# Emotion and attention: predicting electrodermal activity through video visual descriptors

A. Hernández-García
Universität Osnabrück
Institute of Cognitive Science
Osnabrück, Germany
ahernandez@uos.de

F. Fernández-Martínez
Universidad Politécnica de Madrid
Dept. Electrical Engineering
Madrid, Spain
fernando.fernandezm@upm.es

F. Díaz-de-María
Universidad Carlos III de Madrid
Signal Theory and Communications
Madrid, Spain
fdiaz@tsc.uc3m.es

## ABSTRACT

This paper contributes to the field of affective video content analysis through the novel employment of electrodermal activity (EDA) measurements as ground truth for machine learning algorithms. The variation of the electrical properties of the skin, known as EDA, is a psychophysiological indicator widely used in medicine, psychology and neuroscience which can be considered a somatic marker of the *emotional* and *attentional* reaction of subjects towards stimuli. One of its main advantages is that the recorded information is not biased by the cognitive process of giving an opinion or a score to characterize the subjective perception. In this work, we predict the levels of emotion and attention, derived from EDA records, by means of a small set of low-level visual descriptors computed from the video stimuli. Linear regression experiments show that our descriptors predict significantly well the sum of emotion and attention levels, reaching a coefficient of determination $R^2 = 0.25$. This result sets a promising path for further research on the prediction of emotion and attention from videos using EDA.

## KEYWORDS

electrodermal activity, emotion, attention, affective video content analysis, visual descriptors

**ACM Reference format:**
A. Hernández-García, F. Fernández-Martínez, and F. Díaz-de-María. 2017. Emotion and attention: predicting electrodermal activity through video visual descriptors. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 10 pages.
https://doi.org/10.1145/3106426.3109418

## 1 INTRODUCTION AND PREVIOUS WORK

The question of how humans react towards stimuli has been addressed from many and varied perspectives. Psychology, medicine, neuroscience and even marketing are some of the fields which have shown the most interest, particularly in exploring the mechanisms of emotion and attention. According to the so-called *ABC model of attitudes* [3, 7, 22] of social psychology, our "predispositions to respond to some classes of stimuli" or attitudes [38], comprise three components: affect, behavior and cognition. These, in turn, at a simple level can be related to emotion (affect) and attention (behavior and cognition).

Emotion and attention are some of the oldest fields in psychology and they still keep on receiving even increasing interest [28, 36]. The "golden years" of emotion research [20] began with the famous Darwin's 1872 publication of *The Expression of the Emotions in Man and Animals* [11], one of whose contributions was the proposal of a set of basic emotions common to all species and cultures. Likewise, the earliest psychological research works on attention were carried out by the end of the XIX century by Wilhelm Wundt and William James, who defined 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" [24]. The definition of emotion, however, is not easy and there is no consensus about it [35]. Many attempts have been done to define emotion, from the first one by James in his essay *What is an emotion?* [23] to more recent ones [40]. Since proposing a new definition is clearly out of the scope of this paper, for plainness we will pick a simple but enlightening one, given by Antonio Damasio in his celebrated *Descartes' Error*: emotion is "a collection of changes occurring in both brain and body, usually prompted by a particular mental content" [10].

So as to study emotion and attention as well as other mind expressions, there exists a wide range of methods. The last few decades have seen the development of sophisticated measures of the central nervous system (CNS) activity, like hemodynamic and magnetic responses (fMRI), which have yielded revolutionary findings [9]. However, autonomic peripheral electrical measures have also been successfully employed as indicators of psychological states. In particular, electrodermal activity (EDA), also known as galvanic skin response (GSR), has been one of the most extensively utilized methods in psychophysiology [14] after it was first measured by Féré more than 125 years ago [15]. Essentially, electrodermal activity refers to the variation of the electrical properties of the skin due to the secretion of sweat by eccrine glands. Such secretion is under the control of the sympathetic autonomous nervous system (ANS) which, in turn, is influenced by the CNS. What is of most interest in this context is that these electrical variations are considered somatic markers of emotional and attentional expressions, among others [6, 14, 41].

The literature and history of electrodermal activity is fruitful and long, as the 1973 book *Electrodermal Activity in Psychological Research* reflects [37]. EDA has been applied in multiple fields: for instance, in medicine it has been used to predict symptoms of pathologies like schizophrenia [45]; in psychology to demonstrate

*WI '17, August 23-26, 2017, Leipzig, Germany*

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3109418

that fear responses do not require consciousness, studying elicited EDA when subjects were shown very shortly pictures of spiders and snakes [32]; and far from being given up, today it has found a perfect ally in neuroscience [47]. Besides, the modern fields of consumer neuroscience and neuromarketing are making extensive use of EDA, as it is reviewed and validated in [26, 48].

Within this context, a new neuromarketing technology based on EDA, Sociograph [1, 29], emerged aiming at analyzing the emotional and attentional response of groups during specific activities, for example the visualization of audiovisual content. Sociograph is a combination of hardware and software able to register the EDA of up to 128 subjects and process the signals to output a single response referred to as group electrodermal activity (EDAg), related to the emotion and attention experienced by the subjects while they watch the videos.

Whereas the above mentioned approaches typically analyze the psychophysiological reactions to stimuli, in this paper we propose focusing on the stimuli side, videos in this case, considering the reactions as ground truth. We aim at analyzing the potential of low-level characteristics of videos for predicting affective reaction they evoke in the viewers and, as an ultimate goal, to find out how. More specifically, we train a linear regression model using low-level visual descriptors as features and the EDAg as ground truth for the emotion and attention elicited by the videos.

While it is true that emotion and attention are largely influenced by the semantics and the narrative structure of movies [46], formal aspects like film editing also play a key role in how the spectators react, as classics of film-making literature explain [5, 34]. By way of illustration, it is well-known that colors have a strong influence on the emotions [8]. Moreover, we can find grounds for this approach in psychology as well, in the theories on the primacy of affect over cognition proposed by Zajonc [50], who even gave a name, *preferenda*, to those features intrinsic to stimuli which interact more readily with affect, without interference of cognition.

Earliest works employing EDA in the context of visual stimuli date back to 1993 when Lang *et al.* validated its use to characterize the emotions reported by subjects when watching pictures [27]. A similar validation with video content was carried out more recently in [18]. In [43] EDA and content-based descriptors were brought together to model arousal and valence provided by spectators. The MAHNOB-HCI data base [44] and DEAP [25] included EDA recordings together with EEG and other physiological signals from 20 videos and 40 music videos respectively. However, typical uses of these data bases are on predicting the EEG signal, whereas very few attempts have been done to model the EDA responses through content-based audiovisual descriptors [4]. Therefore, to our knowledge, this is one of the first approaches to the problem of affective video content analysis using EDA ground truth data and a set of low-level visual features.

## 2 EDAG AS GROUND TRUTH

Electrodermal activity can be measured by placing a pair of electrodes on the surface of the skin, usually on the fingers. Then, by applying a small current, one can record the voltage across the electrodes, which will vary directly with the skin resistance. This method is called *exosomatic* and is the most widely used [19], 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 [14]. 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” [41].

EDA is most often treated as the superimposition of two components: tonic and phasic activity. Tonic activity (slow changing) is usually referred to as skin conductance *level* (SCL) or EDL. It ranges from 2–20 microsiemens [$\mu S$] depending on the subject. Phasic activity (fast changing) is usually referred to as skin conductance *response* (SCR) or EDR and arises as higher frequency variations on top of SCL. Higher SCL is indicative of increased sympathetic activation or alertness, which denotes more attention and predisposition to receive and analyze information. SCR peaks originate in the presence of relevant stimuli and are indicative of higher emotional state [37], i.e. higher arousal. Additionally, SCR can be elicited in the absence of a stimulus and it is referred in this case as spontaneous or non-specific (NS) activity [14, 41], which can be considered as noise in the measure [30]. In this regard, one of the advantages of the integration of the skin conductance measured on multiple subjects carried out by Sociograph is that it allows removing most NS responses. By way of better illustration of EDA signals, two examples of possible recordings are shown in Figure 1.

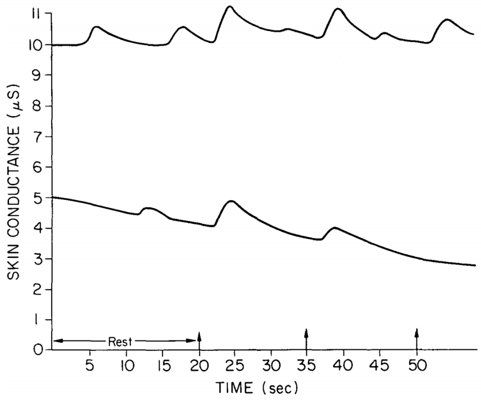


**Figure 1: Two hypothetical EDA recordings during a 20-second period of rest and 3 subsequent stimuli (arrows). SCL can be identified as the slow changing waves and SCR as the peaks originated after the stimuli. The rest of the peaks would correspond to NS-EDA. The figure is taken from [13].**

In order to collect ground truth data for the present study, the EDA was measured on 22 subjects, 10 men and 12 women, with ages between 21 and 59 years old, while they watched a concantenation of videos. The data set consists of 44 videos with an average duration of 44 seconds ($\sigma = 21$) which belong to the awarded spots at the 2002 Cannes International Advertising Festival. The whole sequence of spots was projected in a movie theater while the EDA was recorded on each subject by means of the Sociograph device. Note, however, that for the present analysis the signals of each spot are 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. In order to annotate the videos, 3 metrics were computed for each of them as ground truth:

- Standardized average rate of change of the SCL ($\Delta SCL_{av}$): it represents the average global attention captured along the duration of the video, normalized as standard scores.
- Standardized average SCR ($SCR_{av}$): it represents the average emotional activation elicited by the whole video, normalized likewise.
- Sum of $\Delta SCL_{av}$ and $SCR_{av}$ ($SUM$): as a sum of both previous activation measures, it can be regarded as an indicator of the global impact prompted by the video as it combines emotion and attention responses.

For the sake of readability, we will refer to these metrics simply as attention or SCL for $\Delta SCL_{av}$, emotion or SCR for $SCR_{av}$ and (global) impact for SUM.

## 3 VISUAL 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 visual 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 quality of the annotations, the chances are that the strongest influence corresponds to the choice of the features and their reliability to represent what they aim.

We have extracted 31 low-level visual features overall. Our set of descriptors is a blend of very simple features and some slightly more complex ones inspired by previous works on automatic assessment of perception [12, 31] and by cinematographic and photographic aspects [5]. 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. In order to more clearly present an overview of the extracted descriptors we have organized them into 9 families:

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 predefined 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.

The full list and some details about the definition and the computation procedures of the descriptors has been included for completeness in Appendix A. Some of these features have been validated for aesthetics recognition in [16], where more details can be found and a comparison of their performance was presented in [21].

## 4 REGRESSION EXPERIMENTS AND DISCUSSION OF RESULTS

The most ambitious (but also most challenging) goal we can address with these data is predicting the values of SCL, SCR and SUM by means of regression. There exists a wide variety of regression algorithms, but rather than looking for the highest performance method, our main goal is first demonstrating that it is possible to predict these biometric measurements identified with attention and emotion making use of a small set of low-level visual descriptors extracted from videos. For this reason and given also that the data set is not very large we opt for performing least squares linear regression, one of the simplest regression methods. Besides, simplicity in this case plays in favor of interpretability, which is a very desired property in this type of problem.

### 4.1 Feature Selection

Nonetheless, we do not simply perform linear regression directly on the data set, but we carry out a careful procedure of feature selection in order to reduce the dimensionality of the data set, discard the less relevant descriptors and make the predictions with the most suitable feature subsets. For this purpose we propose a cross-validated filter-based feature selection scheme that uses a correlation metric to rank the features.

Given a set of $N$ examples (videos) $X$ consisting of $D$ dimensions (descriptors), and their corresponding annotations $Y$, the steps of the feature selection procedure are the following:

1. Compute the sample cross correlation between $Y$ and every feature $j$:

$$r_{xy_j} = \frac{\sum_{i=1}^{N} X_{ij} Y_i}{\sqrt{\sum_{i=1}^{N} X_{ij}^2 \sum_{i=1}^{N} Y_i^2}}$$

2. Convert the correlation into an F-statistic according to an F-test:

$$F_j = \frac{(R_j^2 - R_r^2)(N - k)}{(1 - R_j^2)m}$$

where:
$R_j^2$: coefficient of determination of each unrestricted model, consisting of one single coefficient (corresponding to one descriptor). It is equal to $r_{xy_j}^2$ for simple linear regression.

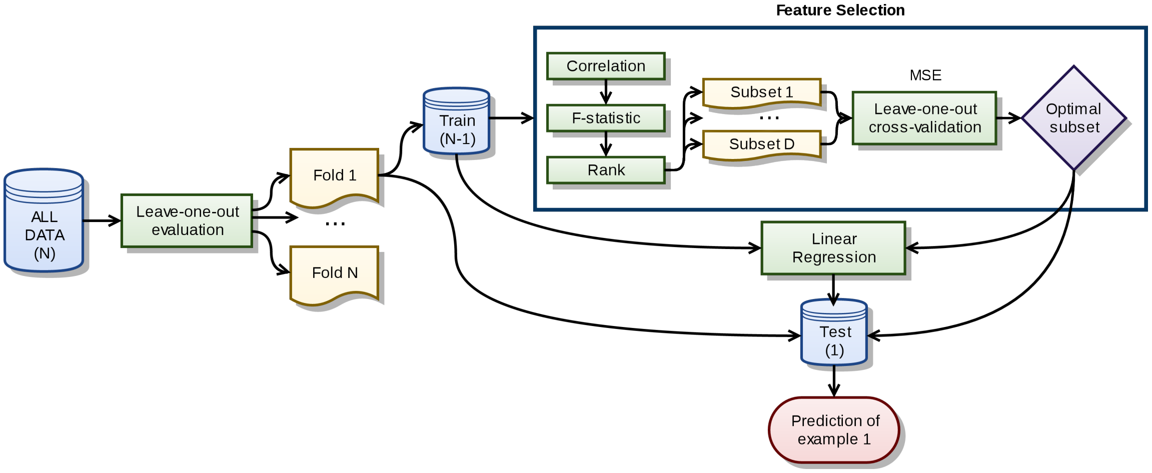


**Figure 2: Block diagram of the whole experimentation process, showing the details of one of the leave-one-out evaluation folds for better illustration. Within each evaluation fold, the train set is used to perform a feature selection procedure which eventually yields an optimal feature subset according to which the remaining test example is transformed and evaluated. Same operations are carried out in every fold so as to get predictions of every video and obtain overall performance measures.**

$R_r^2$: coefficient of determination of the restricted model, in this case a constant regressor (baseline prediction), therefore equal to zero.

$k$: number of estimated coefficients in the unrestricted model, 1 in this case.

$m$: number of restricted coefficients, 1 in this case.

Therefore:

$$F_j = \frac{r_{xy_j}^2}{(1 - r_{xy_j}^2)}(N - 1)$$

$F_j$ can be interpreted as a measure of the benefit that is gained by using feature $j$ for predicting $Y$ in comparison to simply predicting by the average (constant regressor). The p-values of the F-statistic are also computed here.

(3) Rank the features according to $F_j$ and create $D$ subsets of features $X_d$, each containing the $d = 1, 2, ..., D$ features with highest $F_j$.

(4) Perform leave-one-out cross-validation on the $D$ subsets in order to obtain the most suitable subset of features in terms of the mean squared error (MSE) of linear regression.

## 4.2 Evaluation

Finally, we test the generalization performance of the whole model through leave-one-out evaluation due to the inconvenience of creating a separate test set in view of the reduced number of examples. Therefore, on each fold of the leave-one-out we use all but one training examples to derive a reduced feature subset through the feature selection procedure described above and train the linear regression algorithm. Then, a prediction on the remaining test example is made (Figure 2). Note that each fold of the evaluation consists of different training examples and will potentially select different feature subsets. We will take advantage of this information and will provide a detailed analysis of the most valuable descriptors in Section 4.4.

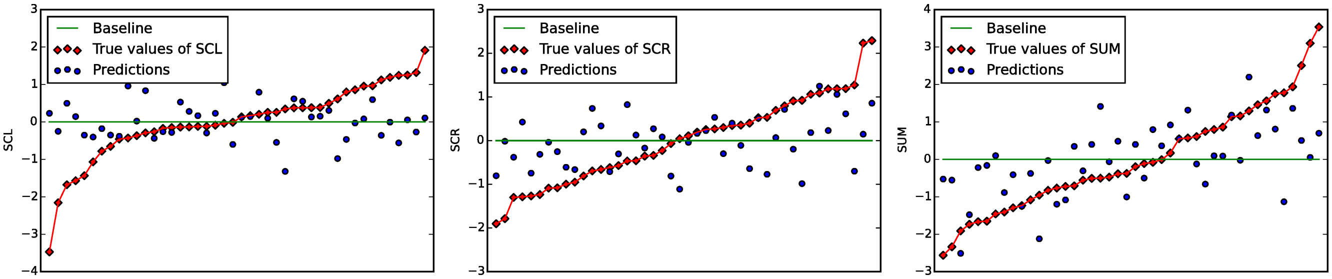
Additionally, one last experiment was performed by training the algorithms with the whole set of data, which would be the model employed to predict new examples. Then, it is possible to interpret this model in terms of both the selected descriptors and the weights assigned by the linear regression. Besides, the train results can be considered an optimistic reference for the performance with this method.

| | Attention (SCL) | | Emotion (SCR) | | Gl. impact (SUM) | |
|---|---|---|---|---|---|---|
| | MSE | $R^2$ | MSE | $R^2$ | MSE | $R^2$ |
| Test | 1.29 | 0.00 | 0.93 | 0.06 | 1.49 | 0.25 |
| Train | 0.88 | 0.09 | 0.68 | 0.30 | 1.10 | 0.45 |
| Ref. | 0.98 | 0.00 | 0.98 | 0.00 | 1.99 | 0.00 |

**Table 1: Mean squared error (MSE) and coefficient of determination ($R^2$) of the results from the leave-one-out evaluation (*Test* in the table, for short) and from the baseline dummy prediction (*Ref.*).**

## 4.3 Discussion

By analyzing Table 1 and Figure 3 we can draw several conclusions: First of all, the global impact (SUM) is the variable that is best predicted. The coefficient of determination for this metric is 0.25, which denotes a significant improvement over the baseline and the mean squared error is significantly lower than the baseline as well. Qualitatively, Figure 3c also shows that the predictions match reasonably well the ground truth annotations. In turn, the prediction of emotion (SCR) provides also an improvement over the baseline, but at a much lower scale than the case of SUM. Finally, no improvement has been achieved when predicting the attention (SCL), especially because the regressors have troubles to predict the extreme values, as shown in the figure.



**(a) Attention (SCL)**

**(b) Emotion (SCR)**

**(c) Impact (SUM)**

**Figure 3: Prediction (in blue) of the 3 global values for each of the videos after leave-one-out regression with the corresponding feature selection procedure. The actual values of the annotations are also shown in red and are sorted to facilitate the interpretation.**

Even though a coefficient of determination of 0.25 might seem far from its theoretical maximum of 1 (perfect prediction), it should not be considered a poor performance. It is important to recall that we are predicting attention and emotion, considered subjective elements of perception, by means of only low-level visual descriptors. It would be too ambitious aspiring to reach a close-to-perfect prediction, since that would mean that neither the aural characteristics (music, sounds, dialogues, etc.) nor the high-level semantics and the narrative structure of the videos have any influence on our perception. Thus, the relevance of these results lies in the fact that it has been indeed possible to explain some of the EDA variance. One proof of the challenging nature of the problem is that the train results do not achieve very high performances either, as it is shown in Table 1.

We can put forward several reasons why SCL could not be predicted and SCR results are modest. One is that predicting the tonic (SCL) and the phasic (SCR) electrodermal activity is challenging because they reflect very subtle and not very general reactions towards stimuli, which seems not to be the case when both metrics are brought together as the SUM metric. Another probable explanation is that these descriptors might indeed be useful for better predicting SCL and SCR, but linear regression is not powerful enough as it is highly probable that the correlation with these features is nonlinear. Besides, the feature selection algorithm is univariate, so it does not explore combinations of descriptors in depth, but it only takes into consideration individual linear correlations. Note that the use of simple methods is deliberate due to the early stage of the research in this field and for interpretability reasons.

## 4.4 Analysis of descriptors

| | Attention (SCL) | Emotion (SCR) | Impact (SUM) |
|---|---|---|---|
| Mean | 4.16 | 2.93 | 5.45 |
| Median | 1 | 2 | 5 |
| Std. Dev. | 4.35 | 2.00 | 0.81 |

**Table 2: Mean, median and standard deviation of the sizes of the feature subsets selected in the folds of the leave-one-out evaluation. Significant consistency differences can be observed between the three metrics.**

One very interesting analysis that can be carried out is looking at the descriptors and trying to infer which of them are most useful for the predictions. For this purpose we can take advantage of the output of the feature selection algorithm. Let us recall that the feature selection procedure is applied to 44 different data sets or partitions, one for each fold within the leave-one-out evaluation. Each time, the correlation-related metric $F_j$ is potentially different for each descriptor, as well as the rank and the final feature subset selected after cross-validation.

By analyzing the sizes of the feature subsets selected across the folds, presented in Table 2, one can make some observations which are in line with the performance results analyzed before. The evaluation of attention yielded very different feature subsets in view of the statistics, with a high standard deviation and dissimilar mean and median. This is another symptom of the poor performance with this metric. Conversely, the evaluation of emotion and global impact provided more regular feature subsets, the mean and the median are similar and especially in the case of global impact (SUM) the deviation is particularly low, which is an indicator of consistency in the most useful descriptors. This is translated into good prediction results, as presented before.

Another interesting branch of the descriptors analysis is looking at the 10 best features in terms of the average F-statistic across the folds (Tables 3, 4 and 5). Again, in the case of attention this analysis offers more evidence of the challenging problem of predicting it: F-statistics are not high, the p-values give low confidence and only one descriptor seems to have relatively high correlation. This suggests again that perhaps a multivariate feature selection technique that explores more complex combinations of features could be of help. The converse case appears in the case of emotion and global impact, where $F_j$ and their p-values suggest higher correlations and better confidence. For the emotion most folds select only the two best descriptors, whereas for the global impact five of them seem to be quite useful most times and indeed provide good predictions. Also in the case of emotion distinct combinations of descriptors could possibly yield bigger subsets and better results. Another observation in this regard is that many of the best descriptors (especially the two best ones) coincide in both emotion and global impact, but global impact, as the sum of attention and emotion interestingly

| Descriptors | Avg. $F_j$ | Avg. p-val | Sel. | Coef. |
|---|---|---|---|---|
| mean_cuts_min | 4.46 (0.65) | 0.043 | 100 % | 0.308 |
| num_cuts | 2.90 (0.44) | 0.100 | 41 % | - |
| violet | 1.47 (0.36) | 0.240 | 43 % | - |
| mean_hrot_ut | 1.50 (0.32) | 0.234 | 39 % | - |
| brown | 1.14 (0.31) | 0.300 | 41 % | - |
| std_intensity | 1.11 (0.33) | 0.307 | 41 % | - |
| green | 1.12 (0.31) | 0.306 | 18 % | - |
| darkblue | 1.05 (0.28) | 0.320 | 18 % | - |
| mean_hrot_lt | 0.97 (0.29) | 0.342 | 14 % | - |
| std_hrot_lt | 0.92 (0.31) | 0.356 | 16 % | - |

**Table 3: Information about the 10 visual descriptors (first column) with highest correlation with respect to the attention (SCL). The columns show: The average F-statistic (standard deviation within parentheses) across all the folds of the leave-one-out evaluation, the average p-value, the percentage of folds where the descriptor was included in the subset selected by the feature selection scheme and the coefficients estimated by the linear regression algorithm of the selected descriptors when trained with all the available data.**

| Descriptors | Avg. $F_j$ | Avg. p-val | Sel. | Coef. |
|---|---|---|---|---|
| std_intensity | 13.56 (1.25) | 0.001 | 100 % | 0.361 |
| std_hrot_lt | 10.32 (1.15) | 0.003 | 98 % | 0.262 |
| std_hrot_ut | 8.84 (0.98) | 0.005 | 23 % | - |
| mean_hrot_lt | 5.24 (0.66) | 0.029 | 16 % | - |
| std_colourfulness | 4.37 (0.66) | 0.046 | 11 % | - |
| std_saturation | 4.14 (0.64) | 0.052 | 20 % | - |
| red | 3.13 (0.53) | 0.088 | 5 % | - |
| main_colour1_pct | 2.83 (0.46) | 0.105 | 7 % | - |
| mean_hrot_ut | 2.68 (0.42) | 0.113 | 9 % | - |
| main_colour2_pct | 2.01 (0.52) | 0.172 | 5 % | - |

**Table 4: Information about the 10 visual descriptors with highest correlation with respect to the emotion (SCR). See the description of the columns in Table 3**

| Descriptors | Avg. $F_j$ | Avg. p-val | Sel. | Coef. |
|---|---|---|---|---|
| std_intensity | 11.14 (1.12) | 0.002 | 100 % | 1.305 |
| std_hrot_lt | 8.61 (1.05) | 0.006 | 100 % | 0.810 |
| num_cuts | 4.38 (0.52) | 0.045 | 100 % | 0.326 |
| std_hrot_ut | 3.49 (0.79) | 0.075 | 95 % | -0.678 |
| std_saturation | 2.87 (0.48) | 0.102 | 98 % | -0.866 |
| mean_cuts_min | 2.60 (0.41) | 0.118 | 25 % | - |
| red | 2.38 (0.45) | 0.136 | 14 % | - |
| brown | 1.95 (0.39) | 0.176 | 2 % | - |
| main_colour1_pct | 1.91 (0.45) | 0.184 | 5 % | - |
| green | 1.66 (0.34) | 0.211 | 0 % | - |

**Table 5: Information about the 10 visual descriptors (first column) with highest correlation with respect to the attention (SUM). See the description of the columns in Table 3**

incorporates some of the best descriptors for attention, like those related to the cuts (change of shot).

Finally, take advantage of having performed predictions through linear regression and analyze its estimated coefficients (reported in the last columns of Tables 3, 4 and 5) when training the algorithm with all the available data. Therefore, we will look again at the model we would use to predict new data. The first interesting observation is that the feature selection algorithm selected the same descriptors as in most of the folds, especially in the cases of emotion and global impact, which is in line with the consistency shown. For simplicity we will analyze only the global impact (SUM) coefficients, since it turned out to be the most successful one and it is the model with a bigger subset.

In Table 5 we can observe that five descriptors were selected, three of them got positive coefficients and two negative ones. For example, the descriptor that describes the variation of the intensity along the video (std_intensity) has a significantly high coefficient, which can be given the interpretation that wide variations in the brightness (intensity) along the video elicit higher emotional and attentional impact on the viewers. The same applies, but at a lower scale, with the number of cuts (num_cuts) and the variation in the utilization of the rule of thirds at the lower third (std_hrot_lt). Surprisingly, the converse behavior appears with the upper third (std_hrot_ut). Finally, according to the coefficient assigned to the variation of the saturation (std_saturation), it is suggested that low variation, i.e. constancy, in the saturation elicits more impact.

## 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 field to the stimuli side relying on EDA as ground truth for emotion and attention assessment. In this paper the stimuli are videos and we investigate if it is possible to predict EDA responses in a group of people by means of the visual 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), emotion (SCR) and global impact (SUM). Then, we extract a set of low-medium level visual descriptors from the videos and train linear regression to predict the EDA responses.

Through the experiments we have shown that there is some correlation between certain visual descriptors and EDA because it is possible to predict with reasonable confidence the SUM measure, considered an indicator of the global impact in terms of emotion and attention. In particular, we have achieved a coefficient of determination of 0.25 with linear regression. This demonstrates that visual characteristics of videos, such as the brightness, color, the changes of shot or the composition, among others, have an influence on the automatic and unconscious emotional and attentional reactions. An analysis of the most valuable descriptors according to the feature selection algorithm has been also provided.

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 [16]. However, finding some correlation with EDA has a great interest because it is a psychophysiological reaction controlled by the autonomous nervous system, thus it is automatic and is directly related to actual emotional and attentional activation, avoiding the implicit bias of opinions and judgments.

Correlation between the set of visual descriptors and SCL and SCR was not so clear as in the case of SUM. One explanation for this is that these measures alone reflect subtleties that are more difficult to capture by simple methods and a relatively small set of features. Therefore, one of the next challenges for the future research will be to define new descriptors and methods with potential to describe and predict with higher accuracy emotion (SCR) and attention (SCL) alone. Furthermore, it will be interesting to explore the correlations of aural descriptors with EDA (they already proved to be significantly useful for video aesthetics assessment [17]) and, of course, perform the experiments on larger and assorted data sets. It will be also very interesting to address the task of modeling the continuous evolution of EDA along videos, paying attention to the descriptors that elicit instantaneous emotional changes, for instance, represented by the phasic activity (SCR).

The chances are that EDA becomes 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. With this work we have shed some light on the kind of visual descriptors and methods that can be useful to predict EDA, as a somatic marker of the reaction of viewers to visual stimuli.

## ACKNOWLEDGMENTS

This research has been possible thanks to the collaboration of Sociograph Neuromarketing, particularly in providing the data set and annotations. 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 and the National Grants RTC-2016-5305-7 and TEC2014-53390-P of the Spanish Ministry of Economy and Competitiveness.

## REFERENCES

[1] Montserrat Aiger, María Palacín, and José-Manuel Cornejo. 2013. Electrodermal signal by Sociograph: methodology to measure the Group Activity. *Revista de Psicología Social* 28, 3 (2013), 333–347. https://doi.org/10.1174/021347413807719102

[2] Masato Asahina, Atsuya Suzuki, Masahiro Mori, Toshihide Kanesaka, and Takamichi Hattori. 2003. Emotional sweating response in a patient with bilateral amygdala damage. *International Journal of Psychophysiology* 47, 1 (2003), 87 – 93. https://doi.org/10.1016/S0167-8760(02)00123-X

[3] Martha Augoustinos, Iain Walker, and Ngaire Donaghue. 2014. *Social cognition: An integrated introduction.* Sage. 114–115 pages.

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

[5] David Bordwell, Kristin Thompson, and Jeremy Ashton. 1997. *Film art: An introduction.* Vol. 7. McGraw-Hill New York.

[6] Wolfram Boucsein. 2012. *Electrodermal activity.* Springer Science & Business Media.

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

[8] Christine Brinckmann. 2014. *Color and empathy: essays on two aspects of film.* Amsterdam University Press.

[9] Roberto Cabeza and Lars Nyberg. 2000. Imaging Cognition II: An Empirical Review of 275 PET and fMRI Studies. *J. Cognitive Neuroscience* 12, 1 (Jan. 2000), 1–47. https://doi.org/10.1162/08989290051137585

[10] Antonio R. Damasio. 1994. *Descartes' error: emotion, reason, and the human brain.* New York: Avon Books.

[11] Charles Darwin. 1872. The expression of the emotions in man and animals. *London, UK: John Marry* (1872).

[12] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. 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, 288–301. https://doi.org/10.1007/11744078_23

[13] Michael E. Dawson and Keith H. Nuechterlein. 1984. Psychophysiological Dysfunctions in the Developmental Course of Schizophrenic Disorders. *Schizophrenia Bulletin* 10, 2 (1984), 204–232. https://doi.org/10.1093/schbul/10.2.204

[14] M. E. Dawson, A. M. Schell, and D. L. Filion. 2000. The electrodermal system. In *Handbook of Psychophysiology, Second Edition*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson (Eds.). Cambridge University Press, Cambridge, 200–223.

[15] Ch Fere. 1888. Note sur les modifications de la résistance électrique sous l'influence des excitations sensorielles et des émotions. *CR Soc. Biol* 5 (1888), 217–219.

[16] F. Fernández-Martínez, A. Hernández-García, and F. Díaz-de-María. 2015. 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.

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

[18] J. Fleureau, P. Guillotel, and I. Orlac. 2013. 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.* 73–78. https://doi.org/10.1109/ACII.2013.19

[19] Don C. Fowles, Margaret J. Christie, Robert Edelberg, William W. Grings, David T. Lykken, and Peter H. Venables. 1981. Publication Recommendations for Electrodermal Measurements. *Psychophysiology* 18, 3 (1981), 232–239. https://doi.org/10.1111/j.1469-8986.1981.tb03024.x

[20] Maria Gendron and Lisa Feldman Barrett. 2009. Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review* 1, 4 (2009), 316–339.

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

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

[23] William James. 1884. What is an emotion? *Mind* 34 (1884), 188–205.

[24] William James. 1890. *The principles of psychology.* New York: Dover.

[25] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Trans. Affect. Comput.* 3, 1 (Jan. 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[26] Mathieu Lajante, Olivier Droulers, Thibaut Dondaine, and David Amarantini. 2012. Opening the "black box" of electrodermal activity in consumer neuroscience research. *Journal of Neuroscience, Psychology, and Economics* 5, 4 (2012), 238.

[27] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (1993), 261–273.

[28] Joseph LeDoux. 2012. Rethinking the emotional brain. *Neuron* 73, 4 (2012), 653–676.

[29] J. L. Martínez and E. Garrido. 2003. Sistema para la medición de reacciones emocionales en grupos sociales. 2 168 928. Patente de invención A6113 5116, 2003-10-1. (2003).

[30] J. L. Martínez, Sergio Monge, and M. Isabel Valdunquillo. 2012. Medición de las respuestas psicofisiológicas grupales para apoyar el análisis de discursos políticos. *Trípodos* 29 (2012), 53–72.

[31] Anush K. Moorthy, Pere Obrador, and Nuria Oliver. 2010. 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, 1–14. http://dl.acm.org/citation.cfm?id=1888150.1888152

[32] Arne Öhman. 1999. Distinguishing Unconscious from Conscious Emotional Processes: Methodological Considerations and Theoretical Implications. In *Handbook of Cognition and Emotion.* John Wiley & Sons, Ltd, 321–352. https://doi.org/10.1002/0470013494.ch17

[33] International Commission on Illumination. 2004. *Colorimetry: technical report.* Commission internationale de l'Eclairage, CIE Central Bureau. http://books.google.es/books?id=P1NkAAAACAAJ

[34] Michael Ondaatje and Walter Murch. 2002. *The conversations: Walter Murch and the art of editing film.* A&C Black.

[35] Luiz Pessoa. 2008. On the relationship between emotion and cognition. *Nature Reviews Neuroscience* 9, 2 (2008), 148–158.

[36] Michael I. Posner, M. Rosario Rueda, and Philipp Kanske. 2007. Probing the Mechanisms of Attention. In *Handbook of Psychophysiology* (third ed.), John T. Cacioppo, Louis G. Tassinary, and Gary Berntson (Eds.). Cambridge University Press, 410–432. http://dx.doi.org/10.1017/CBO9780511546396.018 Cambridge Books Online.

[37] William F. Prokasy and David C. Raskin (Eds.). 1973. *Electrodermal Activity in Psychological Research.* Academic Press. https://doi.org/10.1016/B978-0-12-565950-5.50001-6

[38] M. J. Rosenberg and C. I. Hovland. 1960. Cognitive, affective, and behavioral components of attitude. In *Attitude organization and change*, M.J. Rosenberg, C. I. Hovland, McGuire W., Abelson R., and J. Brehm (Eds.). New Haven, CT: Yale University Press.

[39] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A Metric for Distributions with Applications to Image Databases. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV '98).* IEEE Computer Society, Washington, DC, USA, 59–. http://dl.acm.org/citation.cfm?id=938978.939133

[40] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[41] Henrique Sequeira, Pascal Hot, Laetitia Silvert, and Sylvain Delplanque. 2009. Electrical autonomic correlates of emotion. *International journal of psychophysiology* 71, 1 (2009), 50–56.

[42] Alvy Ray Smith. 1978. Color Gamut Transform Pairs. *SIGGRAPH Comput. Graph.* 12, 3 (Aug. 1978), 12–19. https://doi.org/10.1145/965139.807361

[43] Mohammad Soleymani, Guillaume Chanel, Joep J. M. Kierkels, and Thierry Pun. 2008. Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses.. In *ISM.* IEEE Computer Society, 228–235. http://dblp.uni-trier.de/db/conf/ism/ism2008.html#SoleymaniCKP08

[44] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *Affective Computing, IEEE Transactions on* 3, 1 (2012), 42–55.

[45] Kenneth L. Subotnik, Anne M. Schell, Mark S. Chilingar, Michael E. Dawson, Joseph Ventura, Kimberle A. Kelly, Gerhard S. Hellemann, and Keith H. Nuechterlein. 2012. The interaction of electrodermal activity and expressed emotion in predicting symptoms in recent-onset schizophrenia. *Psychophysiology* 49, 8 (2012), 1035–1038. https://doi.org/10.1111/j.1469-8986.2012.01383.x

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

[47] Daniel Tranel. 2000. Electrodermal activity in cognitive neuroscience: Neuroanatomical and neuropsychological correlates. In *Cognitive neuroscience of emotion. Series in affective science.* Oxford University Press, 192–224.

[48] Yong Jian Wang and Michael S Minor. 2008. Validity, reliability, and applicability of psychophysiological techniques in marketing research. *Psychology & Marketing* 25, 2 (2008), 197–232.

[49] Boon-Lock Yeo and Bede Liu. 1995. Rapid Scene Analysis on Compressed Video. *IEEE Trans. Cir. and Sys. for Video Technol.* 5, 6 (Dec. 1995), 533–544. https://doi.org/10.1109/76.475896

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

# A DETAILS ON DESCRIPTORS

This appendix aims at giving more details about the low-level visual descriptors extracted from the videos for this research. Similarly as in Section 3 we will present these details by organizing the descriptors into 9 families.

## A.1 Intensity

Intensity can be also referred to as brightness. In a picture or frame it is the average value of the pixels of the gray-scale version of the image. Then, such image-level characteristic is extended to the video level by computing the following statistics:

- *mean-intensity*: average intensity along all the frames of the video.
- *std-intensity*: standard deviation of the intensity.

## A.2 Hue

Hue is one of the channels of the well-known color space HSV [42]. Roughly speaking, hue allows identifying colors by an angle from 0 to 360 degrees. As in the case of the intensity, we compute the following statistics:

- *mean-hue*: average of the pixel values of the hue channel of every frame in a video.
- *std-hue*: standard deviation of the hue channel.

## A.3 Saturation

Saturation can be thought as a parameter that measures the purity of the color, i.e. how close to gray a color is. It is expressed as a percentage, being 100 % fully saturation and 0 % a gray tone and corresponds to the S channel of HSV. Again, we obtain the average saturation and the standard deviation along the whole video:

- *mean-saturation*: average of the pixel values of the saturation channel of every frame in a video.
- *std-saturation*: standard deviation of the saturation channel.

## A.4 Entropy

Since entropy is a statistical measure that refers to the randomness of a variable, applied to images it can describe texture. Four features related to entropy are computed:

- *mean-entropy*: average entropy along all the frames of the video.
- *std-entropy*: standard deviation of the entropy.
- *pct-low-entropy-frames*: percentage of low entropy frames. A frame can be regarded as a low entropy one when its entropy value is below a particular threshold. This feature is designed to capture those commercials that insert some extra frames with monochromatic background at the end of the video to show, for instance, the brand logo or any other information.
- *low-entropy-end*: a binary feature that states if the end of the video (i.e. last 10% of frames) is mainly formed by low entropy frames, as previously described. For this feature to be set as 1 at least 85% of ending frames must have low entropy.

## A.5 Temporal segmentation (cuts)

Temporal segmentation is in film-making and publicity the basis of montage, the editing technique that allows the creation of most effects cinema produces. In order to extract features related to this aspect, it is necessary to determine the abrupt transitions between subsequent shots. We have followed the procedure described in [49], which uses the sum of absolute differences (SAD) of the gray intensity, $D$. The final detection consists in setting a threshold of 0.18 on a discrete version of the second derivative of $D$:

$$M(n) = -D''(n+1) = -(D'(n+1) - D'(n)) \tag{1}$$

Then, with this information we define these features:

- *num-cuts*: total number of cuts within a video.
- *longest-shot*: duration in seconds of the longest shot (i.e. a fragment of video between two consecutive cuts).
- *mean-shot-duration*: mean duration of the shots of the video, in seconds.
- *std-shot-duration*: standard deviation of the duration of the shots.
- *mean-cuts-min*: mean density of cuts.

## A.6 Frame-level colorfulness

With this visual characteristic, rather than measuring the intensity or vividness of colors, which is described by previous features, we aim to measure the degree of variation of colors. A picture is said to be colorful when it presents richly varied colors, in contrast to monochromatic or poorly colored images. For this family of features, we compute the colorfulness of every frame and extend it to the temporal dimension by averaging and computing the standard deviation as usual. In order to compute the colorfulness of a frame we calculate the 64-bin color histogram (after conversion to the CIE Lab color space [33]) of each frame and compare it with the histogram of an ideal colorful picture, i.e. uniformly distributed, through the Earth Mover's Distance [39].

- *mean-colorfulness*: mean colorfulness along all the frames of a video.
- *std-colorfulness*: standard deviation of the distribution of the colorfulness along all the frames.

## A.7 Video-level colorfulness

It is an adaptation of frame-colorfulness where instead of computing the colorfulness of each frame, a value of colorfulness is computed for all the of pixels of the video as a whole. That is, we compute one single color histogram and compare it to the ideal color histogram as previously explained. Now it is possible to determine, for instance, the peaks of the histogram, which are indicative of the most predominant colors. The particular features derived from this method are the following:

- *video-colorfulness*: colorfulness computed taking into consideration all the pixels of the video at once.
- *first-color*: index (from 1 to 64) of the color with the highest frequency in the histogram.
- *first-color-freq*: relative frequency in the histogram of the first color.
- *second-color*: index of the color with the second highest frequency in the histogram.
- *second-color-freq*: relative frequency in the histogram of the second color.

## A.8 Color profiles

These features aim at characterizing the similarity of the overall colors present in the video to eight predefined colors, which give in turn the name to the descriptors: *red, green, dark blue, light blue, cyan, violet, brown and grey*. The method for obtaining the value of the features is very similar to the one applied for getting *video-colorfulness*: we compute the color histogram of all the pixels within the video and compare it to the color histograms of the corresponding predefined colors.

## A.9 Rule of Thirds (ROT)

The rule of thirds (ROT) is a very important rule of thumb in visual arts which states that the most important subjects in the image should be placed at the horizontal and vertical imaginary lines that divide the image in thirds. Thirds are used because they approximate the golden ratio, widely present in nature and used already by ancient Greeks in architecture, sculpture and other arts because it gives harmony to the compositions. Here we use descriptors that measure the degree of utilization of the rule of thirds for placing important horizontal lines. This measure consists in comparing, by a sum of absolute differences, the 64-bins color histograms corresponding to the two sub-images that the horizontal line generates. We extract the following features:

- *mean-hrot-lt*: mean value of the previously described feature along all the frames of a video, applied to the comparison between the sub-images below and above the lower third line.
- *std-hrot-lt*: standard deviation of the distribution of the degree of utilization of ROT along all the frames of a video applied to the comparison between the sub-images below and above the lower third line.
- *mean-hrot-ut*: same as mean-hrot-lt but referred to the upper third line.
- *std-hrot-ut*: same as std-hrot-lt but referred to the upper third line.