Affective Video Content Analysis: A Multidisciplinary Insight

Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, Liming Chen

Affective Video Content Analysis: A Multidisciplinary Insight

Yoann Baveye, Christel Chamaret, Emmanuel Dellandréa, and Liming Chen

Abstract—In our present society, the cinema has become one of the major forms of entertainment providing unlimited contexts of emotion elicitation for the emotional needs of human beings. Since emotions are universal and shape all aspects of our interpersonal and intellectual experience, they have proved to be a highly multidisciplinary research field, ranging from psychology, sociology, neuroscience, *etc.*, to computer science. However, affective multimedia content analysis work from the computer science community benefits but little from the progress achieved in other research fields. In this paper, a multidisciplinary state-of-the-art for affective movie content analysis is given, in order to promote and encourage exchanges between researchers from a very wide range of fields. In contrast to other state-of-the-art papers on affective video content analysis, this work confronts the ideas and models of psychology, sociology, neuroscience, and computer science. The concepts of aesthetic emotions and emotion induction, as well as the different representations of emotions are introduced, based on psychological and sociological theories. Previous global and continuous affective video content analysis work, including video emotion recognition and violence detection, are also presented in order to point out the limitations of affective video content analysis work.

Index Terms—Multidisciplinary review, Video content analysis, Cognitive models, Affective computing, State of the art.

1 Introduction

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

Under these circumstances, knowing in advance the emotions that a movie is likely to elicit from its viewers is highly beneficial; not only to improve accuracy for video indexing, and summarization (*e.g.* [1], [2]), but also for mood-based personalized content delivery [3]. While major progress has been achieved in computer vision for visual object detection, scene understanding and high-level concept recognition, a natural further step is the modeling and recognition of affective concepts. This explains why the affective video content analysis research topic has emerged and attracted increasingly more attention in recent years.

- *Y. Baveye is with Ixpel, an expertise cell of CAPACITÉS, and is with the University of Nantes, Centre National de la Recherche Scientifique, LS2N, UMR6004, France.*
  E-mail: yoann.baveye@univ-nantes.fr
- *C. Chamaret is with Technicolor, 975, avenue des Champs Blancs, 35576 Cesson Sévigné, France.*
  E-mail: christel.chamaret@technicolor.com
- *E. Dellandréa and L. Chen are with the Université de Lyon, Centre National de la Recherche Scientifique, Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France.*
  E-mail: {emmanuel.dellandrea, liming.chen}@ec-lyon.fr

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

Affective video content analysis aims at automatic recognition of emotions elicited by videos. There are three perspectives for affective movie content analysis work, each related to specific emotion detection: **intended**, **induced** and **expected** emotion [3]. The *intended* emotion is the emotion that the film maker wants to induce in the viewers. The *induced* emotion is the emotion that a viewer feels in response to the movie. The *expected* emotion is the emotion that the majority of the audience feels in response to the same content. In other words, the expected emotion is the expected value of experienced (*i.e.* induced) emotion in a population. While the induced emotion is subjective and context dependent, the expected emotion can be considered objective, as it reflects the more-or-less unanimous response of a general audience to a given stimulus [3]. Intended and expected emotions do not always match [4], [5]. It is a fact that a movie can be unsuccessful in conveying the intended emotion (*e.g.*, viewers laughing while watching horror movies intended to elicit fear). The degree of effectiveness with which a movie creates the desired emotion in the viewer can be a basic criterion for assessing movie quality [6].

A large number of work has been proposed in the literature to tackle this highly challenging task. However, a review of the state of the art shows that affective video content analysis work from the computer science community benefits very little from the progress made in other research fields, including psychology, sociology, and neuroscience. Indeed, it appears from previous works that the main approach for analyzing the affective content of videos is to use machine learning to build, using a dataset, models such as Hidden Markov Models (HMMs), Support Vector Machines for Regression (SVRs), and more recently Convolutional Neural Networks (CNNs).

The main contribution of this paper is to propose a multidisciplinary state of the art for affective movie content

analysis, in order to promote and stimulate exchanges between researchers from a wide variety of fields. In contrast to previous state of the art studies [7], [8], this work thus confronts the ideas and models from psychology, sociology, neuroscience, and computer science with a focus on the emotions elicited by movies. General psychological emotional theories are described alongside models for aesthetic emotions, focusing on emotions in response to a work of art. With these theoretical models in mind, previous works on continuous and global affective video content analysis are presented, including video emotion recognition and violence detection, which are related to the expected affect estimation from audiovisual features. Studies, known as implicit video affective content analysis, have also been conducted using the spontaneous response of users watching the videos (*e.g.*, facial recordings, physiological responses). However, this topic has already been investigated in [7]. This paper focuses rather on direct video affective content analysis studies, using audiovisual features extracted from the videos to infer a given type of emotion.

The paper is organized as follows. Section 2 describes main psychological emotional theories, gives an overview of neuroscience work, and defines the concepts of aesthetic emotions and emotion induction, as well as different representations of emotions. Section 3 presents previous global and continuous affective video content analysis workand lists current existing affective video databases.Limitations are discussed in Section 3.5, while the paper ends in Section 5 with conclusions.

2 EMOTIONS

More than one hundred different scientific definitions were inventoried [9]. Kleinginna and Kleinginna proposed a consensus by suggesting the following definition considered as one of the most comprehensive definitions of emotions [9]:

> "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive." (pp. 355)

2.1 Main psychological approaches

Decades of research driven by psychologists interested in the comprehension of emotions led to three major theoretical approaches: **basic**, **appraisal**, and **psychological constructionist** [10].

2.1.1 Basic emotions: facial expression and emotion

Basic emotion theorists were originally inspired by the ideas expressed by Darwin. In *The Expression of the Emotions in Man and Animals* [11], Darwin deduced the universal nature of facial expressions. Thus, basic emotion models assume that several emotions are automatically triggered by objects and situations in the same way everywhere in the world. This idea of universality was reused by Ekman and Friesen to develop the Emotion Facial Action Coding System (EMFACS) [12], which is a simplified although less common version of FACS [13] designed explicitly for the coding of facial behavior related only to emotion. Ekman considers emotions to be discrete states that are associated with facial expressions. Consequently, he postulates that there are a fixed number of emotions. Plutchik agreed with Ekman's biologically driven perspective and theorized that basic emotions are biologically primitive and have evolved in order to increase the reproductive fitness of the animal [14]. The term *basic* is used in the literature to describe three perspectives [15]:

- Basic emotions are different from each other and show unique characteristics such as physiological and behavioral responses
- Basic emotions are a learned response to deal with fundamental life tasks
- Basic emotions are elemental, and combine together to form complex emotions

More specifically, the basic emotion theory identifies several criteria for an emotion to be considered basic [15], such as: distinctive universal signals, distinctive physiology, distinctive universals in antecedent events (*i.e.*, some stimuli render a universal response), presence in other primates, quick onset, brief duration. Ekman used these criteria to define six basic emotions: fear, anger, sadness, disgust, joy, and surprise [16]. In later works, Ekman and Heider [17] argued for the inclusion of additional basic emotions, such as contempt. Many other lists of basic emotions have been proposed in the literature [18]. For example, Plutchik identified eight basic emotions, which he grouped into four pairs of polar opposites (joy-sadness, anger-fear, trust-distrust, surprise-anticipation) [14], while Izard *et al.* [19] postulated 10 fundamental emotions, universally discernible in the human facial expression.

2.1.2 Appraisal theories

Appraisal models assume that emotions are triggered by the interpretation of stimulus events and thus can be seen as relevance detectors [20]. Appraisal theories study why individuals react differently when confronted with the same stimulus [21]. Arnold made early contributions to the appraisal theory, suggesting that an initial appraisal triggers the emotional sequence by arousing both the appropriate physiological reactions and the emotional experience itself [22]. Lazarus specified two types of appraisal: primary and secondary appraisals [23]. Lazarus suggested that primary appraisal is focused on the conception of the significance or meaning of the stimulus to the organism, while secondary appraisal aims at assessing the ability of the organism to cope with the consequences of the stimulus. Another influential appraisal theory of emotion is the Component Process Model (CPM), Scherer's major contribution to the study of appraisal processes, described in the remainder of this section.

The CPM postulates that the emotion process is a psychological construct driven by subjective appraisal and is the consequence of synchronized changes in five components corresponding to five distinctive functions [24]: cognitive

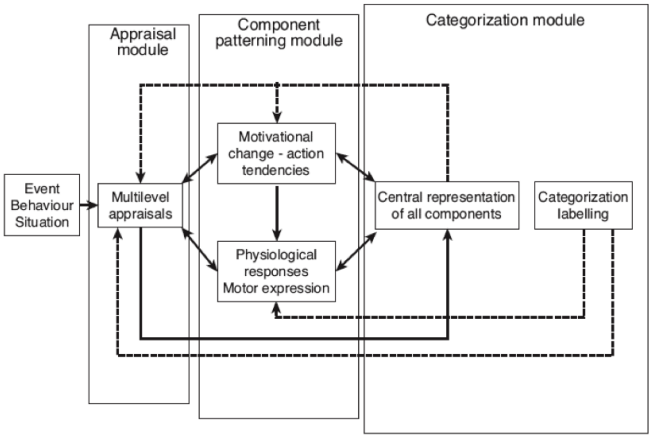

Fig. 1. Architecture of Scherer's Component Process Model of emotion (originally published in [24])

appraisal (evaluation of objects and events), physiological activation (system regulation), motivational tendencies (preparation and direction of action), motor expression (communication of reaction and behavioral intention), and subjective feeling state (monitoring of internal state and external environment). Figure 1 shows the general architecture of the CPM organized into three modules: appraisal, response patterning, and integration/categorization.

The CPM considers the experience of emotions as the result of the recursive multilevel sequential evaluation checking of an emotional stimulus or event. It is a response to the evaluation of a stimulus relevant to major concerns of the organism. The first module, *i.e.*, the appraisal module, is the most important element in the CPM. It determines whether an emotion episode is elicited and its characteristics are based on four major appraisal objectives [21]: Relevance Detection (scan for salient events requiring attention), Implication Assessment (estimation of implication, consequences), Coping Potential Determination (assessment of coping potential, *i.e.*, actions) and Normative Significance Evaluation (compatibility with social norms and self-concept). These objectives are evaluated based on stimulus evaluation checks (SECs) defined as novelty, intrinsic pleasantness, relevance to goals and needs, cause, probable outcomes, failure to meet expectations, conduciveness to goals and need, urgency, control, power, adjustment, internal standards, and external standards. The results of the SECs are highly subjective and are biased by individual differences, moods, cultural values, group pressures, or other context factors [25]. This is why a stimulus may evoke different emotions for different people at different places: it is the evaluation of the events, rather than the events themselves, which determines the characteristics of an emotional episode.

The result of the appraisal updates the motivational state that existed before the occurrence of the emotional stimulus or event. Both the appraisal results and motivational changes affect the automatic nervous system and the somatic nervous system. These components are all continuously updated as events and appraisals change. Scherer claims that, even if most of these components are examined unconsciously through parallel processing, some components may be evaluated consciously to allow more controlled regulation processes [24].

2.1.3 A psychological constructionist model: the Core Affect

Psychological constructionist models presume that emotions can be broken down into primitives that are also involved in other mental states [10]. While appraisal models assume that it is the evaluation of the stimulus that determines the characteristics of an emotional episode, psychological constructionist models assume that an emotion emerges when one's internal state is consciously understood in relation to an event.

Russell describes emotions as a part of a general component process called the Core Affect [26]. The core affect is a primitive, universal and simple neurophysiological state that controls moods, when core affect is experienced as free-floating, and emotions, when core affect can be attributed to some cause, whether based on reality or fiction. In the core affect model, raw feelings (the conscious experience) are a blend of two dimensions: pleasure–displeasure (called pleasure or valence) and activation–deactivation (called arousal or energy).

Psychological constructionist models suggest that an emotional episode is an event that counts as a member of an emotion category. The prototypical emotional episode begins with an antecedent event perceived in terms of its affective quality. The core affect is altered and attributed to the antecedent to enable perceptual cognitive processing of the antecedent with various manifestations: instrumental action, physiological and expressive changes, subjective conscious experiences, and emotional meta-experience. The emotional meta-experience is a self-perception: the emotional episode is noticed by the person who categorizes the emotional episode. Emotion categories are subjective since they are mental representation of emotions. For Russell, emotions respond to the continuous flow of events (whether based on reality or fiction) and are influenced at the same time by the background environment (weather, odors, noise) and the social environment [27].

2.1.4 Summary of the main psychological approaches

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

All three approaches emphasize that emotions are constructed from universal basic biological or psychological parts. However, psychological constructionist and appraisal models differ from basic models insofar as they assume that the social context of the situation and/or cultural differences have an effect on the experience of emotions [10].

Psychological constructionist models and appraisal models both consider emotion as an act of making meaning. However, the meaning analysis is different for each approach. Psychological constructionist models, including Russell's Core Affect [26], assume that an emotion arises when a person's internal state is understood. For appraisal models, emotions are intentional states created by the evaluation of an original stimulus, and not by the internal state of the body. The internal state is only affected by this meaning analysis.

2.2 Neuroscience

Related to the marked improvement in techniques, the neuroscience field has also contributed to the understanding of emotion mechanisms over the last ten years. The findings of specific brain structures and neural networks can be credited to new methods and experimental paradigms in brain imaging techniques. As an example, while electrophysiological methods, such as EEG or ERP, provided high temporal sampling rates (roughly at a millisecond), functional brain imaging methods are able to provide spatial details at a high resolution. Both have allowed a better understanding of when and where emotions occur [28].

By essence, the neuroscience field and its findings oppose experimental psychology by proving that emotional brain mechanisms are typically unconscious. These involuntary processes entail behavioral tendencies thus generating conscious feelings that interact back with emotion [29]. Even if emotional mechanisms originate in the brain, they are expressed and transmitted through the subject body as electrical and/or hormonal activities.

The first theories on brain modeling emerged at the end of the 19^{th} century. As a precursor in the field, James assumed that emotion is created by means of the body's reaction to an external stimulus [30]. Despite being controversial, this theory inspired a number of scientists whose grail was to locate the centrum for emotional perception. Several experiments during the 20^{th} century validated the assumptions that the thalamus is the centrum for emotional perception, while the hypothalamus conciliates emotional expressions [31]. Alternatively, Papez [32] demonstrated a more complex structure with a neural circuit in several brain areas. In addition to the thalamus and hypothalamus, the mamillary bodies, the cingular gyrus and the hippocampus were also concerned by emotional mechanisms. Nowadays, it is universally admitted that the limbic system, enclosing the thalamus, hippocampus, amygdala, hypothalamus and anterior cingulate cortex, is the major brain feature dealing with emotion.

Based on such experiments and deep analysis of brain functions, some computational models were designed in the neuroscience community [33]. Such models reflect the way in which emotional mechanisms are interpreted by authors. Thus, Grossberg and Schmajuk [34] designed a model of conditioned affective states based on antagonistic neural processes, for example fear and relief. They considered the previous state of the model to evaluate the new value of the current emotional event. They simulated the neural circuits involved in emotion elicitation through the different brain areas (*e.g.*, amygdala), which they identified as being stakeholders in emotion expression. Emotional activity proved to be difficult to isolate from other functions. Also, recent research specifically associated emotion with memory [35], attention [36] and decision-making [37]. These other functions are potentially a means of refining the existing computational models and of creating more collaboration between neuroscience and image processing communities.

2.3 Aesthetic emotions

The theories introduced in the previous sections model emotions regardless of the significant stimulus that provoked the emotions. This stimulus could be either a natural phenomenon, the behavior of other people or animals, or even one's own behavior [38]. However, this paper is concerned with the emotions that are provoked by movies. In this section we focus on a particular type of emotions, namely aesthetic emotions, for the reasons explained below.

2.3.1 Definition and models

An aesthetic emotion is an emotional response to a work of art (paintings, pictures, but also songs, movies, *etc*.) which can be described by three characteristics [39]:

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

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

Linking aesthetics and emotion, Leder *et al.* [40] proposed a model of aesthetic and emotional experience of modern art in an artistic context. This model is composed of five main stages resulting in two distinct outputs: an aesthetic judgement which is the result of the evaluation of the cognitive mastering stage and an aesthetic emotion based on the affective state of satisfaction. Like Scherer, the model proposed by Leder postulates both automatic and deliberate processing.

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

However, the Leder's model is too simplified and only considers aesthetic experience as affectively positive and self-rewarding (successful processing). It does not take into account the negative aspect, the antinomy, and the variety of art experience [41]. Several works extended Leder's model to offer a more precise model of aesthetic processing. For example, Marković created a model in which the aesthetic experience is closer to arousal, *i.e.*, the interest for the work of art, than to other dimensions of subjective experience [39].

In this multimodal model, the narrative content (*e.g.*, story, symbolism) and the composition form (*e.g.*, colors, gestures, sounds) both influence the aesthetic emotion, films being more focused on the narrative content.

2.3.2 A sociological perspective

Aesthetic emotions, and, more specifically, emotions provoked by movies, can also be analyzed from a sociological perspective. Hochschild was one of the first to study the sociology of emotions and how shared norms can influence the way we want to try to feel emotions in given social relations [42].

Wiley discussed the differences between the emotions people experience watching a movie, called "movie emotions" and those occurring in our everyday life [43]. He explained that viewers split attention between the movie and the physical environment, which helps regulate distance from the movie and creates an aura of safety and that the effects of movie emotions tend to end with the movie or at least decrease. Furthermore, the viewer is not the subject of the movie emotions, they happen because an identification is built with the character. For Wiley, movie emotions come with clear labels since narratives are written with precise emotional scripts. They are included in the movie with dialogues, clearly structured situations, transparent tendencies and musical cues. Movie emotions can be anticipated since films tend to follow the usual stability-instability-stability rule and because the music is geared to place the viewer in the appropriate emotional channel. These arguments show that it makes sense to investigate computational models based on multimodal features to predict the emotions provoked by movies.

For Wiley, movie emotions are desired: in watching a movie, the viewer wants to feel frequent, dense and almost wall-to-wall emotions [43]. This may result, for example, from a need to escape boredom or to forget the day's upsets. Movie emotions are also quite intense: viewers deal with more emotions than in a comparable period of time of a typical day. Movie emotions can be increased with social aspects (co-viewers' laughs). It is also easier to admit and talk about movie emotions because they are just fantasy so nobody can be blamed for having them.

2.3.3 Types of emotional processes in response to multimedia

Three types of emotional processes in response to multimedia are described in previous work: emotion induction, emotion contagion, and empathic sympathy [44].

Induced emotions are the emotions that viewers feel in response to a multimedia content with respect to their goals and values. For example, the animated movie "Big Buck Bunny"[3] features an evil squirrel tearing up a butterfly. The situation is likely to elicit negative emotions from viewers (disgust, anger, *etc.*), although the imaginary squirrel is enjoying the situation. The negative response from viewers, *i.e.* the induced emotion, is due to their perception of the context according to their goals and values biased by the identification built with one or more characters of the movie.

3. Big Buck Bunny, licensed under the Creative Commons Attribution 3.0 license, (c) copyright 2008, Blender Foundation, www.bigbuckbunny.org

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

Finally, emotions, as defined above, have to be distinguished from other affective phenomena such as moods. Moods are diffuse affect states generally of low intensity, may last for hours or even days, and are often not clearly linked to an event or specific appraisals [38].

2.4 Representations of emotions

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

2.4.1 Categorical

The categorical emotions approach, also commonly referred to as the discrete emotions approach, is very natural since it goes back to the origin of language and the emergence of words and expressions representing clearly separable states. Many discrete categorizations of emotions have been proposed. The basic emotions introduced in Section 2.1.1, such as the six basic universal emotions proposed by Ekman [15], or the eight primary emotions defined by Plutchik [14], are often used in the literature to represent emotions categorically.

This categorical representation is used in many affective movie content analysis work but faces a granularity issue since the number of emotion classes is too small in comparison with the diversity of emotions perceived by film viewers. For example, the list of the six basic emotions introduced by Ekman is criticized because it contains only one positive emotion [45]. If the number of classes is increased, ambiguities due to language difficulties or personal interpretation appear.

2.4.2 Dimensional

Dimensional approaches have also been proposed to model emotions as points in a continuous n-dimensional space. The most famous one is the valence-arousal-dominance space, also known as the pleasure-arousal-dominance (PAD) space, introduced by Russell and Mehrabian [46] and extensively used in research dealing with affective understanding. In this space, each subjective feeling can be described by its position in a three-dimensional space formed by the dimensions of valence, arousal, and dominance. Valence ranges from negative (*e.g.*, sad, disappointed) to positive (*e.g.*, joyous, elated), whereas arousal can range from inactive (*e.g.*, tired, pensive) to active (*e.g.*, alarmed, angry), and dominance ranges from dominated (*e.g.*, bored, sad) to in control (*e.g.*, excited, delighted). Given the difficulty of consistently identifying a third dimension (such as dominance, tension or potency) which differs from arousal, many studies, including this work, limit themselves to the valence and arousal

(VA) dimensions. Indeed, especially when dealing with emotions elicited by videos, valence and arousal account for most of the independent variance [47], [48]. Moreover, psychophysiological experiments have revealed that only certain areas of this two-dimensional space are relevant [49], [50] and that emotions induced by media can be mapped onto a parabolic space created by the arousal and valence axes.

However, this common two-dimensional space is questioned. Fontaine *et al.* demonstrated that two dimensions are not sufficient to represent emotions satisfactorily [51]. They showed that use of at least four dimensions is more appropriate to represent the diversity of emotions (valence, arousal, dominance, and predictability) but that the optimal number of dimensions to be included in a model depends on the purpose of the model. More recently, Joffily and Coricelli [52] proposed a formal definition of emotional valence in terms of the negative rate of change of free-energy over time. They also investigated the relationship between the dynamics of free-energy and some basic forms of emotion such as happiness, unhappiness, hope, fear, relief, disappointment, and surprise. Thus, they claim that free-energy should be considered in order to create a biologically plausible computational model of emotional valence

3 Computational Models and Datasets for Affective Video Content Analysis

Affective video content analysis aims at automatically predicting the emotions elicited by videos. Work on affective video analysis can be categorized into two subgroups: continuous[4] affective video content analysis, which estimates a time-dependent affective score for consecutive portions (*e.g.* each frame or group of frames) of a video, and global[5] affective video content analysis, which assigns an affective score to an entire video. Some work represents emotions in the 2D valence-arousal space or in the 3D valence-arousal-dominance space, while other work represents emotions using discrete categories. Furthermore, the models are sometimes dedicated to specific video categories, *i.e.* music videos or a particular movie genre.

As stated in the introduction, studies on emotion assessment using physiological signals beyond audiovisual (AV) features, *i.e.*, implicit video affective content analysis, are not presented in this paper so as to focus on affective video content analysis work.

3.1 Affective video content analysis: the perspectives

The video investigated with respect to its affective content can be either the stimulus eliciting emotions (known as "affective movie content analysis") or a tool for investigating the emotions expressed by agents (*i.e.*, "video emotion recognition"). While the focus of this paper is affective movie content analysis, research in both affective video content analysis fields will be presented in this section as they are closely related, with emphasis on previous affective movie content analysis work.

Video emotion recognition work aims at automatically estimating the emotion expressed by an agent from a video recording, in relation to an emotion induced by a stimulus. Sometimes, these emotions are not real but played by an actor [53]. The main goal of video emotion recognition models is to enable affective interaction between human beings and computers. Such models are inherently different from affective movie content analysis models. However, temporal modeling of the emotions of such models may be useful for designing new continuous affective movie content analysis frameworks.

Affective movie content analysis work focuses on the video, which is the stimulus of the emotion to be investigated. The emotion to be investigated, related to the intended, induced and expected emotion defined in the introduction, is defined by the perspective of the models and their applications.

Due to the exciting new possibilities offered by such affective computing techniques, they can be naturally applied to help standard multimedia systems [54]. Thus, affective movie content analysis work has a large number of applications, including mood based personalized content recommendation [55] or video indexing [56], and efficient movie visualization and browsing [57]. Beyond the analysis of existing video material, affective computing techniques can also be used to generate new content, *e.g.*, movie summarization [58], or personalized soundtrack recommendation to make user-generated videos more attractive [59]. Affective techniques have even been used to enhance user engagement with advertising content by optimizing the way ads are inserted inside videos [60].

3.2 Continuous affective video content analysis

This section presents previous work on continuous affective video content analysis, including movie content analysis and video emotion recognition. To predict or classify emotions, previous work either directly combines linearly audiovisual (AV) features extracted from the data, or uses machine learning models. Temporal information can be included in the machine learning models, for example using Long Short-Term Memory neural networks, or by simply applying a temporal smoothing to the predicted values. A summary is given in Table 1.

3.2.1 Continuous valence and arousal movie content analysis

Hanjalic and Xu pioneered in [61] analysis of affective movie content by directly mapping video features onto the valence-arousal space to create continuous representations. Based on film theorists' work, they selected low-level features that are known to be related to arousal or valence, such as motion intensity, shot lengths, and audio features (loudness, speech rate, rhythm, *etc.*). They manually designed the functions modeling arousal and valence for consecutive frames based on the selected features and used a Kaiser window to temporally smooth the resulting curves. However, they only offered a qualitative evaluation of their model.

4. The term *continuous* is used in this paper to refer to time-dependent annotations and should not be confused with the *dimensional* annotations referring to the dimensional representation of emotions.

5. Global works are also commonly referred to as *discrete* affective movie content analysis. However, the term *global* was preferred in this paper because the term *discrete* also refers to the categorical representation of emotions.

TABLE 1
Summary of previous work on continuous affective movie content analysis and video emotion recognition

Authors	Method	Output	Ground truth	Annotators	Result
Hanjalic and Xu [61]	No classification: AV features are linearly combined	Continuous induced arousal and valence functions at a frame level	Unknown number of movie scenes and soccer television broadcasts	None	Qualitative evaluation only
Soleymani *et al.* [62]	Bayesian framework using AV and textual features, relying on temporal priors	Videos categorized into three induced emotional classes	21 full-length popular movies	1	Accuracy: 64%, F1 measurement: 63%
Malandrakis *et al.* [6]	Two HMMs using AV features at frame level	Time series of 7 categories interpolated into continuous intended VA curves	30-min video clips from 12 movies	7	Correlation for arousal: 0.54, and valence: 0.23
Nicolaou *et al.* [63]	LSTM-RNN-based framework using facial expression, shoulder gesture, and audio cues	Continuous recognition of emotions expressed by actors in terms of arousal and valence	10 hours of footage from the SAL database capturing the interaction between a human and an operator	4	Correlation for arousal: 0.642, and valence: 0.796
Kahou *et al.* [64]	Combination of multiple deep neural networks for different data modalities	Emotions expressed by the main actor in a video among 7 emotional classes	Short video clips extracted from movies, provided for EmotiW 2013	2	Accuracy: 41.03%

Soleymani *et al.* introduced a Bayesian framework for video affective representation [62] using audiovisual features and textual features extracted from subtitles but also taking into account contextual information (*e.g.* user's personal profile, gender or age). However, as their ground truth was annotated by a single participant only, they did not study the relevance of such contextual information and assumed the model to be personalized for this participant. Arousal estimation, obtained for each shot by means of a Relevance Vector Machine (RVM), is then used as an arousal indicator feature and merged with other content-based features for global scene affective classification thanks to a Bayesian framework. Thus, their framework is a trade-off between continuous and global affective video content analysis. The Bayesian framework relies on two priors: the movie genre prior and the temporal dimension prior consisting of the probability transition between emotions in consecutive scenes. However, movie scenes are categorized into three emotional classes, which is too restrictive. Furthermore, they only provided a qualitative evaluation of the continuous arousal estimation but achieved an accuracy of 63.9% for classification of movie scenes among three emotional classes.

Malandrakis *et al.* also proposed in [6] a continuous affective video content analysis relying on audiovisual features extracted on each video frame, combined at an early stage and used by two Hidden Markov Models (HMMs). These two classifiers are trained independently to model simultaneously the intended arousal and valence. However, HMMs predict discrete labels. Arousal and valence are thus discretized into seven categories and the model only allows transitions between adjacent categories. Finally, they output time series composed of seven categories interpolated into a continuous-valued curve via spline interpolation. However, the continuous curves are thus approximations and cannot recover the precision lost by discretizing the affective space. Their discrete and continuous curves are compared using the leave-one-movie-out approach to the ground truth collected on 30-min video clips from 12 movies. The smoothed predicted curves achieved an average correlation of 0.54 for arousal and 0.23 for valence.

3.2.2 Video emotion recognition

As stated in Section 3.1, video emotion recognition models are inherently different from affective movie content analysis models. While affective movie content analysis models aim at estimating the intended, expected or induced emotion from the intrinsic properties of a movie, video emotion recognition models aim at automatically estimating the emotion expressed by an agent from a video recording, in relation to an emotion induced by a stimulus. Emotion recognition models thus typically rely on facial characteristics such as action units (AU) [65]. Since felt and expressed emotions are linked, the temporal modeling of emotions of emotion recognition models may be of interest for designing new continuous affective movie content analysis frameworks. This is especially true since, as mentioned in Section 2, psychologists suggest that the evaluation of an emotion is a recursive and continuous process.

Nicolaou *et al.* introduced in [63] a framework for continuous prediction of spontaneous affect in the VA space based on facial expression, shoulder gestures, and audio cues. They compared the performance of the bidirectional Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) and Support Vector Machines for Regression (SVR) for continuous spontaneous affect prediction and proposed an output-associative prediction framework. This framework is a modified bidirectional LSTM-RNN taking into account the correlation between the predicted valence and arousal dimensions: it depends on the entire sequence of intermediate output predictions of both dimensions to perform the prediction. They showed that the bidirectional LSTM-RNN outperforms the SVR and that the output-associative prediction framework significantly improves prediction performance. However, to train and evaluate their models, Nicolaou *et al.* used recordings made in a lab setting, using a uniform background and constant lighting conditions. Their model is thus highly sensitive to recording conditions, such as illumination and occlusions.

More recently, Kahou *et al.* [64] designed a framework to assign one of seven acted-out emotions to very short video clips (1 to 2 seconds long) extracted from Hollywood

movies. It was the winning submission in the 2013 Emotion Recognition in the Wild Challenge [66]. Unlike [63], videos provided for the challenge depict acted-out emotions under realistic conditions (large degree of variation in attributes such as pose and illumination). Their framework combines a Convolutional Neural Network (CNN) focusing on capturing visual information in detected faces, a Deep Belief Network for representation of the audio stream, a K-Means based model for extracting visual features around the mouth region, and a relational autoencoder to take into account the spatio-temporal aspects of videos. To efficiently train the CNN, they downloaded two large alternative image databases of facial expressions for seven emotion categories. The final result is a concatenation of models based on a single modality. However, due to the characteristics of the data provided for the challenge, emotional relationships between consecutive video segments are not investigated.

3.3 Global affective movie content analysis

This section focuses on previous work on global affective video content analysis, including valence and arousal estimation but also violence detection. Global work assigns an affective score to an entire video. A summary is given in Table 2.

3.3.1 Global valence and arousal movie content analysis

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

Kang [67] was the first to propose a model where classifiers are adopted for affective analysis. He suggested detecting affective states in movies including "sadness", "joy" and "fear" from low-level features using HMMs. HMM topology is designed such that the only possible transitions between consecutive affective states are those between the neutral state and the other affective states. However, this topology is very restrictive and is unrealistic.

Wang and Cheong introduced features inspired from psychology and film-making rules [68]. One Support Vector Machine (SVM) is especially dedicated to audio cues to obtain high-level audio information at scene level. Each video segment is then classified with a second SVM to obtain probabilistic membership vectors for seven discrete emotional states. Their training data are made up of 36 full-length popular Hollywood movies divided into 2,040 scenes labeled with one or two emotional states by only three annotators. Due to the limited number of annotators, ambiguities arise and make it necessary to assign two labels to some videos. Furthermore, they do not take into account the relations between consecutive scenes.

In the work of Sun and Yu [69], movie units are first represented in different granularities using an excitement curve based on the arousal curve introduced in [61]. Then, four HMMs are trained independently using features extracted on these granularities to recognize one of the four emotional states among "joy", "anger", "sadness" and "fear". Each HMM has the same topology and is composed of two states: a neutral state, and a state representing the emotion assigned to the HMM. Thus, each HMM only computes for a given observation sequence, the state transition probabilities between the neutral state and one of the four emotional states. As in [67], this topology is restrictive and unrealistic. Their ground truth consists of 10 movies labeled at different levels by 30 annotators. Xu *et al.* used a similar approach in [70] sharing the same disadvantages. However, to compute the emotion intensity level they used fuzzy clustering instead of linearly combining audiovisual features. It is a fact that fuzzy clustering is closer to human perception. Five HMMs were then trained using emotion intensity and low-level features to model five emotional classes with different levels of valence. They evaluated the efficiency of their method for several movie genres, where the highest accuracy was obtained for action movies.

Soleymani *et al.* [71] compared in the VA space the values obtained automatically from either physiological responses or from audiovisual features. They showed significant correlations between multimedia features, physiological features and spectators' self-assessments for both valence and arousal. Affective scores are estimated by a linear combination of content-based features and compared to the estimation of affective scores using a linear combination of physiological features only. They showed that the performance of both models is the same and thus that none has a significant advantage over the other. A dataset composed of 64 movie scenes extracted from 8 Hollywood movies was created, from which 42 movie scenes were randomly selected for the training set, while the remaining ones were used in the test set. Movie scenes extracted from the same movie can thus be part of the training set and also of the test set, questioning the reliability of the results.

Zhang *et al.* developed in [56] a personalized affective analysis for music videos composed of SVR-based arousal and valence models using both multimedia features and user profiles. In fact, two nonlinear SVRs are learned for each user, taking into account the underlying relationships between user's affective descriptions and the extracted features. However, SVRs are not able to model the temporal transition characteristics of emotions. Their dataset of 552 music videos is used to train and update the models based on user feedback.

Irie *et al.* [72] proposed an approach based on Latent Dirichlet Allocation (LDA) considering the temporal transition characteristics of emotions. Emotion-category-specific audiovisual features are extracted and transformed into affective audio-visual words using k-mean clustering. These higher level features are then used to classify movie scenes using a latent topic driving model. The model considers the probability of emotional changes between consecutive scenes based on the Plutchik's wheel [14]. However, these probabilities have not been estimated from a real dataset but empirically defined by the authors. The rate of agreement of their model is equal to 85.5%. The good results obtained by their framework may be due to the evaluation protocol. Their data, composed of 206 scenes from 24 movie titles available as DVDs, were randomly selected to form the training and test sets. Consequently, most films appear both in the training and in the test sets as in [71], which biases the results.

More recently, Acar *et al.* proposed the use of CNNs in order to learn mid-level representations for affective classification of music video clips [73]. Two CNNs are learned to output mid-level representation of 5-second music video

TABLE 2
Summary of previous work on global affective movie content analysis and violence detection

Authors	Method	Output	Ground truth	Annotators	Result
Kang [67]	HMMs relying on visual features only	Discrete labels among 3 classes (fear, sadness, joy)	Scenes extracted from six 30min videos	10	Accuracy: ≈79%
Wang and Cheong [68]	Two SVMs using audio cues or AV features	Video scenes classified using 7 emotional classes	2,040 scenes from 36 full-length popular movies	3	Accuracy: 74.69%
Sun and Yu [69]	4 HMMs using AV features	Labels among 4 classes (anger, fear, sadness, joy)	10 popular movies labeled at different levels	30	Precision: ≈68%, Recall: ≈79%
Xu *et al.* [70]	5 HMMs using AV cues	5 affective discrete classes	Videos from 24 movies	?	Accuracy: 80.7%
Soleymani *et al.* [71]	Personalized RVM using AV or physiological cues	1 global arousal and valence score per video	64 movie scenes from 8 popular movies	8	MSEs for each participant
Zhang *et al.* [56]	2 SVRs using both VA features and user profile	One valence and arousal score per music video	552 representative music videos	10 & 27	Performances for two applications
Irie *et al.* [72]	Latent topic driving model using VA features	Probabilities for 9 emotion categories for each scene	206 scenes from 24 movie titles available as DVDs	16	Agreement: 85.5%
Acar *et al.* [73]	Two CNNs and one SVM using AV features	1 of the 4 quadrants of the VA space for each video	Music videos from the DEAP dataset	32	Accuracy: 52.63%
Penet *et al.* [74]	2 Bayesian networks using temporal AV features	Violence probability for each video shot	MediaEval 2011 Affect Task corpus	7	False alarms and missed curves
Eyben *et al.* [75]	SVMs using temporal AV features	Confidence score to classify a video shot as violent	MediaEval 2012 Affect Task corpus	7	MAP@100: 0.398

clips: one uses as input audio features (MFCC), while the other uses as input one color channel, *i.e.*, red, green or blue color channel, of the resized frame in the middle of the video segment. The mid-level representations are then each used in a dedicated SVM, and their predictions fed to a final multi-class audiovisual SVM to output the category of the video clip (one of the four quadrants of the valence-arousal space). Their framework achieves an accuracy of 52.63% on the DEAP dataset [76].

3.3.2 Violence detection

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

Penet *et al.* introduced a framework using both multimodal and temporal information for violence detection [74]. For temporal integration, their model uses the features extracted from the investigated video excerpts but also contextual features extracted from the five previous and next video shots. However, it is not clear why they preferred fixed-length time windows computing features for 10 consecutive video shots. The audio features (audio energy, audio zero crossing rate, *etc.*) , and video features (shot duration, number of flashes, *etc.*) are separately used in two independent Bayesian networks. The two probabilities given by the networks are finally fused using a late fusion approach.

One year later, to classify video excerpts as violent or non-violent, Eyben *et al.* fused by simple score averaging the predictions made by acoustic and visual linear kernel SVMs [75]. To capture temporal dynamics, numerous statistics are computed for the features over fixed size windows. The authors also provided a detailed analysis of the features extracted from the audio and video modalities. They showed that some features are particularly relevant for violence detection: color and optical flow for the video modality, and spectral distribution descriptors and peak-based functional extraction for the audio channel.

While using short fixed-length time windows may be relevant for violence detection, estimating induced emotions requires consideration of longer-term dependencies [77].

3.4 Affective video datasets

Main affective video datasets and their limitations have already been reviewed in recent papers [78], [79]. They outlined that main public affective video datasets, namely the HUMAINE [80], FilmStim [81], DEAP [76], MAHNOB-HCI [82], EMDB [83], and VSD [84] datasets, suffer from limitations preventing their use in training and evaluating computational models in the field of affective content analysis. Indeed, these datasets are either of limited size and content diversity, suffer from copyright issues, or have different emotional labels that are sometimes not representative of the whole range of emotions in movies. However, some attempts are starting to emerge to solve these limitations.

Recently, the LIRIS-ACCEDE dataset has been released publiclyin order to overcome the limitations of the existing affective video datasets and to foster research in affective video content analysis [79]. This dataset has been used in particular in the MediaEval benchmarking initiative to compare computational models estimating the emotions elicited by movies [85]. Based on the FilmStim dataset [81], Guntuku *et al.* proposed the CP-QAE-I dataset to explore the effect of quality of experience, personality and culture on the felt emotions [86]. This dataset is composed of 144 videos, with various bit-rate, frame dimension, and frame rate, derived from 12 original video excerpts. Although limited in size and content diversity, the interest of this dataset lies in the affective self-assessments collected by participants with different cultural backgrounds. It is also interesting to mention the Audio/Visual Emotion Challenge

and Workshop (AVEC) [87], and the Emotion Recognition In The Wild Challenge (EmotiW) workshop [66], both consisting of recognition of the emotions expressed by actors in laboratory environments or in movies using large and public datasets. Such benchmarking campaigns are the key to building more reliable computational models in the future.

3.5 Issues

Due to the constraints on databases presented in Section 3.4, most state of the art work about affective movie content analysis uses a private dataset of a very limited size and content diversity, designed according to their goals and needs (see Tables 1 and 2). Thus, it makes fair comparisons and results reproducibility impossible, preventing achievement of major strides in the field. For example, some work represents emotions in the 2D VA space or in the 3D valence-arousal-dominance space [1], [56], while other work represents emotions using discrete categories [68]. Furthermore, the models are sometimes dedicated to specific video categories, *i.e.*, music videos [56], [73] or to a particular movie genre [88]. It is a fact that choice of representation of emotions is highly dependent on the model goal. For example, if a model targets end-users, natural categorical representation of emotions may be preferred over dimensional representation since the former is simpler to comprehend. On the contrary, continuous representation of emotions may be preferred, for example, for work modeling the transitions between consecutive emotions. Benchmarking previous work is also limited because affective movie content analysis work can be identified either as continuous or global approaches. The choice of one approach over the other is also biased by the model goal. For example, the global approach may be chosen for models dealing with video indexing considering a movie as a whole, while the continuous approach may be preferred for models requiring a fine-grained temporal resolution (*e.g.*, efficient movie visualization and browsing).

The difficulty in collecting reliable affective ground truth is the last issue discussed in this section. The term "ground truth" should be taken with a pinch of salt in affective video analysis work. Indeed, the affective self-assessments collected in previous work dealing with emotions induced by video content are intrinsically biased. This is because they represent the interpretation of the emotional experience that the annotator is currently feeling, which may be different from the emotion felt by the annotator (*i.e.*, the induced emotion). Recording the facial expression of the annotators and collecting physiological signals, such as skin conductance, heart rate, or even electroencephalogram signals, could help to improve the reliability of affective self-assessments. However, the correlation between such modalities and felt emotions is still ongoing [89] and there is no direct evidence that any particular facial movement provides an unambiguous expression of a specific emotion [90]. The continuous ground truth used in previous work (such as [6], [63]) is also biased because of the post-processing steps for taking into account the annotator-specific delays amongst the annotations, and because it is the result of the aggregation of the multiple annotators' self-assessments [91], [92], [93]. To tackle these issues, future work should start investigating individual emotional differences in order to design personalized models, as presented in the next section. To summarize, the "ground truth" used in most affective video analysis work, does not represent the emotions that an annotator has felt during an experiment, but rather represents the emotions that most annotators say they have experienced while watching movies during an experiment. Thus, these values correspond to the interpretation of the expected emotions. Such feedback, requiring a true introspection, have to be carefully considered [94].

4 Bringing psychological theories in computational models

The issues listed in the previous section are closely related to the machine learning-based architectures used to build the computational models estimating the emotions expected or induced by movies. However, psychological theories and computational models rely very little on each other. A Human-Computer Interaction (HCI) system would require implementation of psychological models while considering a large variety of interactions. However, computational models estimating the emotions elicited by movies form a special case where the stimulus is a video scene. Thus, only specific sections of the psychological models could be first investigated in order to maximize the performance of computational models. In this section, we share some insights for integrating ideas expressed in Section 2 by stating that future affective video content analysis work should:

- Integrate the recursive aspect of emotions.
- Emulate bottom-up and top-down processes.
- Predict intermediate dimensions.
- Personalize prediction.

We hope these insights will help build models closer to human perception and thus allow emotions to be modeled with greater reliability and fidelity. Furthermore, some insights are general ideas and are thus also relevant for computational models over and beyond affective movie content analysis.

First, psychologists suggest that evaluation of an emotion is an iterative process. For example, Russell defines the Core Affect as [26]:

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

This process of recursive and continuous evaluations is also the core of the appraisal evaluation [24] and of aesthetic emotions [40]. This is particularly important when dealing with movies because the emotions induced by such types of stimulus are closely linked to the narrative content, and thus to the temporal sequence of events within movies [39]. Furthermore, as stated in Section 2.3.2, the narrative content tends to follow typical rules, which could be learned by a recursive model taking into account the sequence of events. Thus, this key aspect of emotions has to be considered in affective computational models. However, most computational models do not take into account the recursivity of the emotional episode. This is not the case for HMM-based [67], [69], [70] and RNN-based [63] affective frameworks. Indeed,

HMMs are statistical models of sequential data, inherently able to take into consideration consecutive emotional changes through hidden state transitions. However, they are composed of a specific number of discrete states and thus cannot be used to directly infer dimensional scores. Malandrakis *et al.* converted discrete affective curves obtained with HMMs into continuous curves using a Savitzky-Golay filter [6]. However, the continuous curves are thus approximations and cannot recover the precision lost by discretizing the affective space. RNN-based affective frameworks are also able to take into account temporal transitions between consecutive emotions. Combined with LSTM cells, they can efficiently learn long-term dependencies, as shown by previous work on video emotion recognition [63], [95].

Beyond the recursive aspect of the models, their structures themselves can be inspired from neuroscience and psychological work. Indeed, psychological theories describe emotions as generated through the interaction of bottom-up and top-down processes (but with emphasis on the bottom-up process) [24], [40]. This is confirmed by neuroscience work demonstrating that both types of response activate the amygdala (bottom-up responses activating the amygdala more strongly) [96]. Bottom-up processes correspond to the elicitation of emotion in response to inherently emotional perceptual properties of the stimulus. On the other hand, top-down processes are cognitive evaluations occurring, for example, during the relevance to goals and needs SEC defined in appraisal models [21]. Bottom-up emotions help the subject to respond quickly and accurately to emotion-relevant aspects of the environment, whereas top-down emotions help achieve greater flexibility in producing these emotional responses [97]. Such descriptions of the interaction between bottom-up and top-down responses can help future affective movie content analysis work to create computational models with appropriate structures as in other fields of study [98]. For example, saliency models predicting where people are looking in pictures have greatly benefited from the integration of such structures by considering neuroanatomical and psychological work [99].

Psychological theories have also shown that several intermediate dimensions are important characteristics of the emotional experience. For example, Scherer has postulated that dimensions such as novelty (including suddenness, familiarity and predictability), as well as pleasantness, are important factors in the appraisal module for determining the characteristics of the emotional episode and are linked to attention and memory allocation [100]. Thus, predicting intermediate dimensions, such as predictability or novelty, may be of interest to help estimate valence and arousal elicited by movies. Recently, first attempts have been proposed in order to predict the memorability of images [101]. Other high-level intermediate factors can be inspired by Wiley [43] using a sociological perspective, introduced in Section 2.3.2. In particular, information obtained from movie scripts, such as relations among movie characters [102] as well as their personalities [103], could improve future affective movie content analysis work.

Predicting the emotions that most people feel while watching movies (*i.e.*, expected emotions) is extremely difficult to solve with a universal solution. This is because emotional experience is a highly subjective process. Consequently, personalized solutions should also be considered to predict induced emotions (*i.e.*, emotions that a specific viewer feels while watching a movie). This will be possible by collecting self-assessments across highly diverse viewer groups with different cultural and socio-demographical backgrounds. Personalized video affective models would be useful for applications such as personalized content delivery, while general video affective models predicting intended or expected emotions would be applied for video indexing or as a support to help create new video content. This implies the need for different types of ground truth depending on the model goal: annotations from video creators are necessary for evaluating predictions of intended emotion, annotations from individual viewers are necessary for evaluating predictions of induced emotion, and annotations from multiple viewers are necessary for evaluating predictions of expected emotion. Nowadays, most work only considers expected emotions and computes the ground truth as the mean of the individual affective self-assessments. Subjective factors such as previous experience, interest and personal taste play an important role in the experience of an emotion [40]. Future personalized solutions aiming to predict the induced emotions of a given user could thus use basic characteristics as additional features (*e.g.*, age, gender, nationality), but also the results of personality tests previously filled in by users (*e.g.*, measurements of the Big Five personality traits), or the list of the last played movies to take into account previous experience and interest. The environment (*i.e.*, the contextual information) is also a key element for reliably predicting induced emotions. Both the background environment and the social environment influence emotions [27], including for example co-viewers' laughs [43] or group pressures [25]. Despite the importance of such information in predicting induced emotions, currently no affective movie content analysis work analyzes the influence of personalized features.

Previous attempts have been proposed to personalize computational models. As mentioned in Section 3.2.1, Soleymani *et al.* introduced a personalized framework for video affective representation [62] taking into account contextual and personal information. Nevertheless, reliability of the personalization priors was not evaluated due to the difficulty of collecting large samples of individual affective self-assessments. More recently, affective datasets mentioned in Section 3.4, aiming at exploring the effects of personality and culture , are starting to emerge, but are still of limited size and content diversity. However, it is of great interest to start analyzing the effect of personality and culture on the perception of video content quality and affect. First results confirm that personality and culture play an important role and have to be taken into account in future computational models [104], [105]. Bigger affective datasets, composed of self-assessments across highly diverse viewer groups with different cultural and socio-demographical backgrounds, are needed to generalize and confirm these findings. These would pave the way for the next generation of personalized computational models of induced emotions.

5 Conclusion

In this paper, we gave an overview of emotion theories and emotion representations, as well as computational models

for affective video content analysis.

A review of previous work developing computational models to infer emotions showed that they rely on a predefined set of highly problem-dependent handcrafted features requiring strong domain knowledge and that they use private datasets to evaluate their performance, thus making fair comparisons and result reproducibility impossible, and preventing achievement of major strides in the field. Computational models trained using private datasets are often not generalizable and fail when applied to new videos due to the limited size and content diversity of the datasets used for training. Bigger datasets and experimental evaluations on new contents are thus needed to create more robust and generalizable computational models.

Furthermore, previous affective multimedia content analysis work should be designed to be closer to psychological models. In particular, they should model emotions recursively using intermediate dimensions such as predictability or novelty, structured using insights from the neuroscience and vision science fields. Finally, to reliably predict induced emotions, future computational models will need to be personalized in order to better model the individual differences among viewers and to take into account the environment, which plays a major role during emotional processes. Without such personalized solutions, future computational models will fail to predict induced emotions reliably.

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

ACKNOWLEDGMENTS

This work was supported in part by the French Research Agency ANR through the VideoSense project under the grant 2009 CORD 026 02 and the Visen project within the CHIST-ERA program under the grant ANR-12-CHRI-0002-04.

REFERENCES

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

[2] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *Proc. 19th ACM Int. Conf. on Multimedia*, 2011, pp. 1473–1476.

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

[4] P. Philippot, "Inducing and assessing differentiated emotion-feeling states in the laboratory," *Cognition & emotion*, vol. 7, no. 2, pp. 171–193, 1993.

[5] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[6] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2376–2379.

[7] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan 2010.

[8] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. on Affective Computing*, vol. 6, no. 4, pp. 410–430, Oct 2015.

[9] P. R. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion*, vol. 5, no. 4, pp. 345–379, Dec. 1981.

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

[11] C. Darwin, P. Ekman, and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[12] P. Ekman and W. Friesen, "Em-facs coding manual," *San Francisco: Consulting Psychologists Press*, 1984.

[13] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, pp. 384–392, Apr 1993.

[14] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Academic Press New York, 1980, ch. 1, pp. 3–31.

[15] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd, 2005, ch. 3, pp. 45–60.

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

[17] P. Ekman and K. G. Heider, "The universality of a contempt expression: A replication," *Motivation and emotion*, vol. 12, no. 3, pp. 303–308, Sept 1988.

[18] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological review*, vol. 97, no. 3, pp. 315–331, Jul 1990.

[19] C. Izard, F. Dougherty, B. Bloxom, and W. Kotsch, "The differential emotions scale: A method of measuring the subjective experience of discrete emotions," *Vanderbilt University, Department of Psychology, Nashville, TN*, 1974.

[20] N. H. Frijda, *The emotions*, ser. Studies in Emotion and Social Interaction. Cambridge University Press, 1986.

[21] K. R. Scherer, "Appraisal considered as a process of multi-level sequential checking," in *Appraisal processes in emotion: Theory, Methods, Research*, K. R. Scherer, A. Schorr, and T. Johnstone, Eds. Oxford University Press, 2001, ch. 5, pp. 92–120.

[22] M. B. Arnold, *Emotion and personality.* Columbia University Press, 1960.

[23] R. S. Lazarus, *Psychological stress and the coping process.* McGraw-Hill, 1966.

[24] K. R. Scherer, "The component process model: Architecture for a comprehensive computational model of emergent emotion," in *Blueprint for affective computing: A sourcebook*. Oxford University Press, 2010, ch. 2.1, pp. 47–70.

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

[26] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological Review*, vol. 110, no. 1, pp. 145–172, Jan 2003.

[27] J. Russell and J. Snodgrass, "Emotion and the environment," in *Handbook of environmental psychology*, 1987, pp. 245–280.

[28] A. Deak, "Brain and emotion: Cognitive neuroscience of emotions," *Review of Psychology*, vol. 18, no. 2, pp. 71–80, 2011.

[29] D. D. Franks, "The neuroscience of emotions," in *Handbook of the Sociology of Emotions*. Springer, 2006, ch. 2, pp. 38–62.

[30] W. James, "What is an emotion?" *Mind*, vol. 9, no. 34, pp. 188–205, Apr 1884.

[31] J. Morris and R. Dolan, "Functional neuroanatomy of human emotion," in *Human Brain Function (Second Edition)*. Academic Press, 2004, pp. 365–396.

[32] J. W. Papez, "A proposed mechanism of emotion," *Archives of Neurology & Psychiatry*, vol. 38, no. 4, pp. 725–743, Oct 1937.

[33] J.-M. Fellous, J. L. Armony, and J. E. LeDoux, "Emotional circuits and computational neuroscience," in *Handbook of brain theory and neural networks*. MIT Press, USA, 2002, vol. 2.

[34] S. Grossberg and N. A. Schmajuk, "Neural dynamics of attentionally modulated pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing," *Psychobiology*, vol. 15, no. 3, pp. 195–240, Oct 1987.

[35] S. Hamann, "Cognitive and neural mechanisms of emotional memory," *Trends in cognitive sciences*, vol. 5, no. 9, pp. 394–400, Sept 2001.

[36] T. Brosch and J. J. Van Bavel, "The flexibility of emotional attention: Accessible social identities guide rapid attentional orienting," *Cognition*, vol. 125, no. 2, pp. 309–316, Nov 2012.

[37] T. Brosch and D. Sander, "Appraising value: The role of universal core values and emotions in decision-making," *Cortex*, vol. 59, pp. 203–205, Oct 2014.

[38] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, Dec. 2005.

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

[40] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *British Journal of Psychology*, vol. 95, no. 4, pp. 489–508, Nov. 2004.

[41] L.-H. Hsu, "Visible and expression. study on the intersubjective relation between visual perception, aesthetic feeling, and pictorial form," Ph.D. dissertation, Ecole des Hautes Etudes en Sciences Sociales (EHESS), Jun. 2009.

[42] A. R. Hochschild, "Emotion work, feeling rules, and social structure," *American Journal of Sociology*, vol. 85, no. 3, pp. 551–575, Nov 1979.

[43] N. Wiley, "Emotion and film theory," in *Studies in Symbolic Interaction*. Emerald Group Publishing Limited, 2003, vol. 26, pp. 169–190.

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

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

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

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

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

[49] R. B. Dietz and A. Lang, "Aefective agents: Effects of agent affect on arousal, attention, liking and learning," in *Cognitive Technology Conf.*, 1999.

[50] A. Kron, M. Pilkiw, J. Banaei, A. Goldstein, and A. K. Anderson, "Are valence and arousal separable in emotional experience?" *Emotion*, vol. 15, no. 1, pp. 35–44, Feb 2015.

[51] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, Dec. 2007.

[52] M. Joffily and G. Coricelli, "Emotional valence and the free-energy principle," *PLoS Comput Biol*, vol. 9, no. 6, 06 2013.

[53] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. on Multimodal Interfaces*, 2004, pp. 205–211.

[54] R. W. Picard, "Affective computing: From laughter to ieee," *IEEE Trans. on Affective Computing*, vol. 1, no. 1, pp. 11–17, Jan 2010.

[55] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 636–647, Aug 2013.

[56] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.

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

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

[59] R. R. Shah, Y. Yu, and R. Zimmermann, "Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings," in *Proc. 22nd ACM Int. Conf. on Multimedia*, 2014, pp. 607–616.

[60] K. Yadati, H. Katti, and M. Kankanhalli, "Cavva: Computational affective video-in-video advertising," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 15–23, Sept 2014.

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

[62] M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *3rd Int. Conf. on Affective Computing and Intelligent Interaction*, Sep. 2009, pp. 1–7.

[63] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[64] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. on Multimodal Interaction*, 2013, pp. 543–550.

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

[66] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge (emotiw) challenge and workshop summary," in *Proc. 15th ACM Int. Conf. on multimodal interaction*, 2013, pp. 371–372.

[67] H.-B. Kang, "Affective content detection using HMMs," in *Proc. 11th ACM Int. Conf. on Multimedia*, 2003, pp. 259–262.

[68] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[69] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *2nd Int. Conf. on Affective Computing and Intelligent Interaction*, 2007, pp. 594–605.

[70] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proc. 16th ACM Int. Conf. on Multimedia*, ser. MM '08, 2008, pp. 677–680.

[71] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *IEEE Int Symp. on Multimedia*, Dec. 2008, pp. 228–235.

[72] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[73] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *20th Int. Conf. MultiMedia Modeling*, 2014, pp. 303–314.

[74] C. Penet, C. Demarty, G. Gravier, and P. Gros, "Multimodal information fusion and temporal integration for violence detection in movies," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2393–2396.

[75] F. Eyben, F. Weninger, N. Lehment, B. Schuller, and G. Rigoll, "Affective video retrieval: Violence detection in hollywood movies by large-scale segmental feature extraction," *PloS one*, vol. 8, no. 12, Dec 2013.

[76] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: a database for emotion analysis using physiological signals," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

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

[78] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, Jun. 2014.

[79] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. on Affective Computing*, vol. 6, no. 1, pp. 43–55, Jan 2015.

[80] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *2nd Int. Conf. on Affective Computing and Intelligent Interaction*, 2007, pp. 488–500.

[81] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, Nov. 2010.

[82] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

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

[84] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "A benchmarking campaign for the multimodal detection of violent scenes in movies," in *Proc. 12th Int. Conf. on Computer Vision*, ser. ECCV'12, 2012, pp. 416–425.

[85] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen, "The mediaeval 2015 affective impact of movies task," *MediaEval*, 2015.

[86] S. Guntuku, M. Scott, H. Yang, G. Ghinea, and W. Lin, "The cp-qae-i: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia," in *7th Int. Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–7.

[87] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 3–10.

[88] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2005.

[89] M. Soleymani, S. Asghari Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Trans. on Affective Computing*, vol. 7, no. 1, pp. 17–28, Jan 2016.

[90] B. Parkinson, "Do facial movements express emotions or communicate motives?" *Personality and Social Psychology Review*, vol. 9, no. 4, pp. 278–311, Nov 2005.

[91] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

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

[93] M. Nicolaou, V. Pavlovic, M. Pantic *et al.*, "Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1299–1311, July 2014.

[94] R. E. Nisbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes." *Psychological review*, vol. 84, no. 3, pp. 231–259, Mar 1977.

[95] Y. Baveye, "Automatic prediction of emotions induced by movies," Ph.D. dissertation, Dept. Comput. Sci., Ecole Centrale de Lyon, Nov. 2015.

[96] K. N. Ochsner, R. R. Ray, B. Hughes, K. McRae, J. C. Cooper, J. Weber, J. D. Gabrieli, and J. J. Gross, "Bottom-up and top-down processes in emotion generation common and distinct neural mechanisms," *Psychological science*, vol. 20, no. 11, pp. 1322–1331, Nov 2009.

[97] K. McRae, S. Misra, A. K. Prasad, S. C. Pereira, and J. J. Gross, "Bottom-up and top-down emotion generation: implications for emotion regulation," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 3, pp. 253–262, Feb 2012.

[98] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, Dept. Biology, California Institute of Technology, Pasadena, 2000.

[99] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 438–445.

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

[101] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1469–1482, July 2014.

[102] L. Ding and A. Yilmaz, "Learning relations among movie characters: A social network perspective," in *11th European Conference on Computer Vision*, 2010, pp. 410–423.

[103] D. Bamman, B. O'Connor, and N. A. Smith, "Learning latent personas of film characters," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014, pp. 352–361.

[104] S. Guntuku, W. Lin, M. Scott, and G. Ghinea, "Modelling the influence of personality and culture on affect and enjoyment in multimedia," in *Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, Sept 2015, pp. 236–242.

[105] M. J. Scott, S. C. Guntuku, Y. Huan, W. Lin, and G. Ghinea, "Modelling human factors in perceptual multimedia quality: On the role of personality and culture," in *Proc. 23rd ACM Int. Conf. on Multimedia*, ser. MM '15, 2015, pp. 481–490.

Yoann Baveye was awarded his Master's degree in computer science from the University of Rennes, France, in 2012, followed by his Ph.D. in Computer Science from the Ecole Centrale de Lyon, France, in 2015. Since January 2016, he has been a postdoctoral researcher at the University of Nantes, France. His research interests include signal processing, machine learning, computer vision, pattern recognition and affective computing.

Christel Chamaret is currently working as a Senior Scientist at Technicolor Research & Innovation in Rennes, France. Her research focuses on color image processing and, more particularly, on color harmonization and color grading. She has previously worked on visual attention models and quality metrics as well as on aesthetic models. She has also designed and conducted a number of user studies (pair-wise protocol, eyetracking, etc.) to validate computational models and image processing algorithms. She holds two master's degrees, defended in 2003, from the University of Nantes and the Ecole Centrale de Nantes, France as well as a Ph.D. degree defended in 2016.

Emmanuel Dellandréa was awarded his Master and Engineering degrees in Computer Science from the University of Tours, France, in 2000 followed by his Ph.D. in Computer Science in 2003. He then joined the Ecole Centrale de Lyon, France, in 2004 as an Associate Professor. His research interests include multimedia analysis, image and audio understanding and affective computing, including recognition of affect from image, audio and video signals.

Liming Chen was awarded his B.Sc. degree in joint mathematics-computer science from the University of Nantes, France, in 1984, and his M.S. and Ph.D. degrees in computer science from the University of Paris 6, France, in 1986 and 1989, respectively. He first served as an Associate Professor with the Université de Technologie de Compiègne, France, before joining the Ecole Centrale de Lyon, France, as a Professor in 1998. Since 2007, he has been the Head of the Department of Mathematics and Computer Science at ECL. His research interests include computer vision and multimedia, in particular 2-D/3-D face analysis, image and video categorization, and affective computing. He is a Senior Member of the IEEE.