULTIMEDIA '03, pp. 259–262, 2003. [cited at p. 26, 27, 28, and 30]

[57] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, pp. 689–704, June 2006. [cited at p. 27, 28, 30, 31, and 87]

[58] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *Affective Computing and Intelligent Interaction*, vol. 4738, pp. 594–605, 2007. [cited at p. 27, 28, and 30]

[59] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pp. 677–680, 2008. [cited at p. 27, 28, 30, 101, and 102]

[60] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *IEEE International Symposium on Multimedia*, pp. 228–235, Dec. 2008. [cited at p. 27, 28, 29, 74, and 103]

[61] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Transactions on Multimedia*, vol. 12, pp. 523–535, Oct. 2010. [cited at p. 27 and 29]

[62] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *MultiMedia Modeling*, vol. 8325 of *Lecture Notes in Computer Science*, pp. 303–314, 2014. [cited at p. 27, 29, 30, and 31]

[63] C. Penet, C. Demarty, G. Gravier, and P. Gros, "Multimodal information fusion and temporal integration for violence detection in movies," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2393–2396, March 2012. [cited at p. 27 and 29]

[64] F. Eyben, F. Weninger, N. Lehment, B. Schuller, and G. Rigoll, "Affective video retrieval: Violence detection in hollywood movies by

large-scale segmental feature extraction," *PloS one*, vol. 8, no. 12, 2013. [cited at p. 27 and 30]

[65] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *INTERSPEECH*, vol. 2008, pp. 597–600, 2008. [cited at p. 30]

[66] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012. Best of Automatic Face and Gesture Recognition 2011. [cited at p. 30]

[67] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Affective Computing and Intelligent Interaction (ACII)*, 2015. [cited at p. 30]

[68] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005. [cited at p. 31]

[69] M. Horvat, S. Popovic, and K. Cosic, "Multimedia stimuli databases usage patterns: a survey report," in *Proceedings of the 36nd International ICT Convention MIPRO*, pp. 993–997, 2013. [cited at p. 31]

[70] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, vol. 16, pp. 1075–1089, June 2014. [cited at p. 31 and 56]

[71] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *IEEE International Conference on Multimedia and Expo*, pp. 566–569, 2009. [cited at p. 31]

[72] R. W. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Electronic Imaging*, pp. 290–301, 1998. [cited at p. 38]

[73] H. Leventhal and K. Scherer, "The relationship of emotion to cognition: A functional approach to a semantic controversy," *Cognition & Emotion*, vol. 1, no. 1, pp. 3–28, 1987. [cited at p. 38]

[74] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, Apr. 2013. [cited at p. 38, 39, 68, and 69]

[75] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE Multimedia*, vol. 6, pp. 38–53, Sept. 1999. [cited at p. 38]

[76] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. [cited at p. 39]

[77] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus,"

in *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010*, July 2010. [cited at p. 39]

[78] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, pp. 436–465, Aug. 2013. [cited at p. 39, 46, and 56]

[79] Y.-H. Yang and H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011. [cited at p. 39]

[80] P. A. Russell and C. D. Gray, "Ranking or rating? some data and their implications for the measurement of evaluative response," *British Journal of Psychology*, vol. 85, pp. 79–92, Feb. 1994. [cited at p. 39]

[81] S. Ovadia, "Ratings and rankings: reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004. [cited at p. 39]

[82] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: A comparative study of self-reporting," in *Affective Computing and Intelligent Interaction*, vol. 6974, pp. 437–446, 2011. [cited at p. 39]

[83] R. Mantiuk, A. M. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment.," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012. [cited at p. 40]

[84] J. D. Smyth, D. A. Dillman, L. M. Christian, and M. J. Stern, "Comparing check-all and forced-choice question formats in web surveys," *Public Opinion Quarterly*, vol. 70, no. 1, pp. 66–77, 2006. [cited at p. 40]

[85] C. A. R. Hoare, "Algorithm 64: Quicksort," *Communications of the ACM*, vol. 4, p. 321, July 1961. [cited at p. 40]

[86] S. S. Skiena, *The algorithm design manual*. London: Springer, 2nd ed., 2008. [cited at p. 40]

[87] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, pp. 49–59, Mar. 1994. [cited at p. 41, 54, and 66]

[88] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971. [cited at p. 44 and 129]

[89] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, pp. 61–70, Apr. 1970. [cited at p. 44, 52, and 130]

[90] J. J. Randolph, "Free-marginal multirater kappa (multirater κ_{free}): An alternative to fleiss' fixed-marginal multirater kappa." Paper presented at the *Joensuu University Learning and Instruction Symposium*, Oct. 2005. [cited at p. 44, 46, 69, and 130]

[91] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, Mar. 1977. [cited at p. 46 and 69]

[92] J. A. Sloboda, "Empirical studies of emotional response to music.," in *Cognitive bases of musical communication.*, pp. 33–46, American Psychological Association, 1992. [cited at p. 48]

[93] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement.," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008. [cited at p. 48]

[94] K. R. Scherer and M. R. Zentner, "Emotional effects of music: Production rules.," *Music and emotion: Theory and research*, pp. 361–392, 2001. [cited at p. 48]

[95] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, pp. 261–273, May 1993. [cited at p. 54, 59, and 74]

[96] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is stronger than good.," *Review of General Psychology*, vol. 5, no. 4, pp. 323–370, 2001. [cited at p. 57]

[97] G. Peeters and J. Czapinski, "Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects," *European Review of Social Psychology*, vol. 1, pp. 33–60, Jan. 1990. [cited at p. 57]

[98] B. L. Fredrickson, "What good are positive emotions?," *Review of General Psychology*, vol. 2, no. 3, pp. 300–319, 1998. [cited at p. 59]

[99] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Adaptive computation and machine learning, Cambridge, Mass: MIT Press, 2006. [cited at p. 59]

[100] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, pp. 871–880, Dec. 1984. [cited at p. 60]

[101] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [cited at p. 60]

[102] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 709–710, Sept 2013. [cited at p. 66 and 118]

[103] K. R. Scherer, "Appraisal considered as a process of multi-level sequential checking," in *Appraisal processes in emotion: Theory, Methods, Research* (K. R. Scherer, A. Schorr, and T. Johnstone, eds.), pp. 92–120, Oxford University Press, 2001. [cited at p. 66]

[104] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 85–90, Sept 2013. [cited at p. 69]

[105] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of*

the 2nd ACM international workshop on Crowdsourcing for multimedia, pp. 1–6, ACM, 2013. [cited at p. 69]

[106] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos," in *Brain informatics*, pp. 89–100, 2010. [cited at p. 73]

[107] F. Zhou, X. Qu, J. R. Jiao, and M. G. Helander, "Emotion prediction from physiological signals: A comparison study between visual and auditory elicitors," *Interacting with computers*, vol. 26, no. 3, pp. 285–302, 2014. [cited at p. 74]

[108] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 73–78, 2013. [cited at p. 74 and 75]

[109] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pp. 2662–2667, IEEE, 2007. [cited at p. 75]

[110] J. Fleureau, P. Guillotel, and Q. Huynh-Thu, "Physiological-based affect event detector for entertainment video applications," *IEEE Transactions on Affective Computing*, vol. 3, pp. 379–385, July 2012. [cited at p. 75]

[111] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 381–385, 2005. [cited at p. 76]

[112] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, *et al.*, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997. [cited at p. 86]

[113] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *IEEE International Conference on Image Processing (ICIP)*, pp. 1853–1856, Nov. 2009. [cited at p. 86 and 90]

[114] A. Smola and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997. [cited at p. 86]

[115] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004. [cited at p. 86]

[116] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 710–732, July 1992. [cited at p. 86]

[117] D. Hasler and S. Suesstrunk, "Measuring colourfulness in natural images," in *Proc. SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, pp. 87–95, 2003. [cited at p. 87]

[118] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, vol. 1, pp. 419–426, 2006. [cited at p. 87]

[119] O. Le Meur, T. Baccino, and A. Roumy, "Prediction of the inter-observer visual congruency (IOVC) and application to image ranking," in *Proceedings of the 19th ACM International Conference on Multimedia*, pp. 373–382, 2011. [cited at p. 87]

[120] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proceedings of the 10th International Conference on Computer Vision*, vol. 5304, pp. 386–399, 2008. [cited at p. 87 and 88]

[121] Y. Baveye, F. Urban, and C. Chamaret, "Image and video saliency models improvement by blur identification," in *Computer Vision and Graphics*, vol. 7594, pp. 280–287, 2012. [cited at p. 87]

[122] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 624–630, 2006. [cited at p. 88]

[123] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, pp. 1664–1078, 2013. [cited at p. 90]

[124] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," in *16th European Signal Processing Conference, Lausanne, Switzerland*, 2008. [cited at p. 90]

[125] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989. [cited at p. 90]

[126] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5, pp. 555–559, 2003. [cited at p. 90]

[127] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014. [cited at p. 92]

[128] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [cited at p. 94 and 99]

[129] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732, June 2014. [cited at p. 98]

[130] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 802–817, May 2006. [cited at p. 99]

[131] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the ACM International Conference on Multimedia*, pp. 801–804, 2014. [cited at p. 102]

[132] J. Schlüter and S. Böck, "Musical onset detection with convolutional neural networks," in *6th International Workshop on Machine Learning and Music (MML)*, 2013. [cited at p. 102]

[133] J. Allen, "Short-term spectral analysis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977. [cited at p. 103]

[134] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, 2014. [cited at p. 103 and 105]

[135] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning." Book in preparation for MIT Press, 2015. [cited at p. 104]

[136] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [cited at p. 105]

[137] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 933–936, ACM, 2011. [cited at p. 105]

[138] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using eeg signals and facial expressions," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2014. [cited at p. 105]

[139] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5412–5416, 2014. [cited at p. 105]

[140] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013. [cited at p. 105]

[141] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005. [cited at p. 106]

[142] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. [cited at p. 106]

[143] K. Delac, M. Grgic, and S. Grgic, "Independent comparative study of pca, ica, and lda on the feret data set," *International Journal of Imaging Systems and Technology*, vol. 15, no. 5, p. 252, 2005. [cited at p. 109]

[144] A. Bouzalmat, J. Kharroubi, and A. Zarghili, "Comparative study of pca, ica, lda using svm classifier," *Journal of Emerging Technologies in Web Intelligence*, vol. 6, no. 1, pp. 64–68, 2014. [cited at p. 109]

[145] A. M. Martínez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001. [cited at p. 109]

[146] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990. [cited at p. 110]

[147] F. Weninger, "Introducing currennt: The munich open-source cuda recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015. [cited at p. 110]

[148] M. Nicolaou, V. Pavlovic, M. Pantic, *et al.*, "Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014. [cited at p. 119]

[149] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015. [cited at p. 119]

[150] T. Charlesworth, H. Ford, L. Milton, T. Mortensson, J. Pedlingham, J. Knibbe, and S. A. Seah, "Telltale: Adding a polygraph to everyday life," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1693–1698, 2015. [cited at p. 119]

[151] E. Ščiglinskas and A. Vidugirienė, "Investigation on human attentiveness to video clips using neurosky and liris-accede database," in *Augmented and Virtual Reality*, pp. 450–456, Springer, 2014. [cited at p. 119]

[152] M. Soleymani, S. Asghari Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, 2015. [cited at p. 120]

[153] K. R. Scherer, "Emotions are emergent processes: they require a dynamic computational architecture," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3459–3474, 2009. [cited at p. 120]

[154] L. Ding and A. Yilmaz, "Learning relations among movie characters: A social network perspective," in *Computer Vision–ECCV 2010*, pp. 410–423, 2010. [cited at p. 120]