García-Faura, Álvaro; Hernández-García, Alejandro; Fernández-Martínez, Fernando; Díaz-de-María, Fernando; San-Segundo, Rubén. (2019). Emotion and attention: Audiovisual models for group-level skin response recognition in short movies. *Web Intelligence*, 17(1), pp.: 29-40.

DOI: https://doi.org/10.3233/WEB-190398

Web Intelligence 17 (2019) 29–40
DOI 10.3233/WEB-190398
IOS Press



# Emotion and attention: Audiovisual models for group-level skin response recognition in short movies

Álvaro García-Faura [a,*], Alejandro Hernández-García [b], Fernando Fernández-Martínez [a], Fernando Díaz-de-María [c] and Rubén San-Segundo [a]

[a] *Department of Electrical Engineering, Universidad Politécnica de Madrid, Madrid, Spain*
*E-mails: agfaura@die.upm.es, fernando.fernandezm@upm.es, ruben.sansegundo@upm.es*
[b] *Institute of Cognitive Science, Universität Osnabrück, Osnabrück, Germany*
*E-mail: ahernandez@uos.de*
[c] *Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, Spain*
*E-mail: fdiaz@tsc.uc3m.es*



**Abstract.** The electrodermal activity (EDA) is a psychophysiological indicator which can be considered a somatic marker of the emotional and attentional reaction of subjects towards stimuli. EDA measurements are not biased by the cognitive process of giving an opinion or a score to characterize the subjective perception, and group-level EDA recordings integrate the reaction of the whole audience, thus reducing the signal noise. This paper contributes to the fiel of affective video content analysis, extending previous novel work on the use of EDA as ground truth for prediction algorithms. Here, we label short video clips according to the audience's emotion (high vs. low) and attention (increasing vs. decreasing), derived from EDA records. Then, we propose a set of low-level audiovisual descriptors and train binary classifier that predict the emotion and attention with 75% and 80% accuracy, respectively. These results, along with those of previous works, reinforce the usefulness of such low-level audiovisual descriptors to model video in terms of the induced affective response.

Keywords: Electrodermal activity, emotion, attention, affective video content analysis, audiovisual descriptors

## 1. Introduction and previous work

Attention and emotion have been historically approached from many different perspectives and, thus, they have been given many different definitions Emotion is associated with affect, while attention relates to cognition and behavior. The three of them – affect, behavior and cognition – encompass our "predispositions to respond to some classes of stimuli", according to the *ABC model of attitudes* [3,8,30]. These phenomena have been studied under the light several different disciplines, such as neuroscience, psychology, medicine and even marketing.

*The Expression of the Emotions in Man and Animals* [13], published in 1872 by Charles Darwin, is still considered as one of the main publications in the field as well as being the starter of the "golden years" of emotion research [24]. Darwin suggested that there exists a set of basic emotions which is shared by all species and cultures. Concurrently, in the fina years of the XIX century, Wilhelm Wundt and William James conducted the earliest research on attention. They define attention as "taking possession by the mind, in clear and vivid form, of one of what seem several simultaneously possible objects or trains of thought" [32]. As opposed to attention, there is no consensus on how to defin emotion [43]. In 1884, William James was also the firs one to give a definitio of emotion in *What is*

*Corresponding author. E-mail: agfaura@die.upm.es.

*an emotion?* [31]. Since then, many others have been given, some of them quite recently [48]. In this work, we do not intend to explore all definition nor propose a novel one. Consequently, we have chosen the one in *Descartes' Error* by Antonio Damasio, which we consider to be simple and easily understandable: emotion is "a collection of changes occurring in both brain and body, usually prompted by a particular mental content" [12]. As two of the main traditional field in psychology, nowadays both emotion and attention keep attracting researchers' interest [37,45].

Research on psychophysiology has led to the creation of several different techniques to measure the activity of the central nervous system (CNS), making possible the study of mind expressions, including attention and emotion. Some of these techniques are functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), which have been both used in research on emotion [44], yielding in some cases to revolutionary finding [10]. Another method that has been employed in this regard is galvanic skin response (GSR), often also referred to as electrodermal activity (EDA) [16]. EDA, as opposed to the previously mentioned methods, measures the autonomic nervous system (ANS) peripheric responses using electricity. In 1888, Féré firs measured EDA using two electrodes and a galvanometer [18], which is essentially the same procedure implemented in modern devices. Eccrine glands in our skin are in charge of secreting sweat, what varies the electrical properties of our skin. The term electrodermal activity refers to these variations. It is the sympathetic ANS what controls sweat secretion, being in turn influence by the activity of our CNS. The reason why EDA is of interest in this research is because of its proven relation as somatic marker of emotion and attention [7,16,50].

Electrodermal activity has been intensively studied, as reflecte by the 1973 book *Electrodermal Activity in Psychological Research* [46]. Several different applications of EDA have been used in diverse fields psychologists measured the elicited EDA in subjects that were shown pictures of spiders and snakes shortly and demonstrated that consciousness is not required to show fear responses [41]; in medicine, symptoms of mental pathologies like schizophrenia have been predicted using EDA measures [53]; and, more recently, it is neuroscience where EDA have served as a useful research tool [55]. Furthermore, in [34,56] we can see how EDA is now being also used in neuromarketing and consumer neuroscience, showing that its potential applications are broad.

On this matter, there exist technological solutions designed specificall to carry out these kind of studies. An example to this is Sociograph, a novel EDA-based neuromarketing technology [1,38] that aims at registering the attentional and emotional response of groups of people when they are presented some kind of stimuli, for instance, during a screening of any audiovisual composition. The EDA of up to 128 subjects can be registered and processed by the Sociograph technology, which encompass both hardware and software. It outputs a signal that aggregates that of every subject, and which can be referred to as group-level electrodermal activity (EDAg).

Our aim in this paper is to use low-level characteristics of short movies to model the affective reaction they evoke in the audience. This is, we are not analyzing the psychophysiological reactions to videos but, instead, we take these reactions as ground truth to label video clips accordingly. In short, we train classifi cation models using low-level audiovisual descriptors as features and Sociograph's EDAg as ground truth for the elicited attention and emotion in viewers.

It is clear to us that the narrative structure and the semantics of movies largely influenc attention and emotion [54]. However, other formal aspects such as editing undoubtedly influenc how spectators perceive and consequently react to movies, as stated in classic film making literature [6,42]. For instance, color is a visual property that strongly influence emotions [9] as well as music clearly affects the emotional perception of film [11]. Psychology also studied these phenomena. Specificall , Zajonc proposed theories on the primacy of affect over cognition [57]. He named *preferenda* those features intrinsic to stimuli that, with no interference of cognition, interact more directly with affect.

The 1993 work by Lang *et al.* is one of the earliest studying visual stimuli using electrodermal activity [35], in which it proved to be useful to characterize emotions reported by subjects when watching pictures. In [21], a similar experiment applied to videos was carried out. Content-based descriptors and EDA were combined in [51] to model spectators' arousal and valence. DEAP [33] and MAHNOB-HCI [52] are two datasets that include EDA recordings as well as EEG and other physiological signals as reactions to 40 music videos and 20 generic videos, respectively. Nevertheless, most authors have typically focused in the prediction of the EEG signal, while the modelling of EDA responses using content-based audiovisual features has been tackled less often [4]. This paper is an

extension to our previous work on this subject [29], in which it was shown that visual descriptors extracted out of videos were useful to model attention and emotion through simple linear regression. Now, we go further in this line of research by considering also aural descriptors and addressing the problem of classifying short video clips previously labeled in terms of attention and emotion.

## 2. EDAg as ground truth

EDA can be measured accurately and easily by placing a pair of electrodes on the surface of the skin. In this regard, the palms and the feet soles are the places where human body has the highest density of eccrine sweat glands that respond to the emotional stimuli. Measurements are usually conducted on the finger by following the *exosomatic* method where, by applying a small electric current, one can record the voltage across the electrodes, which will vary directly with the skin resistance. This method is the most widely used [22], but it is also possible to follow its reciprocal *endosomatic* method.

Changes on the skin resistance depend immediately upon the amount of sweat within the sweat ducts, which can be regarded as a set of variable resistors wired in parallel [16]. In turn, such eccrine sweating (sometimes referred to as *emotional* sweating [2]) is controlled by the sympathetic innervation which transmits impulses from the CNS as autonomic responses related to "mind components such as emotion, preparation to action and vigilance processes" [50].

EDA is mainly characterized by the superimposition of two activity components: tonic and phasic. Tonic or slow changing activity is usually referred to as skin conductance *level* (SCL) or EDL. SCL values typically vary between 2 and 20 microsiemens [$\mu S$] depending on the subject. Increased sympathetic activation or alertness, which denotes more attention and predisposition to receive and analyze information, is associated with higher SCL. Phasic or fast changing activity is usually referred to as skin conductance *response* (SCR) or EDR, and arises as higher frequency variations on top of SCL. SCR peaks originate in the presence of relevant stimuli and are indicative of higher emotional state [46], i.e. higher arousal. Besides, SCR can be elicited in the absence of a stimulus and it is referred in this case as spontaneous or non-specifi (NS) activity [16,50], which can be considered as noise in the measure [39]. In this regard, the Sociograph technology utilized in our research allows the integration of the skin conductance measured on multiple subjects thus removing most NS responses. For a better illustration of the EDA signals, two examples of possible recordings are shown in Fig. 1.

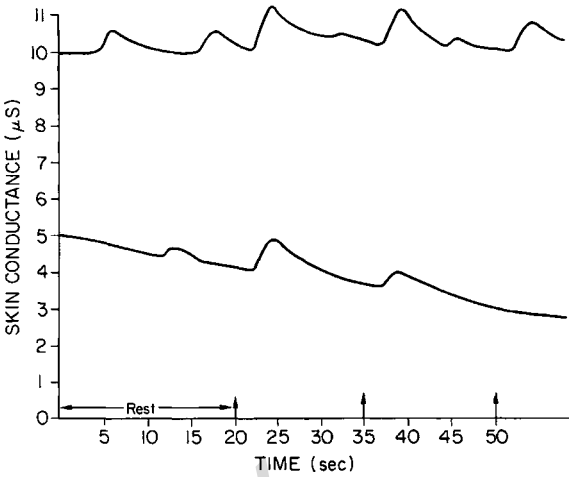


Fig. 1. Two hypothetical EDA recordings during a 20-second period of rest and 3 subsequent stimuli (arrows). SCL can be identifie as the slow changing waves and SCR as the peaks originated after the stimuli. The rest of the peaks would correspond to NS-EDA. The figur is taken from [15].

In order to collect ground truth data for the present study, the EDA was measured on participants while they watched a concatenation of videos. The age ranged from 17 to 60 years old, with an average of 23.11, and there were a 46.4% of male participants. Twelve screening sessions were held, each having a mean number of 22.5 attendees, from a total of 270 different participants. The data set consists of 136 short film from the selected short film at the Jameson Notodofilmfes Short Film Festival 2015. The whole sequence of videos was projected in a movie theater while the EDA was recorded at 1 Hz on each subject by means of the Sociograph device. Note, however, that for the present analysis the signals of each short fil were segmented into shorter clips, aiming to work with videos whose duration is closer to that of a scene rather than with the whole film Each "scene-level" clip is then considered separately. For each of them, Sociograph integrates the signals from all the participants and outputs the separate SCL and SCR signals, which represent, respectively, the attentional and emotional activation of the group along the video with a 95% confidence

### 2.1. Short fil segmentation

Short film were segmented into shorter scene-level clips by means of the SCL signal because of its straightforward interpretability and slow-varying nature. Our goal was to separate those scenes that resulted in increasing attention from the ones that made the audience lose attention.

The procedure is as follows: the firs derivative of the SCL signal is taken, so that instants above zero indicate growing attention and vice versa. However, instead of directly segmenting at those zero-crossing points, we carry out a procedure to limit segmentation noise: we firs apply a median filte and then impose guard intervals between segments, that is, a number of guard samples before and after zero-crossing points that are discarded.

To choose the filte order and the number of guard samples, we take into account the number and nature of segments that certain values for these parameters generate. Our goal is to maximize the ratio of the total number of generated segments to the difference between the number of those of increasing and decreasing attention. In other words, we pick the filte order and number of guard samples that result in the most similar number of clips with increasing and decreasing attention, while maximizing the total number of them. Any growing period of the signal is followed by a decreasing one, so it is reasonable to assume that the original distribution of segments, i.e. without any fil tering nor guard intervals, is also close to 50% of each kind.

The higher the order of the median filte , the lower the number of generated segments, though the longer their duration. Also, the higher the number of guard samples, the fewer generated segments and the shorter the duration of the remaining ones, as we force the SCL derivative to be above or below zero for at least twice the number of guard samples, which are then discarded. We carry out a tuning of the filte order varying it between 1 and 30, while the number of guard samples is manually set to 3. We do not consider a lower number of guard samples for them to fulfil their mission, i.e. not to include instants close to trend changes, when uncertainty is higher. In Fig. 2, the tuning for the filte order can be seen for 3 guard samples, which turned out to be the optimal number.

By applying a median filte of order 12 and 3 guard samples before and after each segment, 537 video clips are generated. We will refer to those video segments simply as «videos», but note that our working unit are these 537 scene-level clips rather than the complete short films Figure 3 illustrates the segmentation process for a sample short fil from our dataset.

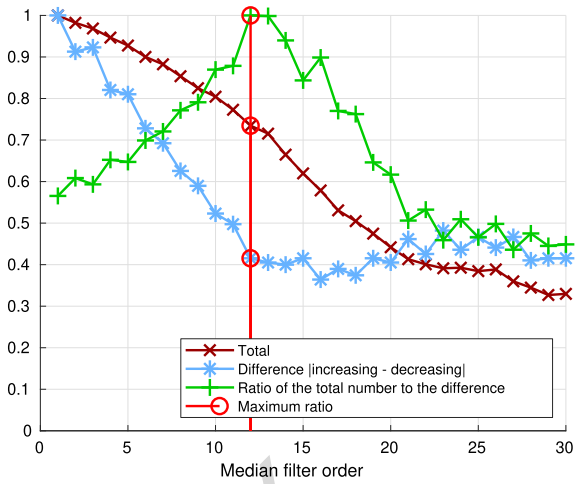


Fig. 2. Variation on the ratio of number of generated video segments to the difference between those with increasing and decreasing attention when varying the median filte order and then applying 3 guard samples before and after each zero-crossing point. For illustrative purposes, every plot has been scaled to have a maximum value of 1.

### 2.2. Annotation

For each video we obtain a ground truth of their elicited emotion and attention by computing a two-fold metric, such that we defin in both cases a binary classificatio problem:

- Sign of SCL slope ($\Delta SCL$): same criterion used to segmentate films There are clips of growing attention and clips for which attention decreases.
- Maximum SCR value ($SCR_{\max}$): the maximum SCR value reached during each clip. We use the maximum value of the signal because of its impulsive nature. Other metrics such as the mean value would imply a temporal integration of the signal, resulting in a representation not much in line with its nature. Having the whole distribution of maximum SCR values, we label segments over the median as “high” and those below as “low”, referring to the signal’s amplitude.

For the sake of readability, we will refer to these metrics simply as attention or SCL for $\Delta SCL$ and emotion or SCR for $SCR_{\max}$.

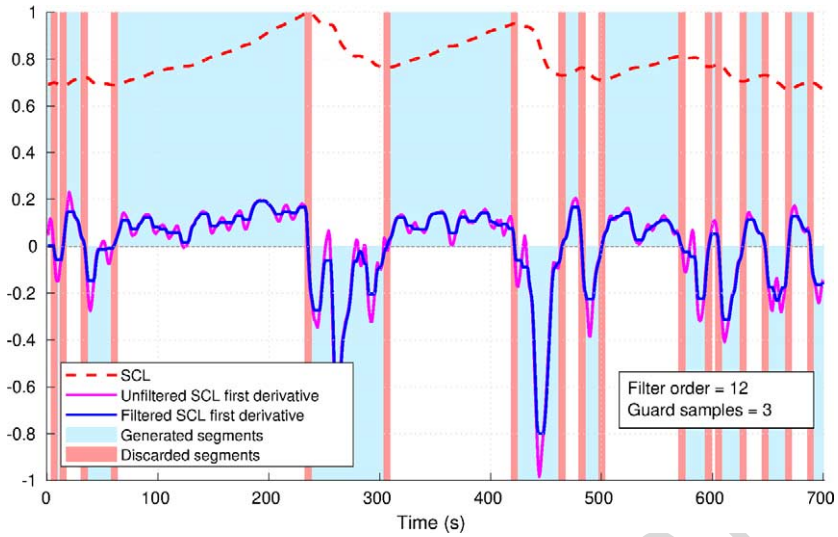


Fig. 3. Representation of the fragmentation procedure. Blue and red background colors indicate included and discarded video segments, respectively. For illustrative purposes, all signals have been scaled to have a maximum value of 1.

## 3. Descriptors

The core element of the task we are addressing, namely the automatic prediction of the attention and emotion elicited by videos, is the set of audiovisual descriptors. Although there are many factors playing a role in the success of assessing the affective value of videos, such as the machine learning algorithms or the reliability of the annotations, the chances are that the strongest influenc corresponds to the choice of the features and their reliability to represent what they aim.

### *3.1. Visual descriptors*

We extract 34 low-level visual features overall. Our set of descriptors is a blend of very simple features and others that are slightly more complex, all inspired by previous works on automatic assessment of perception [14,40] and by cinematographic and photographic aspects [6]. Some descriptors are statistical metrics (mean, deviation, etc.) of the distribution along the video of certain frame-level characteristics while others directly consider all the frames at once or are intrinsically related to the temporal nature of videos. They are the same exact descriptors that we used in our earlier work [29]. The overview of the visual descriptors, organized into 9 families, is as follows:

1. Intensity: statistical features describing the brightness of the frames.
2. Hue: statistical features related to the hue channel of the frames after conversion to the HSV color space.
3. Saturation: statistical features related to the saturation channel of the frames after conversion to the HSV color space.
4. Entropy: statistical features related to the entropy of the frames, as a measure of the amount of texture. Also, there are descriptors related to the amount of low-entropy (highly monochromatic) frames.
5. Temporal segmentation: features describing the number and duration of shots.
6. Frame-level colorfulness: statistical features about the colorfulness of the frames, which refers to the degree of utilization of richly varied colors and is computed by comparing the frame color histogram to a uniformly distributed histogram of an ideally multicolored image.
7. Video-level colorfulness: similar to the frame-level version, but considering all the pixels of the video at once for computing the histogram.
8. Color profiles features characterizing the similarity of the colors within the video to eight predefine colors: red, green, dark blue, light blue, cyan, violet, brown and gray.
9. Rule of thirds: features related to the degree of utilization of the rule of thirds (a well-known photographic composition rule) to place imaginary horizontal lines within the image.

For further details, the reader may refer to our previous work on EDA prediction [29]. Besides, in [19], some of these features were validated for aesthetics recognition and a thorough comparison of their performance was presented in [28].

### 3.2. Aural descriptors

In order to obtain aural descriptors from videos, we make use of MIRtoolbox [36]. We obtain diverse features related to the main musical dimensions of audio: dynamics, rhythm, timbre, pitch and tonality. Moreover, features built out of statistics are also obtained, such as mean, median, standard deviation or kurtosis, among others. These statistics are mostly applied to other features' spectra, envelopes or histograms. Overall, 376 aural features were extracted.

As with the visual descriptors, we also organize the aural features into families, including some psychoacoustic properties that make them especially interesting in this work:

1. Dynamics: it refers to the amplitude of the acoustic intensity variation in a musical piece. Intensity variations can influenc arousal values on the listener [17], even more than others such as *tempo* do [49]. Besides, DRC (Dynamic Range Compression) techniques can affect in a negative way the listener's emotional response [47].
2. Rhythm: it can be define as the temporal arrangement of sounds. Musically, rhythm perception is derived from the alternation of weak and strong sound elements. Rhythmic characteristics of sound cause certain reactions and emotional responses in our body. For instance, *tempo* influ ences our breathing and heart rate [5]. Also, we tend to perceive slow music as sadder than other with a higher *tempo* [23].
3. Timbre: the property of sound that enable us to distinguish between two sounds of equal pitch and loudness. Physically, it depends on the number of harmonics and their amplitude. Some experiments have shown that the perceived emotions in music compositions are independently affected by the timbre of the instruments used [26].
4. Pitch and tonality: pitch is the property of sound that is related to its frequency, making possible a disposition in scales. Usually, music compositions feature a pitch center, sometimes also known as tonic. Besides, it can be said that tonality – although it has been given several different definition – refers to a musical system or arrangement in which hierarchical relations are established between the tonic and the rest of pitches, leading to notes, scales and chords. Using this system, one can defin the key, which is an specifi scale characterized by its tonic. Typically, keys in minor mode, in which dissonant sounds predominate, are perceived as sadder than those in major mode, which are mostly characterized by consonant sounds [23].

## 4. Classificatio experiments and discussion of results

In this section, we present our experimental setup and the classificatio results obtained for both annotation strategies. There exists a wide variety of classifica tion algorithms, but rather than looking for the highest performance method, our main goal is demonstrating the validity of using audiovisual descriptors extracted from videos to predict these biometric measurements identifie with attention and emotion. For that reason, we make use a simple logistic regression classifie . As a baseline, we consider a ZeroR classifie , which always predicts the majority class in training data without relying on any feature. All the results presented in this work are the average of 10 repetitions of a 10-fold cross-validation procedure, and include a 95% confi dence interval. Furthermore, we also compare the results when only visual or aural features are used, and the combination of both types of descriptors. The feature selection procedures as well as the classificatio experiments have been carried out using the machine learning software Weka [27].

As already mentioned, we initially extract 34 visual and 376 aural features. Since this is a considerably high number of descriptors, we carry out a careful procedure of feature selection in order to reduce the dimensionality of the dataset, discard the less relevant or redundant descriptors, and make the predictions with the most suitable feature subsets. For this purpose, we make use of the *SVMAttributeEval* Weka functionality, which employs a Support Vector Machine (SVM) classifie to determine the worth of an attribute. These are ranked by the square of the weight assigned by the SVM, so that we are able to choose how many features we would like to include, starting with the top one. Please, refer to [25] for further information on this attribute selection procedure.

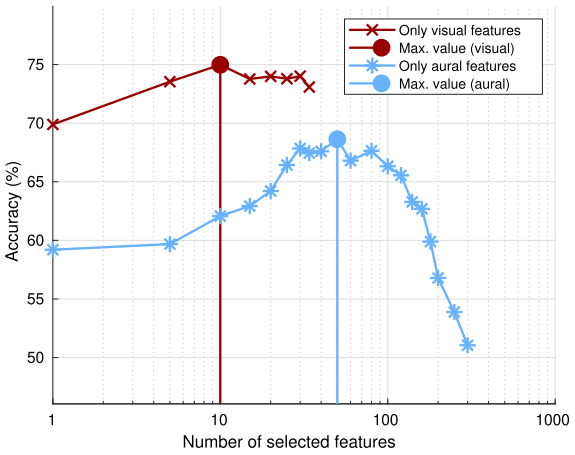


Fig. 4. Accuracy for attention classificatio when using only visual or only aural features to train a logistic regression classifie .

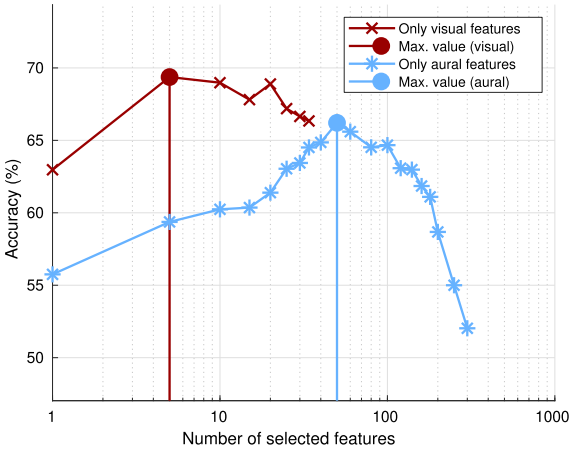


Fig. 5. Accuracy for emotion classificatio when using only visual or only aural features to train a logistic regression classifie .

### 4.1. *Experiments using aural or visual features separately*

One of our goals is to determine what type of attributes are more valuable when trying to model attention or emotion, as well as how many of them are needed. For such purpose, we have separately analyzed how many attributes yield the best accuracy when predicting attention and emotion. We start by choosing only 1 attribute, and then gradually increase the number of selected features up to 34 and 300, respectively. No more than 300 aural features were considered since results clearly showed that the optimal working point was far from that high number of features.

In Fig. 4, it can be observed that when predicting attention, 10 visual features are needed to achieve the best result, an accuracy of 74.98%. With respect to the aural case, the best accuracy obtained is 69.83%, for which 50 aural features were needed.

The results of the emotion prediction are included in Fig. 5. Again, 50 aural features lead to the best result for audio information, 66.21% accuracy and only 5 visual features are needed to reach a 69.83% accuracy. The summary of these results for both categories is presented in Table 1.

### 4.2. *Experiments combining aural and visual features*

Once we have observed how the two types of descriptors behave separately for predicting attention and emotion, we analyze whether the combination of both leads to better results. We have seen that 10 visual features led to the best result for attention and the second best for emotion. Because of that, we will firs experiment by combining the 10 best visual and 10 best aural features. Note that the features are ranked independently for attention and emotion. This means that, even though we keep the 10 best features of each type in both cases, they are not necessarily be the same. Results for these experiments are shown in Table 2.

Table 1

Maximum accuracy obtained when predicting attention and emotion using only visual or only aural features

| Type of features | Attention (SCL) | Emotion (SCR) |
|---|---|---|
| Visual (# features) | 74.98 ± 3.66% (10) | 69.37 ± 3.90% (5) |
| Aural (# features) | 69.83 ± 3.88% (50) | 66.21 ± 4.00% (50) |

Both for attention and emotion, the obtained accuracy is essentially the same as for the case when using only visual features. The addition of the 10 best aural features is not enough to significantl improve the single-modal results.

Secondly, we combine the 10 best visual features with the 50 best aural features, as these obtained the highest accuracy for the audio-only case. This way we are able to study how the inclusion of more audio features affects the previous experiment. Note that features from each type are ranked independently and then combined. In Table 2, we include the obtained accuracy for this experimental setup.

In this case, we observe an accuracy increase of 3.76 for the prediction of attention with respect to the previous experiment, while the emotion prediction only exhibits a minor improvement. If we compare these results to the single-modality ones, we can see that the result for attention is statistically different from that

Table 2
Accuracy obtained when using a number of features selected beforehand

| Classifie | No. of visual features | No. of aural features | Acc. for attention (SCL) | Acc. for emotion (SCR) |
|---|---|---|---|---|
| ZeroR | 0 | 0 | 57.54 ± 4.18% | 49.42 ± 4.23% |
| Logistic | 10 | 10 | 74.91 ± 3.67% | 70.65 ± 3.85% |
| Logistic | 10 | 50 | 78.67 ± 3.46% | 71.98 ± 3.80% |

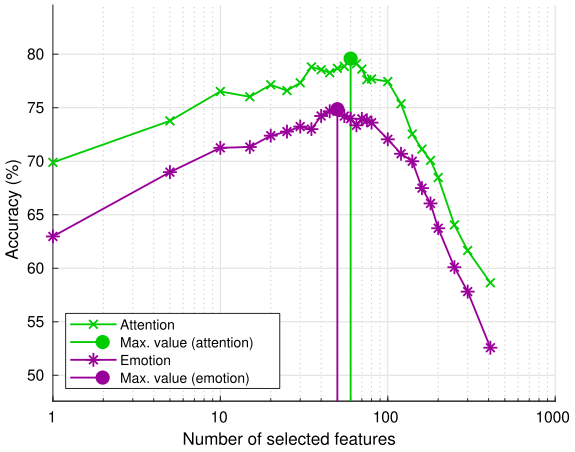


Fig. 6. Accuracy for attention and emotion classificatio when using both visual and aural features to train a logistic regression classifie .

obtained using only aural features. This shows the importance of visual features to model attention: while using the same 50 aural features, a significan increase in accuracy is obtained if considering only 10 additional visual features. For the rest of setups, the improvement is not that wide as to consider it statistically significant

Finally, we perform a similar feature selection process, in this case combining both types of attributes beforehand, such that the algorithm can decide which ones to pick rather than manually forcing a certain number of each kind. The maximum number of features that could be used is 410, resulting from 34 visual and 376 aural features. We seek with fine granularity in values that are likely to be close to the maximum accuracy value, and have reduced granularity for a larger amount of features. Exploring all the possible sets of features could result in an increment of the maximum accuracy, but it would probably be minor, given the previous results.

In Fig. 6, we include the accuracy plots for these experiments for attention and emotion. Besides, Table 3 summarizes them, including the highest accuracy values both for attention and emotion predictions.

With this approach we obtain the best results: 79.59% accuracy for attention prediction and a 74.86% accuracy for emotion prediction. Besides, we can see that the number of selected features of each kind is approximately the same as when selecting them independently. When predicting attention, they match exactly: 10 visual and 50 aural features. In the emotion case, they are also quite similar: 8 visual and 42 aural features, while independently maximum accuracy was obtained for 5 visual and 50 aural features. Since we did not explore all the possibilities, it is reasonable to think that these concordances would be even higher if we carried out that exploration.

Table 3
Maximum accuracy obtained with a logistic regression classifie and using features resulting from the combined selection procedure

| | Attention (SCL) | Emotion (SCR) |
|---|---|---|
| Accuracy | 79.59 ± 3.41% | 74.86 ± 3.67% |
| No. of visual features | 10 | 8 |
| No. of aural features | 50 | 42 |

### 4.3. Discussion

In Tables 2 and 3, we compare the results when using a simple logistic regression classifie with a reference value given by a ZeroR classifie . In all cases, the accuracy improvement with respect to this reference algorithm has been significant up to a 25.44% for emotion prediction and to a 22.05% for attention prediction.

Also, all experiments have confirme that visual features are much more useful to model these signals derived from EDA than aural features. With approximately 10 visual features we could obtain higher accuracy rates than with 50 aural features, which were the ones needed to provide maximum accuracy in both single-modality cases. However, both types of features are complementary, since when combined, better results are obtained.

As we have seen, all our visual features have been specificall tailored to represent aspects that are supposed to be related to film-makin aspects and thus affect attention and emotion. However, most aural features extracted by MIRToolbox are related to music and, though music plays a crucial role in movies, it

may not be present in all clips. Besides, there are other acoustics events such as dialogues that could largely influenc viewers' reaction and that are not related to music.

The combination of features does not significantl improve the results of the visual descriptors alone. This is not the case for aural features, whose results are significantl improved when combined with the more powerful visual descriptors.

When comparing results for attention and emotion, we see that in all cases attention is predicted considerably better. Since video segments were generated out of the SCL signal, and the labeling for attention was done accordingly, it is arguable that the accuracy differences are influence by the segmentation procedure: while the SCL signal shows the same trend for all the duration of the clip, the emotion label is only derived from a peak, which is likely to be a very short event if compared with the full clip. Even so, the results for emotion should not be disregarded, as almost a 75% accuracy is obtained when classifying clips that presented or not a particularly intense emotional response.

## 5. Conclusions

While electrodermal activity (EDA) has been used for long time in psychology and medicine, and more recently in neuroscience and neuromarketing, as a way of measuring the reaction of people towards stimuli, little research has been devoted by the affective computing fiel to the stimuli side relying on EDA as ground truth for emotion and attention assessment. In this paper, we use videos as stimuli, and we investigate whether it is possible to predict EDA responses in a group of people by means of the audiovisual characteristics of the videos.

We make use of Sociograph, a neuromarketing technology that integrates the EDA responses of many individuals and derives, for each video, a value of attention (SCL) and emotion (SCR). We use the SCL signal to segment videos in shorter clips, which are then labeled with regard to the attention slope (increasing vs. decreasing) and the maximum value of emotion or SCR (high vs. low). Then, we extract a set of low-level audiovisual descriptors from the videos and train a logistic regression to classify the EDA responses.

We have shown empirically that both visual and aural descriptors are independently able to predict attention and emotion elicited by short clips. This demonstrates that audiovisual characteristics of videos, such as the brightness, color, rhythm or pitch center, among others, have considerable influenc on the automatic and unconscious human emotional and attentional reactions. Moreover, we have also observed that few visual descriptors achieve slightly better results than several aural features.

We have explored several ways of combining aural and visual features, being the best option to rank all of them together using a feature selection procedure and choose the number of features that leads to best results. This way, we reach a top 79.59% accuracy for attention prediction and 74.86% accuracy for emotion prediction. These results suggest that audiovisual descriptors are useful for modelling attention and emotion derived from EDA. It is also worth mentioning that other more complex classifier could have led to better results, but it was not our goal to look for the best classificatio algorithm, but to show how a simple one can successfully predict both attention and emotion by using valid and informative audiovisual features.

The relationship between visual descriptors and other kind of subjective information like aesthetics or appeal reported deliberately by participants via a score, for instance, had been already demonstrated in previous works [19]. However, findin that signals derived from EDA can be modeled using the audiovisual features of stimuli has a great interest because it is a psychophysiological reaction controlled by the autonomic nervous system, thus it is automatic and is directly related to actual emotional and attentional activation, avoiding the implicit bias of opinions and judgments.

As stated before, the better results obtained for attention with respect to emotion in this work could be influence by the fact that attention was also employed to generate the scene-level clips we use. In the future, it would be interesting to create a segmentation procedure that took into consideration aspects of both SCL and SCR. Furthermore, since all the videos in the dataset are uploaded to YouTube, we could compare how the audience reaction in terms of EDA correlates with the conscious opinion given by the online community, following a similar procedure to the one in [20]. Descriptors derived from YouTube such as likes and dislikes, number of views or number of comments, could also be combined with the audiovisual descriptors we propose in order to create a more accurate model of attention and emotion.

EDA might become a popular measure of ground truth annotation for video content analysis in future research, since it is a relatively straight-forward and non-expensive method for capturing the emotional states

of the human mind, in comparison to more complex methods like fMRI. This work, along with the one it extends [29], reveals which and how audiovisual descriptors are useful to model viewers' reaction to audiovisual stimuli in terms of EDA.

## Acknowledgements

This research has been possible thanks to the collaboration of Sociograph Neuromarketing, particularly in providing the data set and annotations.

The work leading to these results has been supported by ESITUR (MINECO, RTC-2016-5305-7), CAVIAR (MINECO, TEC2017-84593-C2-1-R), and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 641805.

## References

[1] M. Aiger, M. Palacín and J.-M. Cornejo, Electrodermal signal by sociograph: Methodology to measure the group activity, *Revista de Psicología Social* **28**(3) (2013), 333–347. doi:10.1174/021347413807719102.

[2] M. Asahina, A. Suzuki, M. Mori, T. Kanesaka and T. Hattori, Emotional sweating response in a patient with bilateral amygdala damage, *International Journal of Psychophysiology* **47**(1) (2003), 87–93. doi:10.1016/S0167-8760(02)00123-X. http://www.sciencedirect.com/science/article/pii/S016787600200123X.

[3] M. Augoustinos, I. Walker and N. Donaghue, *Social Cognition: An Integrated Introduction*, Sage, 2014, pp. 114–115.

[4] D. Ayata, Y. Yaslan and M. Kamaşak, Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches, 2016, pp. 1–4. doi:10.1109/TIPTEKNO.2016.7863130.

[5] L. Bernardi, C. Porta and P. Sleight, Cardiovascular, cerebrovascular, and respiratory changes induced by different types of music in musicians and non-musicians: The importance of silence, *Heart* **92**(4) (2006), 445–452. doi:10.1136/hrt.2005.064600.

[6] D. Bordwell, K. Thompson and J. Ashton, *Film Art: An Introduction*, Vol. 7, McGraw-Hill, New York, 1997.

[7] W. Boucsein, *Electrodermal Activity*, Springer Science & Business Media, 2012. doi:10.1007/978-1-4614-1126-0.

[8] S.J. Breckler, Empirical validation of affect, behavior, and cognition as distinct components of attitude, *Journal of Personality and Social Psychology* **47**(6) (1984), 1191–1205. doi:10.1037/0022-3514.47.6.1191.

[9] C. Brinckmann, *Color and Empathy: Essays on Two Aspects of Film*, Amsterdam University Press, 2014. doi:10.1515/9789048523269.

[10] R. Cabeza and L. Nyberg, Imaging cognition II: An empirical review of 275 PET and fMRI studies, *J. Cognitive Neuroscience* **12**(1) (2000), 1–47. doi:10.1162/08989290051137585.

[11] A.J. Cohen, Music as a source of emotion in film in: *Music and Emotion: Theory and Research*, 2001, pp. 249–272.

[12] A.R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Avon Books, New York, 1994.

[13] C. Darwin, *The Expression of the Emotions in Man and Animals*, John Marry, London, UK, 1872.

[14] R. Datta, D. Joshi, J. Li and J.Z. Wang, Studying aesthetics in photographic images using a computational approach, in: *Proceedings of the 9th European Conference on Computer Vision – Volume Part III, ECCV'06*, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 288–301. ISBN 3-540-33836-5, 978-3-540-33836-9. doi:10.1007/11744078_23.

[15] M.E. Dawson and K.H. Nuechterlein, Psychophysiological dysfunctions in the developmental course of schizophrenic disorders, *Schizophrenia Bulletin* **10**(2) (1984), 204–232. doi:10.1093/schbul/10.2.204. http://schizophreniabulletin.oxfordjournals.org/content/10/2/204.abstract.

[16] M.E. Dawson, A.M. Schell and D.L. Filion, The electrodermal system, in: *Handbook of Psychophysiology*, J.T. Cacioppo, L.G. Tassinary and G.G. Berntson, eds, 2nd edn, Cambridge University Press, Cambridge, 2000, pp. 200–223.

[17] R.T. Dean, F. Bailes and E. Schubert, Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation, *PloS One* **6**(4) (2011), e18591.

[18] C. Féré, Note sur les modification de la résistance électrique sous l'influenc des excitations sensorielles et des émotions, *CR Soc. Biol* **5** (1888), 217–219.

[19] F. Fernández-Martínez, A. Hernández-García and F. Díaz-de-María, Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials, *Expert Systems with Applications* (2015), 293–305. doi:10.1016/j.eswa.2014.07.033.

[20] F. Fernández-Martínez, A. Hernández-García, A. Gallardo-Antolín and F.D. de María, Combining audio-visual features for viewers' perception classificatio of Youtube car commercials, in: *Proceedings of Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2014.

[21] J. Fleureau, P. Guillotel and I. Orlac, Affective benchmarking of movies based on the physiological responses of a real audience, in: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 2013, pp. 73–78, doi:10.1109/ACII.2013.19. ISSN 2156-8103.

[22] D.C. Fowles, M.J. Christie, R. Edelberg, W.W. Grings, D.T. Lykken and P.H. Venables, Publication recommendations for electrodermal measurements, *Psychophysiology* **18**(3) (1981), 232–239. doi:10.1111/j.1469-8986.1981.tb03024.x.

[23] L. Gagnon and I. Peretz, Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies, *Cognition and emotion* **17**(1) (2003), 25–40. doi:10.1080/02699930302279.

[24] M. Gendron and L.F. Barrett, Reconstructing the past: A century of ideas about emotion in psychology, *Emotion review* **1**(4) (2009), 316–339. doi:10.1177/1754073909338877.

[25] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classificatio using support vector machines, *Mach. Learn.* **46**(1–3) (2002), 389–422. doi:10.1023/A:1012487302797.

[26] J.C. Hailstone, R. Omar, S.M. Henley, C. Frost, M.G. Kenward and J.D. Warren, It's not what you play, it's how you play it: Timbre affects perception of emotion in music, *The quarterly Journal of Experimental psychology* **62**(11) (2009), 2141–2155. doi:10.1080/17470210902765957.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: An update, *ACM SIGKDD explorations newsletter* **11**(1) (2009), 10–18. doi:10.1145/1656274.1656278.

[28] A. Hernández-García, F. Fernández-Martínez and F. Díaz-de-María, Comparing visual descriptors and automatic rating strategies for video aesthetics prediction, *Signal Processing: Image Communication* **47** (2016), 280–288.

[29] A. Hernández-García, F. Fernández-Martínez and F. Díaz-de-María, Emotion and attention: Predicting electrodermal activity through video visual descriptors, in: *Proceedings of the International Conference on Web Intelligence*, ACM, 2017, pp. 914–923. doi:10.1145/3106426.3109418.

[30] E.R. Hilgard, The trilogy of mind: Cognition, affection, and conation, *Journal of the History of the Behavioral Sciences* **16**(2) (1980), 107–117. doi:10.1002/1520-6696(198004)16:2<107::AID-JHBS2300160202>3.0.CO;2-Y.

[31] W. James, What is an emotion?, *Mind* **34** (1884), 188–205. doi:10.1093/mind/os-IX.34.188.

[32] W. James, *The Principles of Psychology*, Dover, New York, 1890.

[33] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt and I. Patras, DEAP: A database for emotion analysis using physiological signals, *IEEE Transactions on Affective Computing* **3**(1) (2012), 18–31. doi:10.1109/T-AFFC.2011.15.

[34] M. Lajante, O. Droulers, T. Dondaine and D. Amarantini, Opening the "black box" of electrodermal activity in consumer neuroscience research., *Journal of Neuroscience, Psychology, and Economics* **5**(4) (2012), 238. doi:10.1037/a0030680.

[35] P.J. Lang, M.K. Greenwald, M.M. Bradley and A.O. Hamm, Looking at pictures: Affective, facial, visceral, and behavioral reactions, *Psychophysiology* **30**(3) (1993), 261–273. doi:10.1111/j.1469-8986.1993.tb03352.x.

[36] O. Latrillot and P. Toiviainen, MIR in Matlab: A toolbox for musical feature extraction, in: *Proceedings of the International Conference on Music Information Retrieval*, 2007.

[37] J. LeDoux, Rethinking the emotional brain, *Neuron* **73**(4) (2012), 653–676. doi:10.1016/j.neuron.2012.02.004.

[38] J.L. Martínez and E. Garrido, Sistema para la medición de reacciones emocionales en grupos sociales. 2 168 928, 2003.

[39] J.L. Martínez, S. Monge and M.I. Valdunquillo, Medición de las respuestas psicofisiológica grupales para apoyar el análisis de discursos políticos, *Trípodos* **29** (2012), 53–72.

[40] A.K. Moorthy, P. Obrador and N. Oliver, Towards computational models of the visual aesthetic appeal of consumer videos, in: *Proceedings of the 11th European Conference on Computer Vision: Part V, ECCV'10*, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–14. ISBN 3-642-15554-5, 978-3-642-15554-3. http://dl.acm.org/citation.cfm?id=1888150.1888152.

[41] A. Öhman, Distinguishing unconscious from conscious emotional processes: Methodological considerations and theoretical implications, in: *Handbook of Cognition and Emotion*, John Wiley & Sons, Ltd, 1999, pp. 321–352. ISBN 9780470013496. doi:10.1002/0470013494.ch17.

[42] M. Ondaatje and W. Murch, *The Conversations: Walter Murch and the Art of Editing Film*, A&C Black, 2002.

[43] L. Pessoa, On the relationship between emotion and cognition, *Nature Reviews Neuroscience* **9**(2) (2008), 148–158. doi:10.1038/nrn2317.

[44] K.L. Phan, T. Wager, S.F. Taylor and I. Liberzon, Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI, *Neuroimage* **16**(2) (2002), 331–348. doi:10.1006/nimg.2002.1087.

[45] M.I. Posner, M.R. Rueda and P. Kanske, Probing the mechanisms of attention, in: *Handbook of Psychophysiology*, J.T. Cacioppo, L.G. Tassinary and G. Berntson, eds, 3rd edn, Cambridge University Press, 2007, pp. 410–432, Cambridge Books Online. ISBN 9780511546396. doi:10.1017/CBO9780511546396.018.

[46] W.F. Prokasy and D.C. Raskin (eds), *Electrodermal Activity in Psychological Research*, Academic Press, 1973. ISBN 978-0-12-565950-5. doi:10.1016/B978-0-12-565950-5.50001-6.

[47] M. Ronan, R. Sazdov and N. Ward, Loudness normalisation: Paradigm shift or placebo for the use of hyper-compression in pop music?, in: *International Computer Music Conference*, 2014.

[48] K.R. Scherer, What are emotions? And how can they be measured?, *Social science information* **44**(4) (2005), 695–729. doi:10.1177/0539018405058216.

[49] E. Schubert, Modeling perceived emotion with continuous musical features, *Music Perception: An Interdisciplinary Journal* **21**(4) (2004), 561–585. doi:10.1525/mp.2004.21.4.561.

[50] H. Sequeira, P. Hot, L. Silvert and S. Delplanque, Electrical autonomic correlates of emotion, *International journal of psychophysiology* **71**(1) (2009), 50–56. doi:10.1016/j.ijpsycho.2008.07.009.

[51] M. Soleymani, G. Chanel, J.J.M. Kierkels and T. Pun, Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses., in: *International Symposium on Multimedia*, IEEE Computer Society, 2008, pp. 228–235, ISBN 978-0-7695-3454-1. doi:10.1109/ISM.2008.14. http://dblp.uni-trier.de/db/conf/ism/ism2008.html#SoleymaniCKP08.

[52] M. Soleymani, J. Lichtenauer, T. Pun and M. Pantic, A multimodal database for affect recognition and implicit tagging, *IEEE Transactions on Affective Computing* **3**(1) (2012), 42–55. doi:10.1109/T-AFFC.2011.25.

[53] K.L. Subotnik, A.M. Schell, M.S. Chilingar, M.E. Dawson, J. Ventura, K.A. Kelly, G.S. Hellemann and K.H. Nuechterlein, The interaction of electrodermal activity and expressed emotion in predicting symptoms in recent-onset schizophrenia, *Psychophysiology* **49**(8) (2012), 1035–1038. doi:10.1111/j.1469-8986.2012.01383.x.

[54] E.S. Tan (ed.), *Emotion and the Structure of Narrative Film: Film as an Emotion Machine*, Lawrence Erlbaum Associates, Inc., 1996.

[55] D. Tranel, Electrodermal activity in cognitive neuroscience: Neuroanatomical and neuropsychological correlates, in: *Cognitive Neuroscience of Emotion. Series in Affective Science*, Oxford University Press, 2000, pp. 192–224.

[56] Y.J. Wang and M.S. Minor, Validity, reliability, and applicability of psychophysiological techniques in marketing research, *Psychology & Marketing* **25**(2) (2008), 197–232. doi:10.1002/mar.20206.

[57] R.B. Zajonc, Feeling and thinking: Preferences need no inferences., *American Psychologist* **35**(2) (1980), 151–175. doi:10.1037/0003-066x.35.2.151.