Recognizing Induced Emotions of Movie Audiences from Multimodal Information

Michal Muszynski, Leimin Tian, Catherine Lai, Johanna D. Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, Guillaume Chanel

Abstract—Recognizing emotional reactions of movie audiences to affective movie content is a challenging task in affective computing. Previous research on induced emotion recognition has mainly focused on using audio-visual movie content. Nevertheless, the relationship between the perceptions of the affective movie content (perceived emotions) and the emotions evoked in the audiences (induced emotions) is unexplored. In this work, we studied the relationship between perceived and induced emotions of movie audiences. Moreover, we investigated multimodal modelling approaches to predict movie induced emotions from movie content based features, as well as physiological and behavioral reactions of movie audiences. To carry out analysis of induced and perceived emotions, we first extended an existing database for movie affect analysis by annotating perceived emotions in a crowd-sourced manner. We find that perceived and induced emotions are not always consistent with each other. In addition, we show that perceived emotions, movie dialogues, and aesthetic highlights are discriminative for movie induced emotion recognition besides spectators' physiological and behavioral reactions. Also, our experiments revealed that induced emotion recognition could benefit from including temporal information and performing multimodal fusion. Moreover, our work deeply investigated the gap between affective content analysis and induced emotion recognition by gaining insight into the relationships between aesthetic highlights, induced emotions, and perceived emotions.

Index Terms—Affective Computing, Implicit Tagging, Emotion Recognition, Multimodal Learning, Multimodal Fusion, Induced and Perceived Emotions, Aesthetic Highlights, Physiological and Behavioral Signals, Crowdsourcing

1 Introduction

MUCH attention has recently been drawn to recognizing emotions induced in movie audiences by affective content due to potential applications, such as emotion-based content delivery [1], video indexing and summarization [2] as well as movie scene design. Nevertheless, recognizing emotions induced by affective movie content remains a challenging task because only weak or moderate correlations between automatic predictions and human annotations have been achieved [3]. There are three most widely used models for defining emotions in current affective content analysis, such as the basic emotion model [4], the appraisal model [5], and the circumplex model [6]. The circumplex emotion model is able to describe compound or subtle emotions and is widely used in annotating movie induced emotions in state-of-the-art studies on affective content analysis [3].

When stimuli is selected to induce emotions, it is assumed that emotions conveyed by the affective content (**perceived emotions** of the stimuli) are always consistent with emotions evoked in the spectators (**induced emotions**) [7]. Moreover, perceived and induced emotions are not usually considered separately in studies on affective content. However, some research on music emotions has attempted to investigate the differences between the perceived emotions of affective content and the induced emotions of music audiences. Moreover, research on music emotions has discovered that emotions perceived from music are not always consistent with the emotions elicited in music listeners [8]. This suggests that distinguishing perceived and induced emotions of movie audiences can be crucial to make significant progress in affective content analysis.

We can distinguish three perspectives on movie emotions,

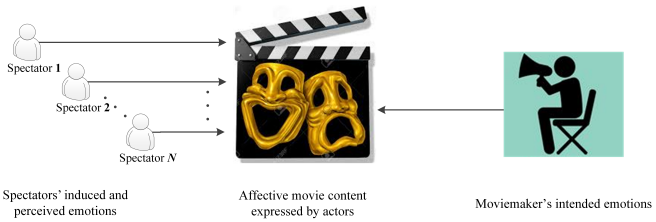

Fig. 1: The three perspectives on movie emotions.

namely the audiences' perspective, the actors' perspective, and the directors' perspective, as shown in Figure 1. Movie audiences perceive and interpret the movie content and emotions expressed by actors playing main characters (perceived emotions). This induces emotional responses in movie audiences (induced emotions). Also, movie directors create scripts with emotional annotations that include their expectations of which emotions should be induced in movie audiences by a particular scene (intended emotions). In this paper we only investigated the audiences' perspective on

Michal Muszynski is with University of Geneva
Leimin Tian is with University of Edinburgh and Monash University
Catherine Lai and Johanna D. Moore are with University of Edinburgh
Theodoros Kostoulas is with University of Geneva and Bournemouth University
Patrizia Lombardo, Thierry Pun, and Guillaume Chanel are with University of Geneva
(e-mail: michal.muszynski@unige.ch; leimin.tian@monash.edu; clai@inf.ed.ac.uk; J.Moore@ed.ac.uk; tkostoulas@bournemouth.ac.uk; patrizia.lombardo@unige.ch; thierry.pun@unige.ch; guillaume.chanel@unige.ch)

movie emotions. Tan [9] argued that the emotions perceived from movies can influence the induced emotional responses of audiences by evoking empathy. Consequently, this could imply a positive correlation between perceived and induced emotions. Nevertheless, Baveye et. al [10] argued that emotions intended by the directors might not always be consistent with emotions that are induced in movie audiences. In fact, the authors did not consider perceived emotions in their studies. We are the first to formally investigate the relationship between perceived and induced emotions of movie audiences. Moreover, we attempt to bridge this gap by carrying out a statistical analysis on emotions perceived from movie content and emotions induced in movie audiences. This work also provides us a fundamental understanding of how affective movie content induces emotions in audiences. Moreover, we could reveal how we use multimodal information on movie content to predict them.

The state of the art research on induced emotion recognition has mainly focused on extracting audio-visual features from video recordings. In [11], the authors used human nonverbal behavior signals, including facial expression, shoulder gesture and audio cues to predict spontaneous affect. It was shown that Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM) outperformed Support Vector Regression (SVR) due to their ability to learn temporal dependencies between multimodal features and emotional scores. In addition, other researchers proposed to use Deep Belief Networks (DBNs) to capture complex non-linear interactions in audio-visual features for emotion recognition [12]. Beyond audio-visual features, movie dialogues have been shown to be effective for violence recognition in movies [13]. Thus, these affective cues of perceived emotions in movies can be used for the recognition of induced emotions.

To investigate the complex relationship between induced and perceived emotions of movie audiences, we selected the Continuous LIRIS-ACCEDE database [3], [14] that contains continuous arousal-valence annotations of emotions induced in movie audiences as well as spectators' physiological and behavioral reactions to movie content. This database has been commonly used in studies on movie induced emotion recognition, including benchmark challenges, e.g., MediaEval2016 [15]. In order to study the relationship between perceived and induced emotions of movie audiences and improve induced emotion recognition, we collected manual transcripts of 8 selected movies from the Continuous LIRIS-ACCEDE database [3], [14]. This includes manual annotations of DISfluency and Non-verbal Vocalisations (DIS-NV) [16] in dialogues to describe lexical movie content. In addition to the extraction of audio-visual features from movie content [17], we collected crowd-sourced annotations on perceived arousal, valence, and power of the movie dialogues to better characterize affective movie content. Also, we used aesthetic highlight annotations to describe the aesthetic part of movie content [18]. Moreover, we extracted statistical features of physiological and behavioral signals to capture spectators' reactions while watching movies. Besides investigating discriminative power of multimodal features, we studied the impact of including temporal information on the recognition performance as well as different fusion strategies for combining multimodal information (i.e., audio, visual, lexical movie content, perceived emotion and aesthetic highlight annotations as well as spectators' physiological and behavioral reactions). This paper is the extension of our previous work [19], addressing the following research questions:

- Are perceived emotions and induced emotions always consistent?
- How can we improve recognition performance of induced emotions?
  - Are there other features beyond the audio-visual movie content that can contribute to induced emotion recognition?
  - Are perceived emotions discriminative for induced emotion recognition?
  - Are fusion of movie content features and spectators' reactions and temporal information on movie content and spectators' reactions beneficial for emotion recognition?

The contribution of the work is below, highlighting the novelty compared to our previous work [19]:

- We provide insights into how movie genres differ in terms of emotions of the audiences and characteristics of the movie dialogues. To do so, we investigate the influence of movie genre on the intensity of movie audiences' perceived emotions, and the amount of DISfluency and Non-verbal Vocalisation (DIS-NV) in movie dialogues.
- We show that discrepancies between perceived emotion annotations are larger than discrepancies between induced emotion annotations for all movies and movie genres.
- We establish the relationship between aesthetic movie content and emotions of movie audiences. In particular, we identify aesthetic highlights as novel high level aesthetic cues that carry information on perceived and induced emotions regardless of the discrepancies between them.
- We propose novel multimodal models for predicting movie induced emotions. These models incorporate perceived emotion annotations in the hierarchical architecture of Long Short-Term Memory Recurrent Neural Networks (LSTM) with movie content features, as well as movie content features and movie audience reactions. We show that recognition of induced emotions benefits from multimodal hierarchical fusion of movie content features and spectators' reactions, and taking into account temporal information when comparing to baseline emotion recognition models, namely, Deep Belief Networks (DBNs) and Support Vector Regression (SVR).

The rest of this paper is organized as follows: Section 2 reviews current affective content analysis studies. Section 3 consists of descriptions of data collection protocols to extend emotional annotations from the Continuous LIRIS-ACCEDE database, as well as basic statistics of new collected annotations. Section 4 corresponds to descriptions of mutlimodal feature extraction. Section 5 provides the descriptions of emotion recognition models. Section 6 consists of an analysis of the relationship between perceived

emotions and induced emotions of movie audiences as well as occurrences of aesthetic highlights in movies. Moreover, Section 6 presents results of unimodal and multimodal emotion recognition. Section 7 discusses the results that are obtained. Section 8 provides conclusions and future directions of our research.

2 Related Work

In Section 2.1 we introduce different approaches to emotion recognition from multimodal signals. In Sections 2.2 and 2.3, we review state-of-the-art work on affective content analysis. Firstly, we discuss previous work on the relationship between perceived and induced emotions to reveal its weaknesses. Secondly, we detail previous studies on induced emotion recognition on the Continuous LIRIS-ACCEDE database and identify their limitations.

2.1 Emotion recognition

The majority of current audio-visual emotion recognition studies have focused on identifying best machine learning models to recognize continuous emotions represented in an arousal-valence space. In [11], [20], [21] the authors applied Support Vector Regression, Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM), and Relevance Vectors Machines to audio-visual features to recognize continuous emotions over time. In addition, the BLSTM models were used to recognize continuous emotions from speech by modelling termporal context of emotions [22]. Also, a reconstruction error based learning framework was introduced to recognize continuous emotions from speech using autoencoders [23]. Other researchers recently introduced an automatic affect sensing system trained on physiological signals in an end-to-end manner [24]. Furthermore, a transfer learning framework has been applied to audio-visual features for basic emotion recognition due to a lack of enough training instances [25].
Also, multi-task and feature learning were used to improved automatic emotion recognition of interacting dyads [26]. Deep Belief Networks were applied to learn a new representation of audio-visual features while multi-task learning was proposed to jointly model recognition of induced and perceived emotions. In addition, multi-clue fusion at the decision level was proposed to model human emotion in the wild from facial appearance texture, facial action, and audio [27]. High level features were extracted from sequential face images by means of recurrent neural networks combined with deep Convolutional Neural Networks (CNNs). The models were previously pretrained on face images. To capture facial actions, a facial landmark trajectory model was proposed based on CNNs and a Support Vector Machine. Also, low-level energy features were extracted from video segments to feed other CNNs. Moreover, many studies have attempted to establish a relationship between information included in stimuli, such as movies, clips, videos, and the affective state of a spectator. Recent work has focused on humans' physiological responses to affective movie content [28], since induced emotions are associated with a subjective sphere of emotions and preferences [29]. Physiological reactions are considered to be related to emotional states of a spectator elicited by movie content [14]. Furthermore, researchers have recently investigated emotion recognition in responses to multimedia content based on electroencephalography and peripheral physiological signals as well as facial expressions [30]. In [14], a weighted mean galvanic skin response profile among movie spectators was proposed. The affective profile of people who watched movies was made by using a single modality, such as electrodermal activity [31] or facial expression of viewers [32]. Also, facial features were extracted to detect spontaneous emotions of viewers who were exposed to movie clips [33]. In addition, observers' physiological signals, such as galvanic skin response and heart rate variability were used to classify the depression levels of multiple people in videos in which the observers did not understand the spoken language [34].
The relationship between information conveyed by the stimuli and spectators' emotional responses has been investigated by the affective computing and mutlimedia community for more than one decade. However, it still remains a challenging task because the performance achieved for predicting induced emotions is still limited [10], [11], [20], [21]. Furthermore, there is a lack of studies on fusion models that can aggregate video content with spectators' physiological and behavioral reactions.

2.2 Perceived vs. Induced Emotions

Listening to music or watching movies can be an emtional experience. Furthermore, we perceive emotions conveyed by the affective content. Characteristics of the stimuli, such as tempo and pitch of music pass information to audiences [8]. On the contrary to perceive emotions, induced emotions are evoked in spectators by the stimuli and are associated with personal experience and individual preferences [29]. For instance, a song that is perceived as happy can induce stronger depression in depressed subjects [8]. Also, previous work on emotions suggests that perceived emotions are more objective than induced emotions [35]. Futhermore, annotators usually have stronger agreement on perceived emotions rather than induced emotions [36]. Previous studies on affective content analysis have not always distinguished the differences between perceived and induced emotions. Even though consistencies between perceived and induced emotions have been shown in some studies [37], reseach on music emotions has discovered fundamental differences between perceived and induced emotions [8]. Previous work on emotions has also suggested that induced emotions could have more intensive arousal and less intensive valence ratings in comparison to perceived emotion ratings of the same stimuli [38]. Furthermore, studies on emotion expression and perception during spoken dyadic interactions proposed a clear distinction between induced and perceived emotions, and revealed complex dependencies between them during dialogues acted by actors [26].
In fact, there has only been limited work [39] that invesitages the relationship between perceived and induced emotions of movie audiences in comparison to studies music emotions. It is important to mention that Hanjalic and Xu [7] assumed positive correlations between perceived and induced emotions of movie audiences. That is why the authors used descriptors of affective content to predict a spectator's

emotional reactions. Also, Benini et al. [40] observed that the annotator agreement on movie emotion descriptions was stronger when movie video features were included in the emotion definition. This could also imply a relationship between movie content and induced emotions. In this work we make the first attempt to study to what extent perceived and induced emotions of movie audiences are consistent.

2.3 Previous Work on LIRIS-ACCEDE Database

As presented in Table 1, previous work on the Continuous LIRIS-ACCEDE database recognized movie induced emotions by means of various regression models, such as Support Vector Regression (SVR) [41], Long Short-Term Memory Recurrent Neural Networks (LSTM) [42], and Convolutional Neural Networks (CNNs) [43]. However, all the models were fed by audio or/and visual features of movie content without taking account spectators' reactions.

The Pearson Correlation Coefficient (CC) is the most commonly reported evaluation metric. Also, the Mean Squared Error (MSE) is often reported [41]. Only weak or moderate correlations have been achieved in state-of-the-art studies on induced emotion recognition based on audio-visual features[1]. This suggests that recognizing induced emotions of movie audiences is a challenging task. It is important to point out that different studies have different experiment settings, e.g., data pre-processing or training-testing set partitions. As a result, their results are not directly comparable. Previous studies have mainly focused on using audio-visual features extracted from movie content [44], [45]. Nevertheless, lexical information from the movie dialogues has largely been overlooked, even though it has been shown that they were important for emotion recognition [49]. Besides movie dialogues, the usefulness of knowledge-inspired affective cues, for example, aesthetic highlights [50] have not been explored for predicting movie induced emotions [14].

TABLE 1: The Continuous LIRIS-ACCEDE database: the state of the art performance in induced emotion recognition.

Model	A-MSE	A-CC	V-MSE	V-CC
AudioVisual SVR [41]	0.326	0.242	0.343	**0.221**
AudioVisual SVR [44]	#	**0.265**	#	0.110
AudioVisual SVR [45]	**0.120**	0.236	**0.099**	0.142
AudioVisual LSTM [46]	0.124	0.054	0.102	0.017
Audio PLS [47]	0.129	-0.072	0.141	-0.062
Visual CNN [3]	**0.021**	0.152	**0.027**	0.197
Visual SVR [3]	0.022	**0.337**	0.034	**0.296**
Visual SVR [48]	0.126	0.056	0.106	0.019

Many previous studies have examined unimodal models for induced emotion recognition [47], [48]. In fact, Baveye et al. [3] used a SVR model with only visual features and achieved the best reported CC (0.337) for induced emotion recognition. Nevertheless, it has been shown that combining multimodal information improved performance for other emotion recognition tasks [51]. That is why we are encouraged to investigate modality fusion strategies that could improve induced emotion recognition. In addition, LSTM models have had low performance for predicting movie induced emotions [46][2]. Nevertheless, the LSTM models have achieved best performance in various emotion recognition tasks because of their ability to take into account temporal information [52]. In particular, Ma et al. [46] predicted movie induced emotions on intervals of 10 seconds. This means that enough temporal information was already provided. However, it is important to mention that the suitable amount of temporal information needed for predicting movie induced emotions remains undefined.

1. The best reported CC for arousal is 0.337, for valence is 0.296 [3]
2. arousal CC is 0.054, valence CC is 0.017 [46]

3 Data set and Annotations

3.1 LIRIS-ACCEDE database

The LIRIS-ACCEDE databases [3] were collected and released to provide researchers resources to work on affective content analysis. In this paper we analyze the Continuous LIRIS-ACCEDE database (C. LIRIS-ACCEDE) [3], [14] that consists of 30 full-length movies. These movies come from 9 movie genres and their total duration is 442 minutes. During annotation collection, these movies were grouped into four sets according to their duration. Each of 10 participants watched selected movies from two sets once and then annotated continuous arousal and valence ratings (value range [-1,1]) of the emotions they felt during watching (induced emotions). Then, the means of these ratings provided by the participants over each second of the movie were used as the gold-standard. A follow-up study displayed these 30 movies to another 13 participants with sensors attached to their limbs. The galvanic skin responses and acceleration signals of these 13 participants were collected during the movie projections.

3.2 Extended Annotations of LIRIS-ACCEDE

We describe below collection of the extended annotations [19] of the C. LIRIS-ACCEDE database with their detailed statistics. These include transcripts of movie dialogues with word timings and affective cue labels in Section 3.3, perceived emotion annotations in Section 3.4, and an analysis of agreement on perceived and induced emotion annotations in Section 3.5. We chose 8 English movies listed in Table 2 which contain significantly more dialogues than the other movies from the C. LIRIS-ACCEDE database, e.g., the movies *Sintel* and *Chatter* [3], [41]. Moreover, these movies come from different movie genres and are in the double-reality art form, where the lead characters exist between two worlds. This is similar to the activity of movie watching in which the real and movie world together create double-reality experience for the movie audiences. For this reason,

TABLE 2: Statistics of selected C. LIRIS-ACCEDE movies.

Movie	Genre	Utterance count	Mean sent. duration (s)
After the Rain (M1)	Drama	77	3.000
First Bite (M2)	Romance	54	2.056
Nuclear Family (M3)	Comedy	147	2.694
Payload (M4)	Adventure	121	2.488
Spaceman (M5)	Adventure	133	2.489
Superhero (M6)	Drama	161	2.832
Tears of Steel (M7)	Adventure	79	2.165
The Secret Number (M8)	Drama	98	2.724

3. http://liris-accede.ec-lyon.fr/database.php

the audiences can empathize more with main movie characters. These are particularly interesting for understanding perceived and induced emotions due to spectators' strong engagement with movies. To sum up, we annotated 118 minutes of movies containing 870 utterances in total.

3.3 Transcription and Affective Cue Annotation

TABLE 3: The Pearson Correlation Coefficient (CC) and the Concordance Correlation Coefficient (CCC) between start and end timings of utterances, words as well as DIS-NV annotations are calculated.

Labels	Start (CC)	End (CC)	Start (CCC)	End (CCC)
Utterance	0.998	0.998	0.997	0.998
Word	0.999	0.999	0.999	0.999
General lexicon	0.989	0.989	0.988	0.988
Filled pause	0.625	0.625	0.560	0.549
Filler	0.920	0.920	0.744	0.744
Stutter	0.916	0.916	0.835	0.836
Laughter	0.635	0.635	0.369	0.369
Audible breath	0.766	0.764	0.620	0.637

The movie transcription and affective cue annotations were collected from two expert annotators. To increase the annotation speed, we first applied the IBM Watson Speech to Text service[4] to audio recordings of movies. This service could provide us automatic speech transcription with word timings. Then, these auto-generated transcripts were manually corrected and annotated by two annotators working in parallel. Each of them annotated five movies. To evaluate the quality of annotations based on annotation agreement, movies *First Bite* and *Spaceman* were annotated by both annotators and then we computed the Normalized Damerau-Levenshtein (NDL) distances [53] of their transcripts, as well as the Pearson Correlation Coefficient (CC) and the Concordance Correlation Coefficient (CCC) of the word timings.

The NDL distance is a common measure of the distance between two strings. It is defined as the minimum number of operations that are required to transform one string to the other. Then it is divided by the length of the longer string of the pair to be normalized. The NDL distance of 0 reveals that the two strings are identical. Thus, values closer to 0 corresponds to strong annotation agreement. We find that 94.8% of the transcribed words are the same for both annotators and the average NDL distance of 0.049. If we suppose that the average length of words is 4 characters in the transcript, an average NDL distance of 0.049 indicates that there is less than one character difference for every five words. In addition, CC and CCC values for the word and utterance timings of the transcript are presented in Table 3. We can see that the utterance and word timings annotated by these two annotators are strongly correlated. Overall, this indicates that the two annotators strongly agreed on movie transcription annotations.

In the follow-up part of our study, the same annotators also annotated the following categories of DISfluency and Non-verbal Vocalisation (DIS-NV) in movie dialogues: filled pauses (e.g., "eh" or "hmm"), fillers (verbal filled pauses), stutters, laughter, and audible breath. Furthermore, remaining words were labelled as general lexicons. The DIS-NVs have been shown to be indicative of speaker emotions in spontaneous dialogues [16]. To evaluate annotation agreement, we divided the annotations into six subsets based on the DIS-NV labels and calculated the CC and CCC of the word timings for each category of DIS-NV labels. Even though the annotation agreement on DIS-NV label timing is lower in comparison to movie transcription, the annotations remain strongly correlated, as presented in Table 3. This suggests that annotating DIS-NV labels is a more subjective task due to some ambiguity made by environmental noise and playing music in the background.

In Tables 4 and 5, we listed the statistics of each category of DIS-NV in each movie and per movie genre. As shown in Table 4, in total there are more disfluencies than non-verbal vocalisations in the movies, and filler is the most common category of DIS-NV. As shown in Table 5, there are fundamental differences in terms of DIS-NV occurrences in different movie genres. Romance movies least contain DIS-NVs, while adventure movies most consist of DIS-NVs, as shown in Table 5. It is worth pointing out that DIS-NVs are indicators of speaker uncertainty. Our observation is that adventure movies have more DIS-NVs than the other movie genres. This may indicate that adventure movies have higher level of uncertainty in the movie dialogues and story development.

TABLE 4: Amount of DIS-NV annotations in movie dialogues.

Movie	General lexicon	Filled pause	Filler	Stutter	Laughter	Audible breath
M1	532	0	9	0	0	0
M2	185	6	2	0	2	0
M3	748	18	0	18	0	5
M4	712	1	15	2	1	3
M5	686	4	5	13	5	5
M6	910	9	16	0	2	8
M7	273	0	6	18	0	7
M8	549	1	12	0	0	3

TABLE 5: Amount of DIS-NV annotations in movie dialogues per movie genre.

Genre	General lexicon	Filled pause	Filler	Stutter	Laughter	Audible breath
Drama	1991	10	37	0	2	11
Roman.	185	6	2	0	2	0
Comed.	748	18	0	18	0	5
Advent.	1671	5	26	33	6	15

3.4 Annotating Perceived Movie Emotions

Emotion annotation is more subjective in comparison to the movie transcription task. Thus, multiple annotators are desired to do the task. Previous work has recommended having more than 6 annotators to achieve reliable emotion annotations [54]. However, the recent development of crowdsourcing tools allows us to have easy access to larger numbers of annotators. To collect a massive amount of annotations efficiently and inexpensively, we annotated

4. https://www.ibm.com/watson/developercloud/speech-to-text.html

perceived emotions of movie audience by means of Amazon Mechanical Turk [5] that is a crowd-sourced online annotation platform. We segmented movies into utterance excerpts using manual transcriptions of utterance timings. Then, we collected at least 10 annotations from different annotators for each excerpt. In our work, we assume that annotators can correctly understand and perceive affective movie content. Mechanical Turk annotators were first instructed to rate emotions expressed by movie characters in arousal, power, and valence dimensions by means of 1 to 9 integer scales. In addition, we provided Mechanical Turk annotators the explanations of each emotion dimension with meaning of the different scores. Each Human Intelligence Task (HIT) contained movie excerpts of 5 continuous utterances from the same movie in their original order to preserve movie context. Each utterance was shown at each of the five video windows in different HITs to reduce cognitive bias. The HITs were launched at random and we tracked annotators of each movie to prevent an utterance from being annotated more than one time by the same annotator. Annotators were only allowed to annotate a movie excerpt after display it. Also, annotators could only submit their ratings after annotating all movie excerpts. To sum up, we published 1809 HITs in total and we collected the annotations of perceived emotions from 129 annotators with various cultural and educational backgrounds [19]. The 1 to 9 scores of the crowd-sourced annotations were normalized to interval [−1,1] to be consistent with the induced emotion annotations from the C. LIRIS-ACCEDE database. We then calculated the means of the arousal, valence, and power annotations collected on each utterance of the movie dialogues. This provided us the perceived emotion annotations of movie audiences at the utterance-level.

TABLE 6: Mean level of movie audience's perceived emotions per movie.

Movie	Arousal	Valence	Power
M1	-0.149±0.142	-0.238±0.155	-0.118±0.138
M2	0.003±0.212	-0.043±0.204	0.055±0.175
M3	0.106±0.301	-0.037±0.385	0.117±0.251
M4	0.073±0.213	-0.045±0.257	0.121±0.210
M5	0.127±0.198	**0.115±0.265**	0.122±0.148
M6	0.127±0.212	0.032±0.254	0.088±0.229
M7	**0.238±0.232**	-0.063±0.228	**0.202±0.183**
M8	0.067±0.199	-0.054±0.160	0.127±0.137

TABLE 7: Mean level of movie audience's perceived emotions per movie genre.

Genre	Arousal	Valence	Power
Drama	0.046±0.222	-0.055±0.235	0.052±0.209
Roman.	0.003±0.212	-0.043±0.204	0.055±0.175
Comed.	0.106±0.301	-0.037±0.385	0.117±0.251
Advent.	**0.134±0.221**	**0.014±0.266**	**0.141±0.183**

In Tables 6 and 7, we report statistics of the perceived emotion annotations for each movie and per movie genre. As shown in Table 6, even though the average perceived emotions vary from one movie to another, the variances are in the same order of magnitude. There are also some movies that are close to the neutral state (value 0) with regard to average perceived emotions, for example, the movie *First Bite*. It means that the movies contain a balanced number of scenes with various emotional tones.

As shown in Table 7, on average adventure movies have higher arousal, valence, and power than the other movie genres. This means that one type of movie events dominates the content of this movie genre. Moreover, the observation that romances are the closest to the neutral state in terms of arousal suggests that there is a balance between the amount of exciting and relaxing scenes in these movies. Besides, comedies include movie scenes with the highest emotional discrepancies between one another.

3.5 Agreement on Perceived and Induced Emotion Annotation

In this section we investigated differences between induced and perceived emotion annotations. Tables 8 and 9 contain the average standard deviations of induced (Ind) and perceived (Per) emotion annotations of multiple annotators per movie and movie genre, respectively. We used the original annotations per second for induced emotions while processing the annotations of perceived emotions per utterance. At an emotion annotation step (a second or an utterance of a movie, respectively), we computed the unbiased standard deviation over all annotators.

TABLE 8: Standard deviations of induced and perceived emotion annotations of multiple annotators per movie. Arousal (Per-A), valence (Per-V), and power (Per-P) dimensions represent perceived emotions while arousal (Ind-A) and valence (Ind-V) dimensions represent induced emotions.

	Per-A	Per-V	Per-P	Ind-A	Ind-V
M1	0.433	0.389	0.374	0.340	0.230
M2	0.404	0.328	0.353	0.239	0.196
M3	0.432	0.425	0.462	0.307	0.319
M4	0.445	0.425	0.421	0.306	0.208
M5	0.390	0.364	0.365	0.294	0.222
M6	0.387	0.456	0.444	0.307	0.253
M7	0.462	0.398	0.430	0.302	0.278
M8	0.439	0.365	0.390	0.264	0.247

TABLE 9: Standard deviations of induced and perceived emotion annotations of multiple annotators per movie genre. Arousal (Per-A), valence (Per-V), and power (Per-P) dimensions represent perceived emotions while arousal (Ind-A) and valence (Ind-V) dimensions represent induced emotions.

	Per-A	Per-V	Per-P	Ind-A	Ind-V
Drama	0.413	0.414	0.412	0.302	0.246
Roman.	0.404	0.328	0.353	0.239	0.196
Comed.	0.432	0.425	0.462	0.307	0.319
Advent.	0.427	0.394	0.401	0.300	0.230

We report the average of the standard deviations for all emotion annotation steps of a movie in Table 8. We observe that the average standard deviation of perceived emotions is larger than the average standard deviation of induced emotions for all movies. This may be due to the use of crowd-sourced annotations for perceived emotions. The perceived

5. https://requester.mturk.com/

emotion annotations were given by 129 untrained annotators from various cultural and educational backgrounds, while the induced emotion annotations were given by 10 trained annotators who were recently graduated master students from France. Therefore, these 10 trained annotators share more similarities in emotion induction and agree more in their annotations. In Table 9 we report the average standard deviations per movie genre. We can similarly infer that the average standard deviation for perceived emotions is larger than that for induced emotions for all the movie genre. Moreover, we see that the annotations of perceived emotions most strongly vary for adventure movies. This may be because the movies *Payload*, *Spaceman* and *Tears of Steel* include action scenes settled down in extraordinary fictional locations (e.g., outer space) and displayed in a spectacular way. These can evoke strong emotional reactions. This conclusion is supported by the high mean level of audience's perceived emotions for adventure movies, as shown in Table 7.

4 Multimodal Feature Extraction

To answer our second research question, we used multimodal signals to improve induced emotion recognition. We revealed complementary information on spectators' induced emotions is encoded in both movie content and spectators' reactions to it [11]. Besides audio-visual features of movie content, we extracted high level affective features, such as lexical features in movie dialogues, aesthetic movie highlight annotations and perceived emotion annotations to describe affective movie content. The lexical features characterize emotions in dialogues expressed by movie main characters while the aesthetic highlight and perceived emotion annotations describe the aesthetic and affective movie content. In addition, we included statistical descriptors of spectators' physiological and behavioral signals to take into account the fact that induced emotions are encoded in movie audiences' reactions. The original arousal-valence annotations from the C. LIRIS-ACCEDE database are provided for each movies at the second level. To capture the suitable amount of temporal information on spectators' physiological and behavioral reactions as well as audio-video movie content with affective cues, we used a 5 second sliding window with a 4 second overlap between neighbouring windows to extract all features.

4.1 Movie Audience Reaction Based Features

To take into account that induced emotions are subjective, we included audience reaction based features, namely statistical features of physiological and behavioral reactions. Also, we assume that each person within a movie audience can display similar behaviors and have similar physiological responses when they are watching a movie together because [18], [50], [55], [56], [57]:

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

The statistical features describe changes and their dynamics in spectators' physiological and behavioral reactions while watching movies. The Galvanic Skin Response (GSR) and ACCeleration signals (ACC) of spectators were filtered by a third order low-pass Butterworth filter with cut-off frequency at 0.3 Hz before feature extraction. These statistical features are mean, median, standard deviation, minimum and maxmimum value as well as minimum and maximum ratio computed over the sliding windows of an original signal and its first and second derivatives [14]. In particular, statistical features were computed over the sliding windows of GSR and ACC signals from sensors attached to each spectator's limbs [50]. Please note that these physiological and behavioral measurements were collected from a different group of participants than those whose induced emotions were annotated as the gold-standard for induced emotion recognition [3], [14].

4.2 Movie Based Features

4.2.1 Audio-Visual Features

We extracted features from the audio-visual movie content by means of the OpenSMILE [17] toolkit. In fact, we computed 1582 InterSpeech2010 Paralinguistic Challenge Low-Level Descriptor audio features [58] and 1793 visual features for each sliding window. The latter include the histograms of Local Binary Pattern, HSV (hue, saturation, and value), as wel as optical flows of each image region [59]. These are considered to be standard benchmark features computed to perform various emotion recognition tasks [49]. Dimensionality reduction was required due to the small number of available instances for model training. This results in less model parameters to tune. To reduce the number of features, we used the ReliefF algorithm [60] and ranked the discriminative power of features for emotion recognition by means of performing regression with 20 nearest neighbours. To do it, we ran ReliefF feature ranking on the remaining 22 movies of the C. LIRIS-ACCEDE database different from the 8 movies on which we performed recognition experiments. This allowed us to incorporate in-domain knowledge and guarantees that testing instances were not included during feature selection. We selected the most discriminative 100 audio features and 100 visual features for arousal and valence prediction, respectively.

The reason that we chose the top 100 features for audio and visual feature sets is to balance the dimensionality between different feature sets and reduce the number of model parameters. This prevents overfitting. In addition, we tested other feature engineering settings, such as selecting more features or performing feature selection on a combined audio-visual feature set. We also applied dimensionality reduction instead of feature selection. We used a linear principal component analysis, a nonlinear principal component analysis with a Gaussian kernel and diffusion maps [61]. In all cases, the first 100 components were sufficient to describe 99 % of the total data variance. However, these did not result in any significant performance improvement.

4.2.2 Lexical Features

It has been shown that lexical features are discriminative for speaker emotion recognition in spontaneous dialogues [51]. The lexical features are DISfluency and Non-verbal Vocalisation (DIS-NV) and Crowd-Sourced Annotation (CSA) features. The former are extracted from manual annotations of DIS-NVs in movie dialogues. The latter are crowd-sourced annotations of arousal, valence and power ratings of 13,915 English lemmas [62]. To extract CSA features, we removed stop words (commonly used words, such as "the", "and", "a", etc.) from the movie transcript and lemmatized the remaining words (e.g., transform "beginning" to "begin") using the Natural Language Toolkit [63]. These are a standard part of pre-processing in Natural Language Processing studies. To compute the feature values, we searched for the lemmas in each sliding window in the dictionary of [62]. Each dictionary entry contains 63 statistics computed over the collected arousal, valence, and power ratings. The statistics are means, standard deviations, and the number of contributing ratings over all the raters and over 6 subsets of raters: male, female, older, younger, high education, and low education, resulting in 21 (3 statistics for the whole set of raters and its subsets) statistics for each emotion dimension. Sums of each of the 63 statistics for all the lemmas in the sliding window are the 63 lexical features.
The six DIS-NV features were computed as the total duration of each category of DIS-NV, including the general lexicons (see Section 3.3) in each sliding window divided by the window length (5s). We did not apply stop word removal or lemmatization for computing the DIS-NV features because these features are based on the duration of words.

4.2.3 Aesthetic Movie Highlights

Aesthetic movie highlights are associated with the occurrences of meaningfull movie scenes defined by experts in terms of art form and content [18]. They are knowledge-inspired features and are more abstract than the audio-visual movie content features. We used annotations of occurrences of six aesthetic highlights (H1, H2, H3, H4, H5, H6) in movies at the time window level. These aesthetic highlights are categorized, as follows:

- Spectacular: technical choices and special effects (H1);
- Subtle: camera use, lighting, and music (H2);
- Character development: main characters' emotional responses to dramatic events (H3);
- Dialogue: clarifying motivation and showing tension among main characters (H4);
- Theme development: unusual close-up and development of the urban theme (H5);
- Any category of highlights above has occurred (H6).

4.2.4 Perceived Emotions

The annotations of perceived emotions of movie audiences were used as high level affective features to recognize induced emotions. The scores in an arousal-valence-power space are averaged and normalized to [−1,1]. Sliding windows were applied to the emotional scores to align them with the features of movie content and movie audience reactions.

5 Recognition Models

In this section we detail LTSM models and their hierarchical architecture to fuse multimodal signals for induced emotion recognition. We select LSTM models instead of BLSTM models to recognize induced emotions from multimodal signals because only previous movie content and emotional states of spectators influence the current emotional states of spectators. This has been supported by previous studies on emotion recognition during interactions between humans and artificial intelligent agents. It was found that BLSTM models did not significantly outperformed LSTM models because future interactions and emotional states cannot influence current interactions and emotional states [64]. Also, SVR and DBN models are described as baseline emotion recognition models. We proposed the hierarchical architecture of LSTM models for fusion of multimodal information based on previous work on emotion recognition [51]. We assume that there is a complex temporal relationship between induced and perceived emotions. This is why we extracted different sets of features that describe affective movie content as well as spectators' physiological and behavioral reactions. We selected LSTM models because of three reasons [11], [51]:

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

However, it is important to mention that building a deep structure (multiple layers) of the LSTM models would require us to have access to massive labelled data. That is why our LSTM models were proposed based on existing LSTM models successfully applied to emotion recognition [51]. We compared our proposed LSTM models to SVR models that are the baseline emotion recognition models [11]. The big advantage of using SVR models is that a small number of training instances is required to find their optimal parameters. However, these SVR models are not able to capture temporal information. Besides SVR models, we compared the proposed LSTM models to DBN models that are able to learn a new representation of data and complex dependencies between them [12]. Nevertheless, temporal information is omitted by the DBN models. Also, a large number of instances is needed to train these models properly. Due to a small number of labelled data available we used the existing machine learning models that could be applied to the induced emotion recognition task.

5.1 Long Short-Term Memory Recurrent Neural Networks

Long Short-Term Memory Recurrent Neural Networks (LSTM) are recurrent neural networks with multiple hidden

layers. This structure allows LSTM models to capture temporal information. It has been shown that a 3 hidden layer hierarchical structure of LSTM models for fusion of multiple modalities improved emotion recognition in spoken dialogues [51]. Moreover, the LSTM model outperformed state of the art algorithms to classify voicing or silence with noise in movies [65].
We built LSTM models using the Keras library [66] for induced emotion recognition. All the LSTM models had three hidden layers with 64 , 32, and 16 neuron units from bottom to top. This architecture with selected hyperparameters was already applied to emotion recognition with success [51]. To avoid overfitting, we used dropout in the first hidden layer with a rate of 0.5 and set the maximum training iteration to 50 epochs with an early stopping tolerance of 10 epochs. The size of mini-batches is 10 due to computational efficiency of training. Other sizes that varied from 3 to 36 were tested. In fact, performance was not influenced by the size selection.
We evaluated three fusion strategies: Feature-Level (FL) fusion (early fusion), Decision-Level (DL) fusion (late fusion), and HierarchicaL (HL) fusion for multimodal emotion recognition [51]. All multimodal features are concatenated in a vector before feeding recognition models when the FL fusion is applied. While using the DL fusion, unimodal recognition models are built for each multimodal feature set and their outputs are incorporated in a decision making module that is another LSTM model. The HL fusion incorporates different multimodal feature sets at different levels of its hierarchy, e.g., aesthetic highlight and perceived emotion annotation based features with noise can be incorporated in lower layers of the LSTM models while more abstract features, e.g., audio and video features can be incorporated in their higher layers.
Furthermore, input neurons of low-level features are connected to the first hidden layer while input neurons of high-level features are directly connected to the second hidden layer for the multimodal HL fusion.
We built multimodal models combining only movie content based features as well as movie content based features with spectators' reactions. The former model uses the descriptors of audio-video content at a higher layer than noisy affective clues because we include in-domain knowledge during feature selection of audio-visual features. The latter model uses features of physiological and behavioral signals at a higher layer than movie based features because movie audiences' reactions are characterized by different dynamics of changes than movie content features.

5.2 Deep Belief Networks

Deep Belief Networks (DBNs) improved emotion recognition, outperforming Deep Neural Networks [67] and Support Vector Machine [68]. It has been shown that two hidden layer DBNs are able to learn a new representation of audio-visual features, capturing complex non-linear dependencies between them [68]. Also, these DBNs are capable of reducing the high dimensionality of the original audio-visual feature space. The structure of DBNs is a stack of multiple restricted Boltzmann machines (RBM). Moreover, the RBMs have drawn increasing attention in current machine learning research because these stochastic graphical models have improved performance in many applications, such as speech recognition and emotion recognition [68], [69]. A basic Bernoulli-Bernoulli RBM (BBRBM) assumes that the input data comes from a binary distribution. This is a crucial limitation. Thus, a RBM assuming that the data are derived from a Gaussian distribution was proposed in [70]. In this paper we only used a Gaussian-Bernoulli RBM (GBRBM) that is a RBM which uses Gaussian distribution for the visible units and binary distribution for the hidden units [71]. Furthermore, a DBN is a stack of multiple RBMs. The hidden units of a learned RBM are used as the visible units of the following RBM. The DBNs are able to learn a high level representation from a large amount of unlabeled instances. Then, relatively small number of labelled data is required for the fine-tunning of the model.
We selected a GBRBM for the input layer with respect to the distributions of physiological and behavioral signals that are better fitted to the Gaussian distribution than the pseudo binary distribution. Other layers were BBRBMs. We learned the DBNs with only 2 hidden layers with 50 and 15 neuron units, respectively, as a result of the limited number of training instances. The size of mini-batch is the number of features divided by 4 due to computational efficiency. The initial learning rate and its upper bound are set to 0.002 for pre-training and the weight-updating ratio is set to 0.1. The cross entropy is used as the loss function. We also applied gradient decent based supervised fine tuning with maximum 100 iterations to find optimal parameters for the whole DBNs. To avoid overfitting on the limited training set, we used a dropout with a ratio of 0.5 for each hidden layer.

5.3 ν-Support Vector Regression

Support Vector Regression models have demonstrated high performance for affect prediction [3], [41], [44], [45]. In this work we used a nonlinear ν-support vector regression (SVR) with a Gaussian kernel as a baseline model for induced emotion recognition [72]. The optimal scaling parameter $\gamma \in \{2^3, ..., 2^{-15}\}$ of the radial basis function, the optimal regularization parameter $C \in \{2^{-5}, ..., 2^{15}\}$, and the optimal parameter $\nu \in (0, 1]$ that controls the number of support vectors were identified by grid search.

6 Experimental Results

6.1 Perceived and Induced Emotions

In this section we respond to our first research question on the relationship between perceived emotions and induced emotions of movie audiences. Please take into account the fact that the induced emotions were annotated at the second level while the perceived emotions were annotated at the utterance level. Also, the perceived emotion annotations are generally longer than one second. That is why we aligned the annotations by computing mean values of induced arousal-valence scores over each movie utterance. This provided us the utterance-level induced emotion annotation. Then, we independently calculated the CC between perceived and induced emotions for each movie. We used a fixed-effects model [73] to analyze the dependence between perceived and induced emotion dimensions described by

CC values. Consequently, we computed weighted average of CC over all 8 movies that is presented in Table 10. To evaluate the practical significance of CC, following Cohen's model [74], we interpret absolute CC values at around 0.1, 0.3, and 0.5 as the small, medium, and large effect sizes, respectively.

As we can see, perceived arousal, valence, and power dimensions are highly positively correlated with each other while induced arousal and valence are moderately negatively correlated with each other. This may be related to the fact that perceived emotion annotation is a more objective task. The negative correlation between induced arousal and valence is consistent with previous work which found a CC of −0.185 between crowd-sourced annotations of induced arousal and valence collected for nearly 14,000 English lemmas [62]. This suggests that induced negative emotions may have stronger arousal than induced positive emotions. However, no definitive conclusions can be made because of the small absolute CC value. Induced valence and perceived emotions have moderately positive correlations, while induced arousal and perceived emotions are weakly or moderately negatively correlated. In particular, perceived arousal and induced arousal are only weakly negatively correlated. These seem that watching too many exciting, pleasant, and dominating scenes in movies may evoke boredom in movie audiences. Nevertheless, movie audiences can feel displeasure during watching movie scenes in which main characters are dominated by dramatic events. This inconsistency between perceived and induced emotion annotations indicates fundamental differences between perceiving emotional movie content and felt emotions by movie audiences. Emotion induction is a complex phenomenon. Various factors other than the emotions conveyed in movie content can influence emotional responses of movie audiences, such as personality, life experience as well as movie and art preferences. Our analysis proves that the assumption that perceived and induced emotions are consistent is not entirely accurate and thus researchers have to take into account this result when designing experiments for affective content analysis research on movies.

TABLE 10: The Pearson correlation coefficient between perceived and induced emotions of movie audience [19] (large magnitudes of the effect size in bold).

Emotion	Per-A	Per-V	Per-P	Ind-A
Per-V	**0.538**	#	#	#
Per-P	**0.652**	0.471	#	#
Ind-A	-0.095	-0.366	-0.170	#
Ind-V	0.243	0.345	0.307	-0.388

6.2 Perceived and Induced Emotions vs. Aesthetic Highlights

In this section we investigated the relationship between aesthetic highlights and both induced and perceived emotions, responding to our first and second research questions. We consider the 8 movies from the C. LIRIS-ACCEDE database as a set of empirical experiments about the given topic. We related the level of induced and perceived emotions of movie audiences with the occurrence of aesthetic highlights in these movies. We calculated effect-size over individual movies. The effect size is the standardized mean difference that is defined as the difference between mean values of continuous emotion annotations of highlight and non-highlight intervals divided by their pooled standard deviation. Positive values indicate a higher level of induced/perceived emotions of highlight scenes in comparison with non-highlight scenes, whereas negative values of the effect size indicate a lower level.

To combine the effect sizes, statistical analysis requires the weighting of each effect size estimate as a function of its precision assuming a fixed-effect model [73]. Here we follow Cohen's benchmarks for the practical significance of the weighted average effect size. We assume that the values around 0.2, 0.5, and 0.8 can be interpreted as the small, medium, and large effect sizes, respectively [74].

We report the weighted average effect size of induced/perceived emotional dimensions for the 8 movies from the C. LIRIS-ACCEDE database in Table 11. Strong emotional reactions may be associated with the occurrence of spectacular highlights H1 in movies, such as adding special effects, changes in saturation of colors, lightening, and camera location. A small positive effect size of induced and perceived arousal and a small negative effect size of induced and perceived valence are observed for spectacular highlights H1. Moreover, a small negative effect size of perceived power is found. It is important to point out that the directions of effects for both induced and perceived arousal/valence are only consistent during highlights H1.

TABLE 11: Dependencies between aesthetic highlights and perceived and induced emotions of movie audience (small, medium, and large magnitudes of the effect size in bold).

	Per-A	Per-V	Per-P	Ind-A	Ind-V
H1	**0.325**	**-0.219**	**-0.378**	**0.481**	**-0.264**
H2	-0.028	**-0.840**	**-0.522**	0.172	0.065
H3	**0.243**	-0.177	0.055	0.148	**0.224**
H4	**-0.201**	**-0.467**	-0.019	0.092	**-0.243**
H5	0.167	**-0.222**	**-0.240**	**0.292**	**0.286**

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

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

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

Theme development highlights H5 partially overlap with other categories of aesthetic highlights, for example, spectacular highlights H1 and character development highlights H3. In particular, the development of a theme is often associated with some changes in emotional states of main characters as their reactions to dramatic events presented in a spectacular or sublime manner. We observe a small negative effect of perceived valence and perceived power. Also, we find a small positive effect of induced arousal and valence. A related point to consider is the incoherence of the effect directions for perceived and induced valence. It means that perceiving negative valence (unpleasantness) can evoke pleasure in the audience.

Essentially, we find aesthetic highlights as high level aesthetic cues that include information on perceived and induced emotions regardless of the discrepancies between them.

6.3 Induced Emotion Recognition

In this section we recognize induced emotions from multimodal information, answering our second research question. The average arousal-valence scores over each window of length 5s are used as the gold-standard induced emotion annotations. We also removed the end credits of each of 8 movies because participants started to remove the wearable sensors at this point, which introduce outliers in the signals. This results in 7103 data instances in total.

Since the small amount of annotated data is available for induced emotion recognition, we performed leave-one-movie-out cross-validation [41], [44]. At each round, instances from one movie are left out as the test set while instances from other movies (7 movies) are used for training. We computed the unweighted average of the MSE as well as the absolute values of the CC and the CCC for arousal (A) and valence (V) prediction. For example, A-MSE refers to the average MSE over leave-one-movie-out cross-validation for arousal prediction. The MSE and CC are the most commonly reported evaluation metrics in the related work (see Section 2.2). A high value of the CC represents a strong linear relationship between values of emotion predictions and annotations. This means that general value changes (increase/decrease trends) in both signals co-occur. A low value of the MSE corresponds to a high quality of the predictive model. The CCC combines the CC with the square difference between the mean of the two compared time series, which makes it sensitive to bias and scaling factors [24]. This measures is commonly applied to multiple unambiguous annotation predictions, for example, induced emotions [24] (see Tables 8 and 9). A large value of the CCC describes a high agreement between values of predictions and annotations. This means that prediction and annotation values are similar to each other and general trend changes in both signals are the same. We used the following validation to investigate the statistical significance of the results. In order to show that our models performed better than a random prediction model, we generated arousal and valence prediction scores at random. Then, we compared predictions of two models with highest CC or CCC values for each experiment to random predictions of arousal and valence scores, respectively. Finally, we compared the predictions of these pairs of models that did not perform randomly (e.g., the SVM models fed by GSR and audio features, respectively, for arousal prediction). All the statistical comparisons were made by means of two-sample Wilcoxon tests with $p < 0.05$ being significant. When we report results for each experiment, numbers in bold italics indicate significantly best performance with ($p << 0.0001$) and numbers in bold indicate significantly best performance with $p < 0.05$.

6.3.1 Influence of history on induced emotion

The original induced emotion annotations, which were provided by the C. LIRIS-ACCEDE database, were annotated at every single second. The average absolute difference between adjacent arousal annotations is 0.006 and between valence annotations is 0.005. These changes in annotations are extremely small considering that the annotation value range is [−1,1]. Previous work has shown that human emotions are context dependent and typically do not change rapidly over a small time interval [49]. However, the suitable amount of temporal context for predicting movie induced emotions remains unknown.

We attempt to identify suitable amount of history for predicting induced emotions by testing LSTM models fed by physiological features with different time steps. We used physiological features because they are representatives of the audience's induced responses [50]. Our experiments show that including features for the past 3 time steps gives better recognition performance than shorter or longer time steps. Thus, all the LSTM models in this work used a time step of 3. Recall that our feature vectors are extracted over a 5 second sliding window with 4 seconds overlap. With 3 history feature vectors the model will have 8 seconds of temporal information (including the current window).

6.3.2 Unimodal induced emotion recognition

The results of our unimodal induced emotion recognition experiments are shown in Table 12 in which we report the average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction. As we can see for arousal and valence prediction, the SVR model achieved the best performance on physiological features and perceived emotion features measured by the CC and CCC, respectively. This means that physiological signals and perceived emotions provide discriminative information on induced emotions. Moreover, the SVM is able to capture the dependencies between changes in physiology and emotional states of spectators. As shown in Table 12, the SVM can only predict an increase or decrease of arousal and valence intensity from GSR signals with respect to the CC values. Besides, the values of the CCC suggest that the same SVR model is able to predict induced emotions from perceived emotion annotations in terms of upward and downward trends and values as well. Nevertheless, the large values of MSE suggest that there is a need to improve learning of this model for these emotion recognition tasks. To prove the statistical significance of the results, we first referred the predictions of two SVR models with the highest performance to predictions of a random prediction model for each experiment. As a result, we showed that SVR predictions were significantly

different from random predictions ($p << 0.0001$). Then, we compared the arousal and valence predictions of these SVR models. We found that all of them were significantly different ($p << 0.0001$), except for the CC of valence prediction from GSR and visual features ($p = 0.7584$). As shown in Table 12, the DBN model best performed induced emotion recognition using audio features of movie content with regard to the values of CC. This means that trends in arousal and valence intensity over time are easily captured. Moreover, the values of the CCC suggest that the DBN is also able to accurately predict the values of arousal scores. However, this is not the case for valence prediction. The DBN achieved the highest values of the CCC for valence prediction from aesthetic highlight annotations. Firstly, we referred the predictions of two DBN models with the highest performance to random arousal and valence predictions for each experiment. We showed that these DBN models performed significantly different from a random prediction model ($p << 0.0001$). Then, we compared arousal and valence predictions of these DBN models. We found that all of them were significantly different with $p << 0.0001$.

TABLE 12: Performance of unimodal induced emotion recognition using SVR, DBN, and LSTM models is reported.

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

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

The LSTM model could predict induced arousal from audio features with regard to the CC values, as shown in Table 12. However, the values of the CCC suggest that the features of behavioral signals are the most discriminative at least for induced arousal prediction. Moreover, the LSTM model best performed valence prediction from the physiological signals. The values and trends of valence intensity were captured by the LSTM model fed by the GSR features. This is confirmed by the high values of the CC and CCC, respectively. To validate the results of two LSTM models with the highest performance, we first compared their predictions to random arousal and valence predictions for each experiment. We proved that the predictions of these LSTM models performed significantly better than random predictions ($p << 0.0001$). We then compared the predictions of these LSTM models. As a result, we observed that all of them were significantly different with $p << 0.0001$. However, there was an exception for the CC of valence prediction based on GSR signals and perceived emotion annotations ($p = 0.4782$).

It is important to mention that our results are not directly comparable with previous work due to different data processing procedures, such as the use of the overlapping window and different settings of cross validation, e.g., the number of folds and the size of training and testing sets. Nevertheless, we can see that we outperformed the state of the art recognition models (Table 1) for valence prediction by means of the LSTM models with the statistical features of GSR signals (a CC of 0.432).

6.3.3 Multimodal induced emotion recognition

We report the average of the MSE as well as CC and CCC absolute values over leave-one-movie-out cross-validation for arousal (A) and valence (V) prediction. Tables 13 and 14 present the results of multimodal induced emotion recognition experiments. We consider fusion of all the audio-video features with high-level affective clues, such as audio, video, CSA, DIS-NV features as well as aesthetic highlight and perceived emotion annotation based features. Moreover, we investigated the fusion of all the movie content based features mentioned above with physiological and behavioral responses of movie spectators. We compared the proposed hierarchical fusion (LSTM-HL) architecture of LSTM models to baseline fusion strategies for LSTM models, such as feature-level fusion (LSTM-FL) and decision-level fusion (LSTM-DL) (see Section 5). Also, we examined the recognition performance of SVM and DBN models when the FL fusion was applied.

As seen in Table 13, the LTSM model with the FL fusion best performed induced arousal recognition from movie content based features with respect to the CC values. It means that trend changes in arousal intensity could be easily captured by this model. Nevertheless, the values of CCC suggest that the proposed hierarchical fusion architecture of the LSTM model could best predict induced arousal in terms of trends and values. Besides, the LSTM-HL did not succeed in recognizing induced valence. The LSTM-DL reached the highest value of the CC. Actually, the LSTM-FL outperformed the other fusion strategies and predictive models and could the most accurately predict the values and trend fluctuations for induced valence according to the CCC values. Generally, all LSTM models outperformed SVR and DBN models for induced emotion recognition from movie based features.

As shown in Table 14, the SVR model could be the most accurate predictor of trend changes in induced arousal intensity from fusion of both movie content features and movie audience reactions. However, the large value of the MSE indicates that the SVR model was not able to predict arousal values as well as slight increases and decreases

in trends. Furthermore, the LSTM-HL achieved the highest value of the CCC. This means the accurate prediction of downward/upward trend changes in induced arousal intensity as well as its values. Also, the LSTM-HL best performed induced valence recognition that is confirmed by values of the CC and CCC, respectively. The results that are obtained suggest that the proposed hierarchical architecture of LSTM models for fusion of movie content features and movie audience reactions is well designed to predict the intensity of induced arousal and valence. To prove the statistical significance of the results obtained from multimodal fusion, we referred the arousal and valence predictions of two multimodal fusion models with the highest performance to predictions of a random prediction model. We observed that all of them performed significantly different with $p << 0.0001$. Next, we compared arousal and valence predictions of these pairs of the multimodal fusion models fed by movie content based features as well as movie content based features and statistical features of audience reactions, respectively. We remarked that all of them were significantly different with $p << 0.0001$.

TABLE 13: Performance of multimodal induced emotion recognition from movie content based features using SVR, DBN, and LSTM models is reported.

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

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

TABLE 14: Performance of multimodal induced emotion recognition from audience reaction and movie content based features using SVR, DBN, and LSTM models is reported.

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

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

7 Discussion

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

7.1 Limitations of our study

Induced emotions can be expressed through different multimodal channels. The importance of these channels is not the same for induced emotion recognition. Different spectators can have different physiological and behavioral responses to the same stimuli. These can be affected by ambient temperature, body postures, gestures as well as attention and mental effort. Furthermore, induced emotions can vary from one person to another due to many factors e.g., personal life experience. Recording and combining multimodal signals of a group of subjects still remain a big challenge due to a lack of access to non-obstructive and reliable sensors. This limits the feasibility of running a large scale experiment in a cinema theater. Measurements of physiological and behavioral signals are often corrupted due to electrode contact noise and sensor device failures during data collection. This results in incomplete data.

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

7.2 Available modalities and sample size

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

Since spectators were watching movies in a cinema theater, the galvanic skin response and acceleration measurements of each spectator could be only collected due to technical constraints and the number of resources available. Our unimodal experiments on induced emotion recognition confirm that spectators have similar physiological responses and display similar behaviors during watching movies. However, the behavioral features are less discriminative than the physiological features for induced emotion recognition. This outcome might be influenced by the placement of sensors. The sensors were attached to spectators' hands when the experiment was conducted. We do not observe that spectators often make some limb movements when they are watching movies.

The inter-annotation agreement for induced and perceived emotions is low. To reduce this variability in the gold standard, the dynamics of changes in annotations could be considered instead of emotion intensity. Moreover, some outlier annotations might be removed, and identifying and correcting annotators' biases can be applied.

7.3 Model selection

The results that we obtained show that the small amount of labelled instances available for emotion recognition can significantly limit the quality of model training and the performance of the emotion recognition system. Because of that, using existing architectures with hyperparameters of emotion recognition models is strongly suggested when we do not have access to a large amount of labelled data to build a emotion recognition system. Model selection is strictly associated with the amount and type of available multimodal data that are recorded and annotated as well as evaluation metrics. The CC could be selected when the goal

is only to capture trend changes in induced emotions by using models. However, the CCC is a more suitable measure to evaluate the quality of models since it describes if models are able to capture changes in trends and estimate values of emotion intensity.

When physiological and behavioral reactions are not recorded and high-level affective cues are not annotated, it is recommended that induced emotions should be recognized by DBN models fed by audio movie features. If physiological or behavioral measurements are available, the results suggest that LSTM models should be applied due to their capabilities of capturing long term dependencies in movie audience reactions. As seen in Table 12, SVR and LSTM models with statistical features of GSR and ACC signals achieved the highest performance of emotion recognition regarding CC or CCC values. It means that dynamic changes of physiological and behavioral reactions are highly discriminative to recognize emotions induced in movie audiences. Besides, when it is only possible to run crowdsourcing annotation experiments, SVM models should be learned on high-level affective cues, such as annotations of aesthetic highlights in movies or perceived emotions of movie audience (see Section 6.3.2).

Our multimodal experiments on induced emotion recognition show that our LSTM models benefit from including temporal information and combining knowledge-inspired affective cues with audio-visual movie content and movie audience responses. As shown in Tables 13 and 14, the fusion of spectators' physiological and behavioral reactions with movie content features improves emotion recognition. However, the hierarchical incorporation of multimodal features is required to increase the performance since movie content features and spectators' reactions do not have the same dynamics of changes in temporal patterns. There is a need to work on LSTM architectures to incorporate high-level affective cues with audio-visual movie content features since the proposed hierarchical fusion did not improve induced valence recognition (see Section 6.3.3).

The SVM and DBN models could not capture consecutive emotional states and reactions of spectators because they do not take into account temporal information. This is why the LSTM models could outperform them. Also, feature fusion by means of these baseline models does not allow multimodal features to be incorporated at different stages of modelling. Thus, multilevel fusion is desired to fuse features with different temporal dynamics, e.g., audio-video features of movie content and statistical features of spectators' physiological and behavioral reactions. The last but not least limitation is that these basic models cannot deal with noisy features and temporal evolution of the probability distribution of movie content features and statistical features of movie audience reactions. The probability distribution varies from one movie to another because measurements of physiological and behavioral signals are corrupted by electrode contact noise and they are subject-dependent. Furthermore, audio-video features are contaminated with movie background noise. On the contrary, the LSTM models are able to operate on different scales of time which limits the influence of variability of spectators' physiological and behavioral signals and movie content. Also, noisy features can be filtered out by learning a new representation in the first layer of LSTM models.

8 Conclusion

This work clarifies the difference between perceived and induced emotions of movie audiences and serves as a reference for future affective content analysis studies. We extend annotations on the C. LIRIS-ACCEDE database and find that perceived and induced emotions of movie audiences are not always positively correlated regarding our first research question. Although the inconsistency was observed on a fairly small movie data set, it should be taken into account when selecting stimuli for emotion induction. There is more to be considered than simply assuming that the perceived emotions of the stimuli are consistent with the emotions induced in spectators. To expand our understanding of perceived and induced emotions and address our second research question, we used perceived emotions to predict induced emotions. Moreover, perceived and induced emotions of the movie audiences are associated with the occurrences of aesthetic highlights in movies. These highlights are considered to be high level affective cues for induced emotion recognition. The improvement of performance using multimodal hierarchical fusion leads us to the conclusion that adding other modalities, such as facial expressions, heart rate, and electroencephalography signals of spectators could result in a large increase of performance. Also, our promising model can be scalable to a larger movie set and thus its architecture and generalization can benefit from a larger number of labelled instances available for training. Nevertheless, there is a need to deeply study in which layer of the model audio-video features and affective cues should be incorporated.

In the future, we will be studying the advantages of using transfer learning between different emotion recognition tasks. The pretrained models on other emotion recognition challenges, e.g., emotion recognition of individuals watching short videos, could be applied to induced emotion recognition of movie audiences. Also, we will attempt to improve performance by means of learning feature representations and designing new architectures of multimodal recognition models. Moreover, we plan on conducting further investigations into how emotions and affective cues differ from one movie genre to another, e.g., action, crime, epics, historical, horror, etc. Studies on different movie emotion perspectives may make a major contribution to cinematography research as well as help moviemakers to design affective content with better alignment of intended and induced emotions.

Acknowledgment

Michal Muszynski is funded by the Computer Vision and Multimedia Laboratory, the University of Geneva. This work is partially supported by grants from the Swiss Center for Affective Sciences and the Swiss National Science Foundation. Leimin Tian was funded by School of Informatics, the University of Edinburgh while conducting the initial experiments, and by the Faculty of IT, Monash University during revision of this study. We want to thank Mohammad Soleymani, Anna Aljanaki, and Soheil Rayatdoost for their generous help on Amazon Mechanical Turk experiments.

REFERENCES

[1] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[2] 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.

[3] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *ACII2015*. IEEE, 2015, pp. 77–83.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, "Appraisal theories of emotion: State of the art and future development," *Emotion Review*, vol. 5, no. 2, pp. 119–124, 2013.

[6] J. Ressel, "A circumplex model of affect," *J. Personality and Social Psychology*, vol. 39, pp. 1161–78, 1980.

[7] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[8] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, no. 1_suppl, pp. 123–147, 2001.

[9] E. S.-H. Tan, "Film-induced affect as a witness emotion," *Poetics*, vol. 23, no. 1-2, pp. 7–32, 1995.

[10] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Transactions on Affective Computing*, 2017.

[11] 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, no. 2, pp. 92–105, 2011.

[12] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[13] G. Gninkoun and M. Soleymani, "Automatic violence scenes detection: A multi-modal approach," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, 2011.

[14] T. Li, Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Continuous arousal self-assessments validation using real-time physiological responses," in *ASM2015*. ACM, 2015, pp. 39–44.

[15] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, C. Chamaret, and E. Lyon, "The mediaeval 2016 emotional impact of movies task," in *MediaEval2016*, 2016.

[16] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *ACII2015*. IEEE, 2015, pp. 698–704.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," in *ICMI2010*. ACM, 2010, pp. 1459–1462.

[18] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Dynamic time warping of multimodal signals for detecting highlights in movies," in *INTERPERSONAL2015*. ACM, 2015, pp. 35–40.

[19] L. Tian, M. Muszynski, C. Lai, J. Moore, T. Kostoulas, P. Lombardo, T. Pun, and G. Chanel, "Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same?" pp. 28–35, 2017.

[20] M. 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.

[21] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "Lstm for dynamic emotion and group emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 451–457.

[22] M. Wöllmer, F. Eyben, S. Reiter, B. W. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies." in *Interspeech*, vol. 2008, 2008, pp. 597–600.

[23] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*, 2017.

[24] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 985–990.

[25] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, 2017.

[26] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 217–224.

[27] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-clue fusion for emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 458–463.

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

[29] C. Plantinga, "Art moods and human moods in narrative cinema," *New Literary History*, vol. 43, no. 3, pp. 455–475, 2012.

[30] E. Kroupi, J.-M. Vesin, and T. Ebrahimi, "Phase-amplitude coupling between EEG and EDA while experiencing multimedia content," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 865–870.

[31] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *ACII2013*. IEEE, 2013, pp. 73–78.

[32] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 505–523, 2011.

[33] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, 2017.

[34] J. Plested, T. Gedeon, X. Zhu, A. Dhall, and R. R. Geocke, "Detection of universal cross-cultural depression indicators from the physiological signals of observers," in *Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017 Seventh International Conference on*. IEEE, 2017, pp. 185–192.

[35] G. Matthews, D. M. Jones, and A. G. Chamberlain, "Refining the measurement of mood: The UWIST mood adjective checklist," *British journal of psychology*, vol. 81, no. 1, pp. 17–42, 1990.

[36] Y. Song, S. Dixon, M. T. Pearce, and A. R. Halpern, "Perceived and induced emotion responses to popular music," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 4, pp. 472–492, 2016.

[37] K. Knautz and W. G. Stock, "Collective indexing of emotions in videos," *Journal of Documentation*, vol. 67, no. 6, pp. 975–994, 2011.

[38] K. Kallinen and N. Ravaja, "Emotion perceived and emotion felt: Same and different," *Musicae Scientiae*, vol. 10, no. 2, pp. 191–213, 2006.

[39] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2085–2098, 2014.

[40] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1356–1370, 2011.

[41] 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, no. 1, pp. 43–55, 2015.

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

[43] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[44] T. Anastasia and H. Leontios, "AUTH-SGP in MediaEval 2016 emotional impact of movies task," in *MediaEval2016*, 2016.

[45] S. Chen and Q. Jin, "RUC at MediaEval 2016 emotional impact of movies task: Fusion of multimodal features," in *MediaEval2016*, 2016.

[46] Y. Ma, Z. Ye, and M. Xu, "THU-HCSI at MediaEval 2016: Emotional impact of movies task," in *MediaEval2016*, 2016.

[47] A. Jan, Y. F. A. Gaus, F. Zhang, and H. Meng, "BUL in MediaEval 2016 emotional impact of movies task," in *MediaEval2016*, 2016.

[48] Y. Liu, Z. Gu, Y. Zhang, and Y. Liu, "Mining emotional features of movies," in *MediaEval2016*, 2016.

[49] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[50] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals," in *QoMEX2015*. IEEE, 2015, pp. 1–6.

[51] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," *SLT2016*, 2016.

[52] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *AVEC2016*. ACM, 2016, pp. 97–104.

[53] G. V. Bard, "Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric," in *ACSW2007*, vol. 68. Australian Computer Society, Inc., 2007, pp. 117–124.

[54] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[55] M. Muszynski, T. Kostoulas, G. Chanel, P. Lombardo, and T. Pun, "Spectators' synchronization detection based on manifold representation of physiological signals: Application to movie highlights detection," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 235–238.

[56] M. Muszynski, T. Kostoulas, P. Lombardo, T. Pun, and G. Chanel, "Synchronization among groups of spectators for highlight detection in movies," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 292–296.

[57] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Films, affective computing and aesthetic experience: Identifying emotional and aesthetic highlights from multimodal signals in a social setting," *Frontiers in ICT*, vol. 4, p. 11, 2017.

[58] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, S. S. Narayanan *et al.*, "The INTERSPEECH 2010 paralinguistic challenge." in *Interspeech*, vol. 2010, 2010, pp. 2795–2798.

[59] F. Eyben, F. Weninger, M. Wöllmer, and B. Schuller, "open-source media interpretation by large feature-space extraction," 2016.

[60] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[61] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[62] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[63] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[64] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[65] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 483–487.

[66] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[67] E. M. Albornoz, M. Sánchez-Gutiérrez, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Spoken emotion recognition using deep learning," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2014, pp. 104–111.

[68] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.

[69] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5688–5691.

[70] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[71] T. Yamashita, M. Tanaka, E. Yoshida, Y. Yamauchi, and H. Fujiyoshii, "To be Bernoulli or to be Gaussian, for a Restricted Boltzmann Machine," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 1520–1525.

[72] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[73] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd, 2009.

[74] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*, 2nd ed. Routledge, 1988.

Michal Muszynski received his B.Sc. in Automatic Control and Robotics, M.Sc. in Automatic Control and Robotics and B.Sc. in Mathematics at the Warsaw University of Technology, Poland in 2011, 2012 and 2014, respectively. He then received Ph.D. in Computer Science at the University of Geneva, Switzerland in 2018. He is now a post-doctoral research fellow at the Neurology and Imaging of Cognition Laboratory in the Department of Fundamental Neurosciences within the University of Geneva, Switzerland. His research interest include affective computing, signal processing, and pattern recognition. In particular, he is working on physiological signal analysis for social interactions and aesthetic highlight detection in movies.

Leimin Tian Leimin achieved her Bachelor of Engineering in 2012 in Intelligent Science and Technology at School of Computer Science, Beijing University of Posts and Telecommunications. She then went to the Institute of Language, Cognition and Computation, School of Informatics, the University of Edinburgh and achieved her Master of Science in 2013 and her PhD in 2018. She is now a research fellow at the Computer Human Interaction and Creativity (CHIC) group at Monash University. Since undergraduate, Leimin has been working on Affective Computing. In particular, she is working on exploring effective approaches to combine human knowledge on emotions and the power of computational models to realize emotion-aware human-computer/robot interaction.

Catherine Lai is a Post-doctoral Research Associate working at the Centre for Speech Technology at the University of Edinburgh, U.K. She previously received a M.Sc. in Computer Science from University of Melbourne, Australia and a Ph.D in Linguistics at the University of Pennsylvania, U.S.A. She has worked on a wide range of topics in spoken language processing and linguistics including multimodal meeting summarization, emotion recognition, modelling of native and non-native prosody, and spoken dialogue semantics.

Johanna D. Moore is a Professor and Head of the School of Informatics, the chair of Artificial Intelligence at the University of Edinburgh, and a Fellow of the Royal Society of Edinburgh and the British Computing Society. She has over 25 years of experience in the areas of dialogue systems, natural language generation, and multimodal interaction. She is a past President of the Association for Computational Linguistics and past Chair of the Cognitive Science Society. She is currently Associate Editor for the journals: Cognitive Science and Speech Communication.

Theodoros Kostoulas holds Electrical and Computer Engineering Degree and Ph.D. in Electrical and Computer Engineering, University of Patras, Greece, 2012. Since 2005 he has been continuously participating in a number of research and development projects and he has long experience in affective computing, signal processing, machine learning and data resources design. From 2013-2017 he was postdoctoral researcher at the Computer Vision and Multimedia Laboratory and the Swiss Center for Affective Sciences, Geneva, Switzerland. Following the postdoctoral activities, he is now Lecturer at Bournemouth University, UK (2017 -), where he is leading the Affective Computing and Multimodal Interaction research topic of the Engineering of Social Informatics Research Group (ESOTICS). He has more than 30 peer-reviewed articles.

Patrizia Lombardo has taught at Princeton University, University of Southern California Los Angeles, University of Pittsburgh and University of Geneva. She is now Emeritus Professor in the Department of French at the University of Geneva where she has been leading since 2009 the Project "Aesthetic Emotions and Affective Dynamics" at the Interdisciplinary Center for Affective Sciences . She has published scholarly articles on French 19th and 20th century literature, literary criticism and literary theory, comparative literature, aesthetics, history of ideas, and theories of emotions.

Thierry Pun is full professor at the University of Geneva, Switzerland, head of the Computer Vision and Multimedia Laboratory (CVML,http://cvml.unige.ch/). His current research interests are in affective computing and emotions analysis, social signal processing and multimodal interaction. He is involved in projects concerning affective computing and gaming, physiological signals analysis for emotion assessment and social interactions analysis, multimodal interfaces for blind and elderly users, affective brain-computer interaction. Thierry Pun has authored or co-authored over 400 full papers as well as eight patents (http://scholar.google.ch/citations?user=sR12P9MAAAAJ&hl=fr). He was one of the general chairs and organizers of ACII 2013 - Affective Computing and Intelligent Interaction in Geneva, Switzerland. T. Pun has been instrumental in the creation of two spin-offs of the University of Geneva, and has participated in and/or led a number of research projects, Swiss and European, financed by public and private entities. He is an elected member of the Swiss Academy of Engineering Sciences SATW.

Guillaume Chanel holds a Ph.D. in Computer science, University of Geneva, 2009, where he worked on machine learning for the automatic assessment of emotions based on EEG and peripheral signals. From 2009 to 2010 he was at the KML-Knowledge Media Laboratory, Aalto University, Helsinki, Finland, studying the physiological correlates of social processes taking place between players during video-gaming. Now a senior researcher and lecturer jointly affiliated with CISA and with CVML, his research investigates how machines can learn to behave in a social and affective environment. He is particularly interested in the use of multimodal and physiological measures for improving man-machine and human remote interactions. Examples of his research include: dynamic adjustment of games mechanic based on players' emotions, inclusion of physiological emotional cues in mediated social interactions, movie highlight detection based on spectators' social reactions and adaptation of human social behaviors through machines.