ORIGINAL RESEARCH
published: 12 June 2017
doi: 10.3389/fict.2017.00011

Films, Affective Computing and Aesthetic Experience: Identifying Emotional and Aesthetic Highlights from Multimodal Signals in a Social Setting

*Theodoros Kostoulas[1,2,3], Guillaume Chanel[1,2], Michal Muszynski[1], Patrizia Lombardo[2,4] and Thierry Pun[1,2]**

[1] Computer Vision and Multimedia Laboratory, University of Geneva, Geneva, Switzerland, [2] Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland, [3] Faculty of Science and Technology, Bournemouth University, Bournemouth, United Kingdom, [4] Department of Modern French, University of Geneva, Geneva, Switzerland

OPEN ACCESS

Edited by:
*Mohamed Chetouani,
Université Pierre et Marie Curie,
France*

Reviewed by:
*Mannes Poel,
University of Twente, Netherlands
Khiet Phuong Truong,
University of Twente, Netherlands
Gualtiero Volpe,
University of Genoa, Italy*

***Correspondence:**
*Thierry Pun
thierry.pun@unige.ch*

Specialty section:
*This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in ICT*

Received: 18 November 2016
Accepted: 28 April 2017
Published: 12 June 2017

Citation:
*Kostoulas T, Chanel G, Muszynski M, Lombardo P and Pun T (2017) Films, Affective Computing and Aesthetic Experience: Identifying Emotional and Aesthetic Highlights from Multimodal Signals in a Social Setting.
Front. ICT 4:11.
doi: 10.3389/fict.2017.00011*

Over the last years, affective computing has been strengthening its ties with the humanities, exploring and building understanding of people's responses to specific artistic multimedia stimuli. "Aesthetic experience" is acknowledged to be the subjective part of some artistic exposure, namely, the inner affective state of a person exposed to some artistic object. In this work, we describe ongoing research activities for studying the aesthetic experience of people when exposed to movie artistic stimuli. To do so, this work is focused on the definition of emotional and aesthetic highlights in movies and studies the people responses to them using physiological and behavioral signals, in a social setting. In order to examine the suitability of multimodal signals for detecting highlights, we initially evaluate a supervised highlight detection system. Further, in order to provide an insight on the reactions of people, in a social setting, during emotional and aesthetic highlights, we study two unsupervised systems. Those systems are able to (a) measure the distance among the captured signals of multiple people using the dynamic time warping algorithm and (b) create a reaction profile for a group of people that would be indicative of whether that group reacts or not at a given time. The results indicate that the proposed systems are suitable for detecting highlights in movies and capturing some form of social interactions across different movie genres. Moreover, similar social interactions during exposure to emotional and some types of aesthetic highlights, such as those corresponding to technical or lightening choices of the director, can be observed. The utilization of electrodermal activity measurements yields in better performances than those achieved when using acceleration measurements, whereas fusion of the modalities does not appear to be beneficial for the majority of the cases.

Keywords: aesthetic experience, synchronization, physiological and behavioral signals, affective computing, social setting, highlights detection

1. INTRODUCTION

Aesthetic experience corresponds to the personal experience that is felt when engaged with art and differs from the everyday experience which deals with the interpretation of natural objects, events, environments and people (Cupchik et al., 2009; Marković, 2012). The exploration of the aesthetic experience and emotions in a social setting can provide the means for better understanding why humans choose to make and engage with art, as well as which features of artistic objects affect our experience.

People exposed to a piece of art can be, in fact, exposed to images, objects, music, colors, concepts, and dialogs. This exposure has an obvious temporal dimension (for example, in movies, music or literature) or an unapparent one (for example, when observing a painting). At the same time, the aesthetic emotions evoked during such an exposure are depicted in the heterogeneous multimodal responses (physiological and behavioral) of the person(s) engaged with a piece of art. Aesthetic experience and Aesthetic emotions are held to be different from everyday experience and emotions (Scherer, 2005; Marković, 2012). In a recent study, an attempt to examine the relation of Aesthetic and everyday emotions is made (Juslin, 2013). Those efforts attempted to define the emotions which might appear when someone is exposed to musical art pieces.

From an affective computing point of view, understanding people responses to art in a social setting can provide insight regarding spontaneous uncontrolled formulations and group behavior in response to some stimuli. This work focuses on understanding people responses to movie artistic stimuli using multimodal signals. To do so, two categories of highlights which are linked with the aesthetic experience while watching a movie are defined: **Emotional** highlights and **Aesthetic** highlights. Their definition follows below:

- Emotional highlights in a given movie are moments that result in high or low arousal and high or low valence at a given time to some audience.
- Aesthetic highlights in a given movie are moments of high aesthetic value in terms of content and form. These moments are constructed by the filmmaker with the purpose of efficiently establishing a connection between the spectator and the movie, thus enabling the spectator to better experience the movie itself.

These definitions rely on (a) a well-established arousal-valence emotion model and (b) the objective identification of moments which are constructed for keeping a person engaged in an aesthetic experience. Though the study of aesthetic emotions, such as those of "being moved," "wonder," and "nostalgia," cannot be realized in this work, since there are no available annotated data for doing so, the exploration of the people responses during an aesthetic experience can be one more step for uncovering the nature of aesthetic emotions.

Emotional highlights can be indicated by annotating a given movie in a two-dimensional space (arousal-valence) from multiple persons and averaging the outcome. Moments of high or low arousal or valence can be determined by comparing to the median over the whole duration of the movie. On the other hand, *aesthetic highlights* follow an objective structure and taxonomy. This is illustrated, along with a description of the different types, in **Figure 1**. This taxonomy was constructed considering the various film theories and utilizing the experts feedback to construct a tier-based annotation process (Bazin, 1967; Cavell, 1979; Deleuze et al., 1986; Deleuze, 1989; David and Thompson, 1994). There exist two general categories of aesthetic highlights (H): highlights of type Form (H1, H2) and highlights of type Content (H3, H4, H5). Form highlights correspond to the way in which a movie is constructed, i.e., the manner in which a subject is presented in the film. Content highlights correspond to the moments in a given film where there exists an explicit development of the components of the film. Such components can be the actor's characters, dialogs developing the social interaction of the characters, development of a specific theme within the movie.

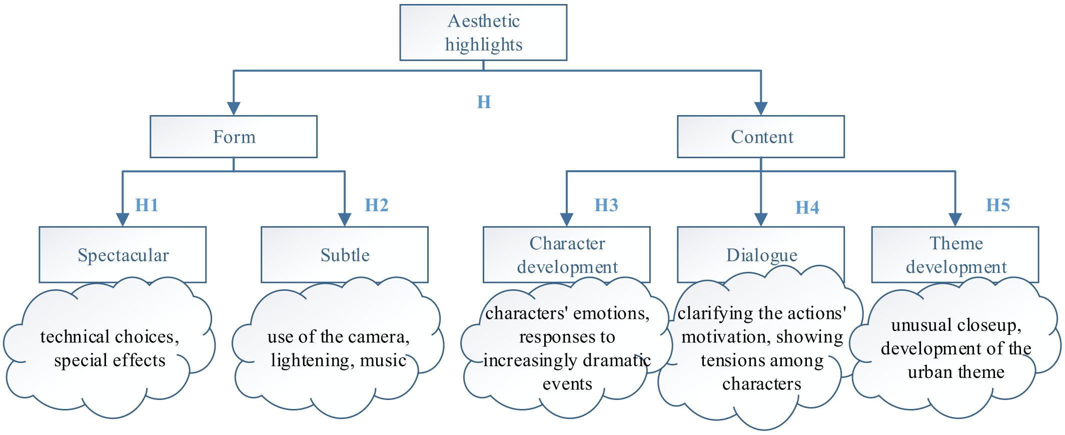
Aesthetic highlights
H
Form
Content
H1
H2
H3
H4
H5
Spectacular
Subtle
Character development
Dialogue
Theme development
technical choices, special effects
use of the camera, lightening, music
characters' emotions, responses to increasingly dramatic events
clarifying the actions' motivation, showing tensions among characters
unusual closeup, development of the urban theme

FIGURE 1 | Aesthetic highlights definition and descriptive examples for each highlight type.

The work described in this manuscript involves uncovering people responses during those highlights and addressing the following research questions:

1. Can emotional and aesthetic highlights in movies be classified from the spectators social physiological and behavioral responses?
2. Which methods can be used to combine the information obtained from the multimodal signals of multiple people, in order to understand people responses to highlights?

1.1. Related Work

In the area of affective computing, a number of implicit measures had been used for modeling people's reactions in some context, such as therapy (Kostoulas et al., 2012; Tárrega et al., 2014), entertainment (Chanel et al., 2011), and learning (Pijeira-Díaz et al., 2016). The common signals selected to be analyzed mostly originate from the autonomous peripheral nervous system (such as heart rate, electrodermal activity) or from the central nervous system (electroencephalograms). Also, behavioral signals have been used in the past for the analysis of emotional reactions through facial expressions, speech, body gestures, and postures (Castellano et al., 2010; Kostoulas et al., 2011, 2012). Moreover, various studies investigated the use of signal processing algorithms for assessing emotions from music or film clips (Lin et al., 2010; Soleymani et al., 2014) using electroencephalogram signals.

With the purpose of characterizing spectators' reactions, some efforts to create an affective profile of people exposed to movie content using a single modality (electrodermal activity) were made in Fleureau et al. (2013). More recent work toward detecting aesthetic highlights in movies included the definition and estimation of a reaction profile for identifying and interpreting aesthetic moments (Kostoulas et al., 2015a), or the utilization of dynamic time warping algorithm for the estimation of the relative physiological and behavioral changes among different spectators exposed to artistic content (Kostoulas et al., 2015b). Other efforts toward identifying synchronization among multiple spectators had been focused in representing physiological signals on manifolds (Muszynski et al., 2015) or on applying periodicity score to measure synchronization among groups of spectators' signals that cannot be identified by other measures (Muszynski et al., 2016). Further, recent attempts which study the correlation of the emotional responses with the physiological signals in a group setting had indicated that some emotional experiences are shared in the social context (Golland et al., 2015), whereas others were focused on analyzing arousal values and galvanic skin response while movie watching (Li et al., 2015) or on the identification of the movie genre in a controlled environment (Ghaemmaghami et al., 2015).

The work conducted so far, specifically by the authors of the current work, show significant ability to recognize aesthetic highlights in an ecological situation and suggest that the presence of aesthetic moments elicit multimodal reaction in some people compared to others. Yet, the relation of the aesthetic moments defined by experts and the emotional highlights, as those can be defined in an arousal-valence space, has not been explored. Further, the manifestation of the different reactions to emotional and aesthetic highlights to different types of movie genres is not studied to this date. This would allow confirming whether the selected methods are appropriate for studying aesthetic experience. Further, it would allow uncovering the differences among the different movie genres and understanding people responses to some types of movie stimuli.

The article is structured as follows. In Section 2, the material and methods designed and implemented are described. In Section 3, the experimental setup and results are included. The results and future research direction are discussed in Section 4.

2. MATERIALS AND METHODS

We make the assumption that the responses of people in a social setting can be used to identify emotional and aesthetic highlights in movies. The signals selected to be used were electrodermal activity and acceleration measurements. The choice of these measurements was motivated by two factors: first, the need of studying physiological and behavioral responses and the suitability of those modalities for emotional assessment based on the current state of the art. Second, the resources available, i.e., for one part of the dataset used in this study we had to use a custom-made solution for performing such a large-scale experiment, which was not possible to support all possible modalities.

In order to answer to the **first research question**, we propose a **supervised highlight detection system** and evaluate a binary classification problem (highlight versus non-highlight), toward uncovering the discriminative power of the used multimodal signals. Specifically, we examine the performance of a supervised emotional/aesthetic highlights detection system, trained and evaluated on a given movie (movie-dependent highlight detection).

In order to answer the **second research question** and gain insight on the people responses to emotional and aesthetic highlights, we propose the utilization of two **unsupervised highlight detection systems**: The first one measures the distance among the multimodal signals of the spectators at a given moment using the dynamic time warping algorithm. The second one is capturing the reactions of the spectators at a given moment, using clustering of multimodal signals over time.

In all the experiments conducted, binary problems are considered. Those problems correspond to the task of detecting whether there is a highlight or not from multimodal signals of multiple people. Among the different classes (in our cases emotional or aesthetic highlights), there is an overlap (e.g., in a given moment, we can have more than one highlight). In this work, we focus on studying the responses of people independently of those overlaps.

2.1. Supervised Highlight Detection System

The supervised highlight detection framework illustrated in **Figure 2** was designed and implemented. The knowledge repository consists of (a) annotated movies in terms of emotional and aesthetic highlights and (b) synchronized multimodal measures of spectators watching these movies. During the training phase,

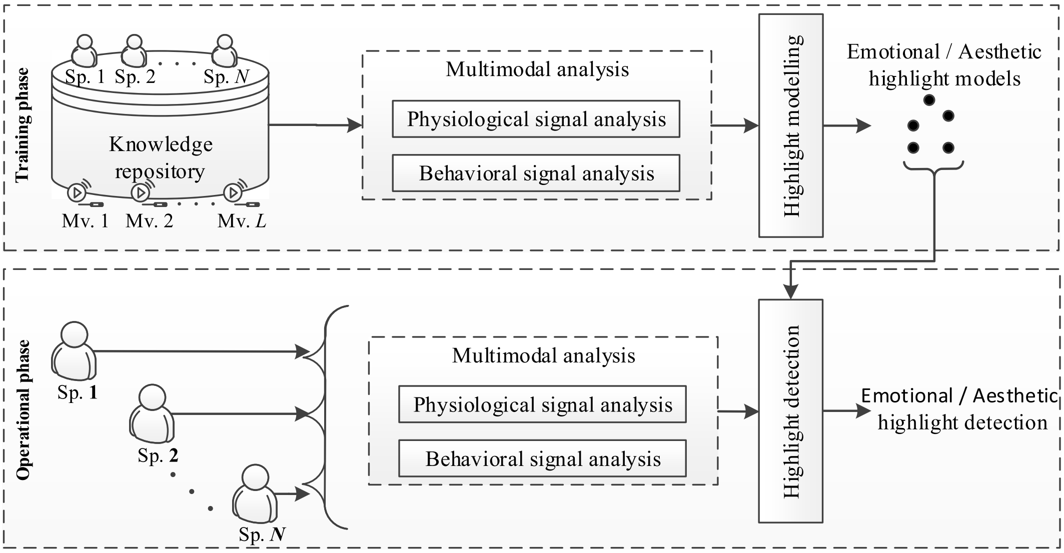
Training phase
Sp. 1
Sp. 2
Sp. N
Knowledge repository
Mv. 1
Mv. 2
Mv. L
Multimodal analysis
Physiological signal analysis
Behavioral signal analysis
Highlight modelling
Emotional / Aesthetic highlight models
Operational phase
Sp. 1
Sp. 2
Sp. N
Multimodal analysis
Physiological signal analysis
Behavioral signal analysis
Highlight detection
Emotional / Aesthetic highlight detection

FIGURE 2 | Supervised highlight detection system.

the data from the knowledge repository are initially subject to multimodal analysis: Let Sp_i be one spectator watching a movie with $i = 1, 2, \ldots, N$. A sliding window d of constant length k is applied to the input signals. A constant time shift s between two subsequent frames is determined. The behavioral and physiological signals are initially subject to lowpass filter, to account for the noise and distortions, as well as capturing the low frequency changes that occur in acceleration and electrodermal activity signals. The resulting signal is then subject to feature extraction and emotional/aesthetic highlight modeling. During the operational phase the physiological and behavioral signals are subject to the same preprocessing and feature extraction processes. A decision regarding whether a signal segment belongs to a highlight or not is made by utilizing the corresponding highlight models created during the training phase.

2.2. Unsupervised Highlight Detection Systems

Two unsupervised highlight detection systems were implemented following our work described in Kostoulas et al. (2015a,b) (refer to **Figure 3**).

The first unsupervised highlight detection system computes the pairwise dynamic time warping distances among the multimodal signals of all possible pairs of spectators (Kostoulas et al., 2015b). This process results in a vector which is indicative of the distances among the signals of all possible pairs over the duration of the movie. This vector is fed to the highlight detection component, which post-processes the input vector either by creating the corresponding highlight models or by applying a measure (such as mean and median) at a given time for estimating the degree of existence of a highlight. In the present article, the median distance over all possible pairs at a given time is used as a measure, toward accounting for the distribution of the scores among different pairs of spectators.

The second unsupervised highlight detection system processes the feature vectors extracted from the multimodal signals and clusters them for splitting them in two clusters over the duration of a given movie (Kostoulas et al., 2015a). There are significant changes in the acceleration signal when a movement occurs, and the same applies for the galvanic skin response signal when a person is reacting to some event. We make the assumption that those periods can be identified by clustering our data in two clusters. The two clusters would correspond to periods of reactions and relaxations. We expect that the moments that people react are shorter than the moments that people relax-do not react. Therefore, the cluster which contains the majority samples includes samples from relaxation periods, i.e., periods that no observable activity can be detected on the acquired multimodal signals. On the other hand, the cluster with fewer samples assigned to it includes samples from reaction periods. The vector resulting from the concatenation of the assigned clusters over time can, therefore, be considered as reaction profile of the given set of spectators over the duration of the movie. This profile is then processed by the highlight detection component for computing a measure of the groups reaction (e.g., the percentage of spectators belonging to the reaction cluster).

The advantage of the first unsupervised system is that it can identify moments where the distance among the signals of all possible pairs of spectators is increasing or decreasing. This can be considered as a measure of dissimilarity among the multimodal signals of multiple spectators. The advantage of the second system is that it can efficiently identify moments were multiple reactions from the groups of spectators are observed. This can be considered as a measure of reactions of groups of people or relaxations.

2.3. Datasets

Multimodal signals (behavioral and physiological) from multiple spectators watching movies are utilized. In this study, we used two

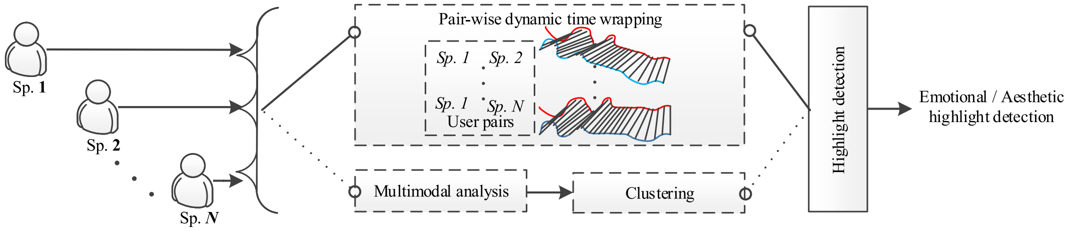

FIGURE 3 | Unsupervised highlight detection systems.

datasets, the first one annotated in terms of aesthetic highlights and the second one annotated in terms of both emotional and aesthetic highlights. The reason for using both datasets was to study different movie types-genres and to evaluate the suitability of our methods for them. The two datasets are described, briefly, below.

The first dataset corresponds to recordings of 12 people watching a movie in a theater (Grütli cinema, Geneva) (Kostoulas et al., 2015a). In this dataset (hereafter "Taxi" dataset), the selection of the movie (Taxi Driver, 1976) was done with respect to its content of aesthetic highlights. The electrodermal activity (sensor recording from the fingers of the participants) and acceleration (sensor placed on the arm of the participant) signals are used in this study. The sensor used was realized as part of a master thesis (Abegg, 2013). The duration of the movie is 113 min. The total number of spectators was 40.

The second dataset (hereafter "Liris" dataset) is part of the LIRIS database (Li et al., 2015). Physiological and behavioral signals (electrodermal activity and acceleration used in this study) were collected from 13 participants in a darkened air-conditioned amphitheater, for 30 movies. The sensor recording those modalities was placed on the fingers of the participants. The sensor used was the Bodymedia armband (Li et al., 2015). The total duration of the movies is 7 h, 22 min, and 5 s. The following genres were defined in this dataset: Action, Adventure, Animation, Comedy, Documentary, Drama, Horror, Romance, and Thriller.

The *emotional highlights* are determined by the annotation of the movies by 10 users in the arousal-valence space. Further information can be found in Li et al. (2015). Annotation of *aesthetic highlights* was realized by an expert assisted by one more person. The aesthetic highlights are moments in the movie which are constructed in a way to engage aesthetic experience and are, in those terms, subject to objective selections. The annotation represented the judgment of the movie based on a neutral aesthetic taste. Since the movies included in the "Liris" dataset were not annotated with respect to their aesthetic highlights, annotation in terms of form and content (as illustrated in **Figure 1**) is performed. Similarly to previous work (Kostoulas et al., 2015a), the annotation has been realized using open-source annotation software (Kipp, 2010). The result of this annotation process is shown in **Table 1**. There, the average number of continuous pieces characterized as highlights within a movie and their average duration are illustrated.

Regarding ethics, these experiments belong to the domain of computer science and multimodal interaction. Their goal is to facilitate the creation and access to multimedia data. As far as the data collected in Geneva, Switzerland, are concerned, this study was done in compliance with the Swiss law; no ethical approval was required for research conducted in this domain. Moreover, the data collection process and handling were carried out in accordance to the law on public information, access to documents and protection of personal data (LIPAD, 2016). All participants filled in the appropriate consent forms which are stored in the appropriate manner in our premises. All participants were informed that they could stop the experiment at any time. Their data are anonymized and stored on secured servers. As far as the data included in the second dataset are concerned, we dealt with properly anonymized data, where the participants had to sign a consent form and were informed regarding the protection of their anonymity (Li et al., 2015).

TABLE 1 | Average aesthetic highlights statistics for the Liris dataset.

Highlights / Statistics	Number	Average duration (s)
H1	5.37	24.43
H2	4.73	18.10
H3	4.47	28.77
H4	2.63	29.43
H5	5.37	24.93

3. RESULTS

3.1. Experimental Setup

The knowledge repository utilized in this study is divided in two parts as described in Section 2.3. The "Taxi" dataset includes acceleration (3-axes) and electrodermal activity signals acquired from 12 participants, sampled at 10 Hz. The "Liris" dataset, also, includes acceleration (magnitude) and electrodermal activity signals. The signals are segmented in non-overlapping windows of 5 s length which results to sequences of non-overlapping frames, to account for an effective experimental setup and ensure no training-testing overlap in the movie-dependent task. The number of samples per class and per experiment conducted is indicated in **Table 2**. The indication "NaN" refers to the case where no samples of this highlight type were annotated.

The information from multiple spectators was included in the implemented systems by concatenating the feature vectors calculated for each one of them to one feature vector. In all experiments, three sub-cases were considered for examining the effect of each modality on identifying highlights: (a) utilizing the electrodermal activity modality, (b) utilizing the acceleration

TABLE 2 | Number of samples per classification/detection problem.

Highlights	Genre	Action	Advent.	Animat.	Comed.	Docum.	Drama	Horror	Roman.	Thriller	LIRIS—total	Taxi
Total samples		**617**	**540**	**517**	**1,058**	**60**	**819**	**911**	**408**	**301**	**5,231**	**1,351**
Emotional	Arousal	309	272	258	529	30	409	454	204	152	2,617	–
	Valence	307	264	262	525	31	411	454	201	149	2,604	–
Aesthetic	H1	57	44	70	86	4	64	100	33	34	492	86
	H2	24	49	52	73	NaN	66	16	42	32	354	169
	H3	87	62	71	142	2	117	135	45	26	687	129
	H4	76	68	29	155	NaN	94	82	23	24	551	122
	H5	58	60	60	133	2	99	122	48	38	620	70
	H	201	185	193	417	8	274	314	157	103	1,852	401

TABLE 3 | Signal parameters extracted.

Signal	Functionals
s *Ds* *D2s*	mean median std min max minRatio maxRatio

modality, or (c) fusion of the two modalities at the feature level (i.e., concatenating the corresponding feature vectors).

3.1.1. Supervised Highlight Detection

In order to evaluate the supervised highlight detection system, the multimodal data of each movie were split in training and testing sets (70 and 30%, respectively), randomly selected 10 times. The training and testing sets are non-overlapping, but contain samples from the same spectators and are, in those terms, person-group dependent. For each of the experiments conducted, only movies which contained enough samples I ($I > 10$) were considered. This was done to ensure the appearance of a minimum number of instances, with respect to the duration and type of highlights, for training the corresponding models.

The signals were subject to lowpass Butterworth filter of order 3 and cutoff frequency 0.3 Hz. The functionals shown in **Table 3** were applied (Wagner, 2014). For the electrodermal activity signal, the functionals are applied to the original signal *s*, to its first derivative *Ds* and to its second derivative *D2s*. For the acceleration signals, the same process is applied to each of the signals corresponding to the *x*, *y*, and *z* axes or to the magnitude signal (in the "Liris" dataset).

Each binary classifier is a support vector machine (SVM). We relied on the LibSVM (Chang and Lin, 2011), implementation of SVM with radial basis kernel function (RBF) (Fan et al., 2005). When building the binary classifiers, the class imbalance was handled by utilizing the priors of the class samples: setting the parameter C of one of the two classes to wC, where w is the ratio of number of samples of the majority class to number of samples of the minority class. The optimal γ parameter of the radial basis kernel function considered here and the C parameter, were determined by performing a grid search $\gamma = \{2^3, 2^1, \ldots, 2^{-15}\}$, $C = \{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ with 10-fold cross validation on the training set. In order to take into account that the number of samples per class is not similar, the primary performance measure was the balanced accuracy over classes.

3.1.2. Unsupervised Highlight Detection

In order to evaluate the unsupervised highlight detection systems, the experimental setup followed in Kostoulas et al. (2015b) and Kostoulas et al. (2015a) was utilized. In summary, for the one-dimensional electrodermal activity signal the DTW algorithm was applied (Müller, 2007), whereas for the acceleration signals the same algorithm is applied to the multidimensional signal composed of the *x*, *y*, and *z* axes or the magnitude signal. For the clustering method, the EM algorithm (Dempster et al., 1977) for expectation maximization is utilized for the unsupervised clustering of the spectators data. For the parameters of the EM algorithm, the values for the maximum number of iterations and allowable standard deviation were set to 100 and 10^{-6}, respectively.

3.2. Experimental Results

In this section, we describe the results of the evaluation of the methods described in Section 2. For the supervised highlight detection system, balanced accuracy is used as performance measure, to account for the unbalanced number of sample per class. For the unsupervised methods, area-under-curve (AUC) was the preferred performance measure. This was motivated by the fact that it can provide feedback regarding the suitability of the performance measure, as well as the performance of the detection system. For example, when using the distance among the signals of the spectators as a measure, if the AUC is significantly higher than 0.5 for one type of highlight, this means that the distance among the multimodal signals for this highlight type increases and we can use this distance to detect highlights of this type.

3.2.1. Supervised Highlight Detection

In **Table 4**, the results of the evaluation of the supervised highlight detection system is illustrated (two sided Welch's t-test, $a = 0.05$ was applied to each result described below and statement made.[1]) Results for the emotional highlights are not included for the "Taxi" dataset, since there are no available annotated data for arousal and valence. As shown in **Table 4** the detector of emotional/aesthetic highlights shows, overall, a significant ability to recognize

[1] Data following normal distribution as tested using one-sample Kolmogorov-Smirnov test with significance level set at 0.05

highlights in movies. The electrodermal activity modality appears to be the most appropriate modality for detecting highlights in movies, whereas the feature-level fusion of modalities does not seem to be beneficial in for the detection of highlights, at least for the "Liris" dataset.

However, this is not the case for the "Taxi" dataset, where the fusion of modalities improves the system's performance for highlights of type 4 ($p < 0.01$), as well as overall (H) ($p < 0.01$).

TABLE 4 | Balanced accuracy (%) of emotional and aesthetic highlights detection, for the two datasets (Liris, Taxi) using electrodermal activity (GSR), acceleration (ACC), and fusion of modalities (FUSE).

Highlights / **Dataset**		**Liris**			**Taxi**		
Modality ⟶		**GSR**	**ACC**	**FUSE**	**GSR**	**ACC**	**FUSE**
Emotional	Arousal	79.94	60.22	76.40	–	–	–
	Valence	80.70	59.54	78.23			
Aesthetic	H1	71.34	55.84	68.96	73.20	65.49	70.29
	H2	67.25	54.93	63.34	70.58	70.37	71.64
	H3	68.08	55.57	66.37	61.39	64.34	69.94
	H4	74.66	58.14	71.95	65.75	66.42	72.60
	H5	63.23	57.35	62.20	67.89	70.45	70.19
	H	62.90	54.68	61.91	65.67	67.14	71.48

Results correspond to the evaluation of binary classification problems (highlights vs. non-highlight).

Further in this dataset the acceleration modality appears to be significantly less discriminative only in the case of H1 ($p < 0.01$). This would suggest that the placement of the acceleration sensor plays an important role (e.g., on the finger, the arm, the back of a person, etc.). Further, the utilization of 3-axes is beneficial, compared to using the magnitude of the acceleration signal.

3.2.2. Unsupervised Highlight Detection

In **Tables 5–10**, the results (AUC) for movie independent unsupervised highlight detection are included. Results are not included for the "Taxi" dataset since thorough investigation of this dataset is made in Kostoulas et al. (2015a,b). The performance of the proposed architectures was evaluated for the different type of movie genres, as those are defined within the Liris dataset. Dark gray and light gray cells highlight the significantly better and worse (respectively) performances, when compared with random (i.e., AUC = 50%), with significance level $a = 0.05$ (Bradley, 1997).

Tables 5 and **6** show a significant ability of both the DTW and clustering methods to predict arousal and valence highlights based on the electrodermal activity modality. Yet, this is not the case for several types of aesthetic highlights. Overall, we can observe that in Action films there is some decreased overall distance among the signals of all possible pairs of spectators (AUC below 50%), and some strong reactions in the case of H1 (clustering method, AUC significantly higher than random). Similar behavior of the proposed systems is observed for animation, adventure and

TABLE 5 | Area under curve (%) for highlights detection using electrodermal activity and the DTW method.

Highlights / **Genre**		**Action**	**Advent.**	**Animat.**	**Comed.**	**Docum.**	**Drama**	**Horror**	**Roman.**	**Thriller**
Emotional	Arousal	49.43	61.57	50.94	50.74	63.22	56.28	67.69	27.72	49.63
	Valence	66.04	46.94	48.30	49.79	32.59	43.48	27.22	70.13	35.09
Aesthetic	H1	39.41	63.56	48.45	53.04	40.18	52.46	57.60	36.36	53.91
	H2	27.61	41.82	61.22	44.86	NaN	56.74	26.84	36.17	53.91
	H3	56.76	52.18	52.96	45.47	61.21	56.02	62.29	28.76	51.69
	H4	42.60	39.42	49.92	45.44	NaN	55.48	54.38	67.00	63.27
	H5	36.52	50.92	46.76	38.85	5.17	48.40	40.65	56.42	43.46
	H	41.96	46.18	47.85	44.83	35.34	51.93	51.92	42.10	53.22

TABLE 6 | Area under curve (%) for highlights detection using electrodermal activity and the clustering method.

Highlights / **Genre**		**Action**	**Advent.**	**Animat.**	**Comed.**	**Docum.**	**Drama**	**Horror**	**Roman.**	**Thriller**
Emotional	Arousal	52.01	59.93	39.48	50.60	46.44	48.55	71.94	44.30	39.65
	Valence	59.56	43.51	48.84	48.56	33.59	46.65	22.76	59.23	44.13
Aesthetic	H1	61.66	65.17	50.73	70.07	37.05	53.88	67.48	27.34	63.55
	H2	56.47	45.33	67.79	53.19	NaN	47.27	46.76	46.83	47.17
	H3	45.00	53.97	44.71	44.07	63.79	42.22	67.96	33.15	54.29
	H4	38.12	49.61	28.21	44.76	NaN	45.82	36.09	60.84	57.38
	H5	51.16	46.09	44.72	42.94	70.69	41.90	52.24	50.01	53.34
	H	49.67	47.74	49.22	51.07	52.64	44.03	53.19	37.33	52.38

TABLE 7 | Area under curve (%) for highlights detection using acceleration and the DTW method.

Highlights	Genre	Action	Advent.	Animat.	Comed.	Docum.	Drama	Horror	Roman.	Thriller
Emotional	Arousal	50.07	37.97	37.10	55.10	55.00	47.18	52.81	43.09	45.65
	Valence	57.37	45.17	52.05	51.28	34.71	50.20	41.75	56.34	50.19
Aesthetic	H1	63.09	48.58	50.94	59.59	73.21	52.95	53.96	31.07	45.30
	H2	72.15	52.44	57.86	51.77	NaN	45.60	57.88	57.31	47.04
	H3	50.89	36.65	37.62	49.77	33.62	44.30	51.18	36.82	43.52
	H4	51.20	43.47	40.24	47.49	NaN	55.50	41.78	64.90	48.33
	H5	47.95	45.89	36.30	50.03	43.97	39.71	42.78	45.11	38.64
	H	57.98	41.67	45.00	48.83	56.25	47.01	47.51	41.39	43.63

TABLE 8 | Area under curve (%) for highlights detection using acceleration and the clustering method.

Highlights	Genre	Action	Advent.	Animat.	Comed.	Docum.	Drama	Horror	Roman.	Thriller
Emotional	Arousal	45.89	43.74	35.18	51.53	59.06	43.81	57.80	42.41	39.66
	Valence	52.69	48.63	54.22	50.91	40.49	52.89	39.27	54.87	52.24
Aesthetic	H1	56.64	56.34	49.14	60.37	60.04	47.74	56.11	36.33	44.27
	H2	66.00	52.22	63.20	49.15	NaN	46.14	54.32	54.19	45.38
	H3	50.50	46.83	37.67	48.93	23.71	43.81	56.79	36.52	45.23
	H4	45.70	43.12	46.49	48.65	NaN	51.60	43.70	63.55	56.06
	H5	53.64	45.71	38.31	51.38	23.71	37.47	47.14	46.86	35.99
	H	53.96	42.23	47.27	50.49	40.75	44.74	49.37	42.33	45.45

TABLE 9 | Area under curve (%) for highlights detection using fusion of modalities and the DTW method.

Highlights	Genre	Action	Advent.	Animat.	Comed.	Docum.	Drama	Horror	Roman.	Thriller
Emotional	Arousal	40.15	48.66	39.01	53.38	45.33	42.77	57.84	38.95	42.76
	Valence	53.24	49.89	53.25	53.92	27.92	58.55	35.73	57.11	49.66
Aesthetic	H1	49.22	60.02	47.40	43.05	28.57	58.16	55.75	32.17	47.26
	H2	54.10	47.24	59.81	48.82	NaN	50.73	44.98	60.38	47.22
	H3	54.40	46.28	47.66	42.95	55.17	44.36	60.69	47.48	48.52
	H4	47.95	36.92	59.33	48.76	NaN	57.37	47.44	62.00	67.54
	H5	45.55	45.94	38.27	45.15	2.59	46.31	44.64	48.11	38.32
	H	50.80	43.30	47.53	46.01	26.68	47.56	51.54	44.53	48.67

comedy movie genres. In horror and romance genres, slightly opposite behavior is observed, which was expected, considering the form and content of these types of movies. In the meantime, Thriller movies are characterized by low AUC scores for valence detection using DTW, which suggests that people respond in a synchronized manner during those moments. Moreover, the low AUC for arousal detection using the clustering method suggests that people can be in relaxation period during those moments.

In **Tables 7** and **8**, we illustrate the performance of the systems when using the acceleration modality. Overall, the significant higher than 0.5 AUC values for highlights of type form (H1 and H2) for some movie genres (Action, Animation, Comedy) indicate that there is increased distance among the signals of the recorded spectators, as well as reactions (clustering method). This observation could result from the fact that only part of the spectators do react. However, this is not the case for Romance films. Romance films are characterized by more relaxation periods which is supported by the low scores of the AUC for the DTW method (we expect that the distance is not increased when being relaxed), as well as with the low scores of AUC for the clustering method (i.e., reaction profile suggesting the absence of any reactions during this type of highlights). Yet, dialog scenes (H4) are

TABLE 10 | Area under curve (%) for highlights detection using fusion of modalities and the clustering method.

Highlights	Genre	Action	Advent.	Animat.	Comed.	Docum.	Drama	Horror	Roman.	Thriller
Emotional	Arousal	51.46	51.54	36.48	46.76	41.11	50.00	65.55	45.05	42.61
	Valence	57.82	43.85	52.24	50.49	44.66	46.81	29.70	54.96	45.86
Aesthetic	H1	61.43	67.55	50.34	65.29	45.98	55.60	62.04	30.90	54.47
	H2	57.12	48.32	64.98	55.49	NaN	49.25	54.04	54.19	46.56
	H3	48.37	47.77	40.64	50.04	16.38	44.90	62.56	37.93	52.94
	H4	38.83	42.44	32.88	43.71	NaN	50.69	38.00	69.71	63.52
	H5	53.38	51.49	38.78	43.74	31.90	43.38	49.58	49.80	43.86
	H	51.22	44.14	46.23	49.66	33.41	46.20	51.58	43.38	50.89

causing reactions to the spectators for Romance movies, which is something that one would expect.

Tables 9 and **10** illustrate the evaluation of the proposed systems when fusion of the modalities is applied. Overall, the fusion of modalities does not appear to be beneficial for the majority of the considered binary problems. The reason for this failure is due to the fact that the acceleration modality does not seem to convey important information for highlights of any type for the Liris dataset (refer also to **Table 4**). Moreover, the short duration of movies might result into less robust clusters (as far as the clustering method is concerned) for the increased feature vector (since we apply feature-level fusion).

4. DISCUSSION

Overall, the use of the electrodermal activity modality appears to be better for detecting highlights in movies. This is certainly the case for the Liris dataset. However, the placement of the acceleration sensor, which is not separated from the electrodermal activity one (on the fingers of the participants), can be a valid explanation for the unstable performance of the proposed methodologies, compared to their performance on the Taxi dataset. A first hint about the reason for the inefficiency of the proposed architectures when using the acceleration modality can be seen from the performance of the supervised emotional and aesthetic highlight detection systems. However, further research would be needed in order to identify the optimal design of the experimental setup for capturing highlights in movies and assessing the emotions of the spectators using the acceleration modality. Though, the utilization of the electrodermal activity modality leads in general to better performance, one should keep in mind that, currently, electrodermal activity sensors are hardly found in the majority of the hand-held devices and everyday objects. However, acceleration sensors are generally installed in every smartphones, making it a promising modality for massive utilization in social experiments and real-life applications. The choice of such unobtrusive sensors is further motivated by the intended application of this work: assist filmmakers in selecting the appropriate methods when creating the movie or understanding how people respond during an aesthetic experience. Transition from controlled experiments to large scale ones, with multiple movies, in real-life settings, would require ready-to-access sensor data.

Emotional highlights are generally characterized by strong reactions, as indicated by the reaction profile of multiple users, as well as the increased distance among their signals. Previous work on continuous arousal self assessment validation (Li et al., 2015) has shown that there is a temporal correlation between the electrodermal activity signal and continuous arousal annotations. The achieved results in our work are in line with the previous findings and confirm the usability of the electrodermal activity modality for detecting arousal.

Aesthetic highlights share some patterns with the emotional ones in the case of highlights of type form (i.e., H1 and H2), whereas this is not the case for highlights of type content (i.e., H3, H4, and H5). The observed results are in line with previous research in aesthetic highlights detection (Kostoulas et al., 2015a,b; Muszynski et al., 2016). Specifically, it is indicated that electrodermal activity and acceleration signals can be used for detecting some types of highlights, especially the ones of H1 and H2, where the use of special effects, lightening techniques, and music are expected to significantly affect the reactions of the spectators. However, since the present work includes the evaluation of a different database, no generalization can be made, due to the different type of movie, small sample size, environment where the experiment was conducted, as well as sensors used. Yet, some observations (such as the opposite behavior of the proposed architectures in romance and horror movie genres), as well as the aforementioned previous research, indicate that the proposed unsupervised architectures are able to discover shared patterns and synchronization measures among groups of people, in a social setting.

5. CONCLUSION

In this work, the definition of emotional and aesthetic highlights was introduced, in order to study the aesthetic experience while watching a movie. According to this definition, emotional highlights are subjective and correspond to moments of high or low felt arousal or valence while watching a movie. Aesthetic highlights are moments of high aesthetic value in terms of content and form and are constructed by the filmmaker with the purpose establishing and maintaining a connection between the spectator and the movie.

In response to the need of studying people responses to those highlights in a social setting, this article studies supervised and unsupervised highlight detection systems. In general, the proposed architectures depict significant capability to detect emotional and aesthetic highlights, both in movie dependent and movie independent modes. In response to the first research question set within this work, the present findings suggest that it is possible to detect emotional and aesthetic highlights in movies of different genres using multimodal signals in a social setting. In response to the second research question, the proposed architectures depict significant capability to efficiently combine information and provide insight on people responses to those highlights, along different movie genres.

As far as the modalities utilized are concerned, the utilization of electrodermal activity measurements results in better performances than those achieved with acceleration signals. In the meantime, fusion of the modalities does not appear to be beneficial for the majority of the cases, possibly due to the placement of the sensors. One main limitation of this work corresponds to the number of available annotated data, i.e., the availability of more labeled data would possibly allow us to observe social interactions that are now not visible. However, when considering multiple modalities and multiple people, the question of how feasible it is to conduct large scale experiments arises, mainly due to the available resources at a given time.

Future work includes collecting multimodal data in real-life settings toward building more robust models. Further, possible future access to cost-effective sensors might enable the usage of modalities that are currently expensive to use, such as electroencephalograph signals. In order to better interpret aesthetic experience and its components, annotating existing and prospective databases in terms of their content in aesthetic emotions would be necessary. Moreover, deep exploration of the aesthetic experience would require accounting for different forms and content of art, e.g., music, literature, and paintings.

AUTHOR CONTRIBUTIONS

TK analyzed the content included in this manuscript, conducted the experiments, developed part of the materials, the methods, and wrote the manuscript. GC assisted in the development of the overall content included in this article. MM assisted in the development of the content included in this article, including development of part of the material. PL assisted the development of the material, supervised their development, with emphasis on the highlights definition. TP participated to the definition of the research questions, coordinated the work conducted, and assisted in the formulation of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the “Grütli cinema, Geneva, Switzerland.” The development of the ideas as well as most of the research work was accomplished while the first author was with the University of Geneva.

FUNDING

This work is partially supported by grants from the Swiss Center for Affective Sciences and the Swiss National Science Foundation.

REFERENCES

Abegg, C. (2013). *Analyse du confort de conduite dans les transports publics*. Thesis. University of Geneva, Geneva.

Bazin, A. (1967). in *What is Cinema?* trans. H. Gray (Berkeley, CA: University of California Press), 14.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 1145–1159. doi:10.1016/S0031-3203(96)00142-2

Castellano, G., Caridakis, G., Camurri, A., Karpouzis, K., Volpe, G., and Kollias, S. (2010). “Body gesture and facial expression analysis for automatic affect recognition,” in *Blueprint for Affective Computing: A Sourcebook*, eds K. R. Scherer, T. Baenziger, and E. B. Roesch (Oxford, UK: Oxford University Press), 245–255.

Cavell, S. (1979). *The World Viewed*, enlarged Edn. Cambridge: Harvard University.

Chanel, G., Rebetez, C., Bétrancourt, M., and Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 41, 1052–1063. doi:10.1109/TSMCA.2011.2116000

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 27. doi:10.1145/1961189.1961199

Cupchik, G. C., Vartanian, O., Crawley, A., and Mikulis, D. J. (2009). Viewing artworks: contributions of cognitive control and perceptual facilitation to aesthetic experience. *Brain Cogn.* 70, 84–91. doi:10.1016/j.bandc.2009.01.003

David, B., and Thompson, K. (1994). *Film History: An Introduction*. New York: MacGraw-Hill.

Deleuze, G. (1989). in *Cinema 2: The Time-Image*, trans. H. Tomlinson and R. Galeta (London: Athlone).

Deleuze, G., Tomlinson, H., and Habberjam, B. (1986). *The Movement-Image*. Minneapolis: University of Minnesota.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* 39, 1–38.

Fan, R.-E., Chen, P.-H., and Lin, C.-J. (2005). Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.* 6, 1889–1918.

Fleureau, J., Guillotel, P., and Orlac, I. (2013). “Affective benchmarking of movies based on the physiological responses of a real audience,” in *2013 Humaine Association Conference on (IEEE) Affective Computing and Intelligent Interaction (ACII)* (Geneva: IEEE), 73–78.

Ghaemmaghami, P., Abadi, M. K., Kia, S. M., Avesani, P., and Sebe, N. (2015). “Movie genre classification by exploiting MEG brain signals,” in *Image Analysis and Processing ICIAP 2015* (Genova: Springer), 683–693.

Golland, Y., Arzouan, Y., and Levit-Binnun, N. (2015). The mere co-presence: synchronization of autonomic signals and emotional responses across co-present individuals not engaged in direct interaction. *PLoS ONE* 10:e0125804. doi:10.1371/journal.pone.0125804

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions. *Phys. Life Rev.* 10, 235–266. doi:10.1016/j.plrev.2013.05.008

Kipp, M. (2010). “Anvil: the video annotation research tool,” in *Handbook of Corpus Phonology*, eds J. Durand, U. Gut, and G. Kristoffersen (Oxford: Oxford University Press).

Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, P. (2015a). “Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals," in *2015 Seventh International Workshop on, IEEE Quality of Multimedia Experience (QoMEX)* (Messinia: IEEE).

Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2015b). "Dynamic time warping of multimodal signals for detecting highlights in movies," in *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence, ICMI 2015* (Seattle: ACM), 35–40.

Kostoulas, T., Ganchev, T., and Fakotakis, N. (2011). "Affect recognition in real life scenarios," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, eds A. Esposito, A. M. Esposito, R. Martone, V. C. Müller, and G. Scarpetta (Berlin, Heidelberg: Springer), 429–435.

Kostoulas, T., Mporas, I., Kocsis, O., Ganchev, T., Katsaounos, N., Santamaria, J. J., et al. (2012). Affective speech interface in serious games for supporting therapy of mental disorders. *Expert Syst. Appl.* 39, 11072–11079. doi:10.1016/j.eswa.2012.03.067

Li, T., Baveye, Y., Chamaret, C., Dellandréa, E., and Chen, L. (2015). "Continuous arousal self-assessments validation using real-time physiological responses," in *International Workshop on Affect and Sentiment in Multimedia (ASM)*, Brisbane.

Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., et al. (2010). Eeg-based emotion recognition in music listening. *IEEE Trans. Biomed. Eng.* 57, 1798–1806. doi:10.1109/TBME.2010.2048568

LIPAD. (2016). *Law on Public Information, Access to Documents and Protection of Personal Data*. Available at: https://www.geneve.ch/legislation/rsg/f/s/rsg_a2_08.html

Marković, S. (2012). Components of aesthetic experience: aesthetic fascination, aesthetic appraisal, and aesthetic emotion. *Iperception* 3, 1–17. doi:10.1068/i0450aap

Müller, M. (2007). "Dynamic time warping," in *Information Retrieval for Music and Motion* (Berlin, Heidelberg: Springer), 69–84.

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

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

Pijeira-Díaz, H. J., Drachsler, H., Järvelä, S., and Kirschner, P. A. (2016). "Investigating collaborative learning success with physiological coupling indices based on electrodermal activity," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, UK: ACM), 64–73.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Soc. Sci. Inf.* 44, 695–729. doi:10.1177/0539018405058216

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

Tárrega, S., Fagundo, A. B., Jimnez-Murcia, S., Granero, R., Giner-Bartolom, C., Forcano, L., et al. (2014). Explicit and implicit emotional expression in bulimia nervosa in the acute state and after recovery. *PLoS ONE* 9:e101639. doi:10.1371/journal.pone.0101639

Wagner, J., Kim, J., and André, E. (2005). "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on* (Amsterdam: IEEE), 940–943.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.