Multimodal Engagement Classification for Affective Cinema

Mojtaba Khomami Abadi[1,3], Jacopo Staiano[1], Alessandro Cappelletti[2], Massimo Zancanaro[2], Nicu Sebe[1]
[1]Department of Information Engineering and Computer Science, University of Trento, Italy
[2]Fondazione Bruno Kessler (FBK), Trento, Italy
[3]Semantic, Knowledge and Innovation Lab (SKIL), Telecom Italia

***Abstract*—This paper describes a multimodal approach to detect viewers' engagement through psycho-physiological affective signals. We investigate the individual contributions of the different modalities, and report experimental results obtained using several fusion strategies, in both per-clip and per-subject cross-validation settings. A sequence of clips from a short movie was showed to 15 participants, from whom we collected per-clip engagement self-assessments. Cues of the users' affective states were collected by means of *(i)* galvanic skin response (GSR), *(ii)* automatic facial tracking, and *(iii)* electroencephalogram (EEG) signals. The main findings of this study can be summarized as follows: *(i)* each individual modality significantly encodes the level of engagement of the viewers in response to movie clips, *(ii)* the GSR and EEG signals provide comparable contributions, and *(iii)* the best performance is obtained when the three modalities are used together.**

I. Introduction

From the gaming domain [1], [2] to visual arts [13] the design and validation of systems for automatic *engagement* assessment has drawn significant attention from the Affective Computing community. Motivated by the fall in production costs on one hand, and by the advent of Smart-TVs and web-series streamed on the internet, this paper adds a further scenario to such research efforts: *affective cinema*. We envision a novel movie-telling paradigm, making use of both hyper-narrative screenwriting (that is, multiple coherent narrative paths for a given movie) and viewers sensing through selected affective channels.

In a gaming scenario, as suggested in [2], the player's engagement (i.e., positive excitation) is crucial to a game's success: it should be maintained while two main factors (the player's skills and the game difficulty) vary over time, in order to prevent the player falling in a state of *anxiety* (i.e., negative excitement) or *boredom* (i.e., negative calm). Conversely, in the scenario we propose, the viewer's engagement represents a variable that screen-writers must take into account during the scripting process: by defining segments of interest within the stimulus (i.e., the script and thus the movie) during which the viewers engagement response is automatically assessed, they will be able to use this response to drive the movie narrative and thus provide a novel and adaptive experience to the audience.

In this study we investigate the effectiveness of different psycho-physiological channels, and combinations thereof, for classifying a viewer's level of engagement while watching a movie. The modalities we take into account are galvanic skin response (GSR), electroencephalogram (EEG), and facial motion tracking. Evaluation is carried out under several experimental settings on a 15-subjects dataset.

This paper aims to:

1) bridge the gap between affective computing and hyper-narrative movies;
2) introduce a multimodal system that combines face analysis with EEG and GSR;
3) present a systematic analysis of the importance of the channels of information and their best combination for detecting a viewer's level of engagement.

The main findings can be summarized as follows: *(i)* if taken independently, the different channels employed (GSR, EEG, Facial Motion) are found to significantly encode the viewers' level of engagement; *(ii)* contributions from GSR and EEG are found to be comparable, and complemented by Facial Motion; *(iii)* the best performance is obtained when considering all channels under a late fusion strategy.

This paper is structured as follows: section II summarizes the previous research efforts in the contexts of both engagement assessment through psycho-physiological channels and of structuring hyper-narrative movies; section III provides an overview of the experimental protocol we followed; section IV describes the data pre-processing and feature extraction steps taken; finally, after reporting experimental results in section V, we discuss them along with the future research directions in section VI.

II. Related works

Our case study is a system in which the viewer is in fact considered more as a sensor than as an interactor, driving the narrative flow implicitly through the psycho-physiological signals sensed. In this section we thus briefly review significant research efforts in the domains of Affective Computing and Hyper-Narrative Movies. While these domains are quite clearly separated in terms of practices and goals, we choose to hereby bridge them, since they both are crucial to our case study.

The affective states induced in viewers by multimedia contents depend on several psychological and contextual factors, and are thus highly subjective. Nonetheless, the creators of such contents have a strong interest to convey certain feelings at given moments in the experience. Investigating the relations between the observed affective response on one hand, and the ones the creators intended to elicit on the other, is therefore a very interesting line of research.

Several efforts from the Affective Computing community have focused on this problem [7], [22], [25]. Commonly, such works have shared the assumption that the range of human emotions is defined, as demonstrated in [5], on a plane whose two dimensions are represented by *arousal* (i.e., excitement *vs* boredom), and *valence* (i.e., positive *vs* negative) [6]. Moreover, they often shared the selection of presentation stimuli or relied on ratings of contents on the arousal/valence plane from large pools of subjects, in order to ensure consistency of the observed/expected emotions.

In [20] several psycho-physiological signals, such as heartbeat rate, skin temperature and conductivity, facial electromyography (EMG), have been found to correlate with affective states in response to visual content. In particular, increases of heartbeat and respiration rates were found to be linked to states of excitement, and physiological responses to visual contents eliciting anger or fear were significantly different from the response to neutral contents. Soleymani *et al.* [21] contributed the MAHNOB-HCI multimodal database to the Affective Computing community: it consists of face videos, audio, physiological signals, and eye-gaze patterns, recorded on 27 participants who were presented with 20 emotional clips in one task, and 14 short videos and 28 images in the other. Furthermore, Koelstra *et al.* [10] assembled the DEAP dataset, in which they presented viewers with 40 one-minute music video-clips and collected their ratings on arousal, valence, liking, and dominance. During the experiment, they recorded blood volume pressure, respiration rate, EEG, GSR, skin temperature, and electrooculogram (EOG) patterns, to be used in a binary classification task; the performance obtained was above 60% of mean accuracy.

Lisetti and Nasoz [15] presented 29 participants with video-clips evoking fear, anger, amusement, surprise and sadness, and induced frustration by asking them to solve difficult mathematical questions on the fly. They recorded physiological responses through an armband sensing heartbeat rate, GSR, temperature, EMG and heat flow, and obtained over 80% accuracy in classification experiments. Staiano *et al.* [23] exploited facial motion features to detect interaction difficulties in a User Experience (UX) evaluation scenario.

Moving to the Media Studies domain, the pioneering work *Kinoautomat*[1] by Raduz Činčera is considered the first example of interactive movie: exhibited in 1967, the viewers' experience would be interrupted at five binary decision points in which the audience could choose the narrative path to follow. Other media, from video-games, to adventure-games, to "choose-your-own" paperbacks, have seldom adopted such strategies and users have gradually become accustomed to it.

From a theoretical point of view, it is argued [19] that hyper-narrative works should be "*consonant with, rather than alien to human cognitive, affective and sensual faculties*" in order to be deeply engaging.

Recently, the use of computational models of surprise and suspense for narrative generation [16], [17] has been proposed. In this study, we take a further step in bridging the research experiences gained within these two very different communities, as quickly summarized above: findings coming from Affective Computing researches can open valuable novel scenarios in the Movie Studies field, and nonetheless in the Movie industry (e.g. along with the rise of Smart-TVs); conversely, Affective Computing researchers will benefit from a wealth of high level affective content directly proportional to the level of interest such scenarios will spawn in the Movie Studies community.

[1] http://www.kinoautomat.cz

III. Experimental Protocol

In this section, we *(a)* detail the experimental set-up and the protocol followed, *(b)* report the analysis of the self-reported engagement assessments, and *(c)* describe the modalities and correspondent features used in this study.

A. Experimental set-up and protocol

1) Material and setup: The study was conducted in a lab environment with the aim of investigating to what extent the psycho-physiological features may be used to infer the viewers' level of engagement. The subjects were tested individually, sitting on a comfortable sofa, facing the monitor (see Figure 1).

For designing the system we tried to minimize invasiveness and hardware costs. The system is composed of a consumer-level personal computer, equipped with a standard webcam, and commercially available EEG and GSR sensors. Stereo speakers were also placed in the acquisition room for rendering the audio.

2) Protocol and stimuli: For the purpose of this study, 15 participants were shown a sequence of 11 movie clips of different lengths – 7 short ones (μ=26s, σ=14s) and 4 long ones (μ=118s, σ=69s), composing one of the possible narrative paths of a short movie. At the end of each clip, the experience was interrupted and users were prompted with a simple slider interface in order to self-report their level of engagement in the continuous range of -10 to 10.

Fig. 1. A viewer, wearing the headphone-like EEG sensor on his head and the watch-like GSR sensor on his left hand while watching the movie.

B. Analysis of self-assessment ratings

In this section, we analyze the ratings provided by the participants for the movie clips. Participants' ratings are

- considered a conscious reflection of their level of engagement while viewing the stimuli, and should therefore be correlated with their physiological responses;
- ultimately used for classification purposes, and hence the distribution of the ratings and the correlation of the

ratings with other artifacts (e.g. clips' length) should be carefully analyzed;

- indicative of whether the presented stimuli can effectively elicit engagement in viewers.

1) Subjects' ratings and the clips' length: To have a better sense of the distribution of participants' ratings, we report in *Table I* some statistical measures over the median[2] and standard deviation of each subject's ratings. While Table I suggests that the participants rated longer clips with higher level of engagement, a correlation analysis shows that there is no correlation between the ratings of individual subjects and the length of the clips (mean of p-values=0.58, std of p-values=0.30).

TABLE I. MEAN, STANDARD DEVIATION (STD), MINIMUM, AND MAXIMUM VALUES OVER MEDIAN/STANDARD DEVIATION OF INDIVIDUAL SUBJECTS' RATINGS.

The measure over each subject's ratings	Mean	STD	Min	Max
Median - all clips	+1.4	2.6	-3.6	+4.6
STD - all clips	2.8	1.6	0.8	6.3
Median - short clips	+0.5	3.4	-6.0	+4.6
STD - short clips	2.7	1.4	1.0	5.5
Median - long clips	+2.0	3.7	-3.9	+9.8
STD - long clips	2.8	1.9	0.3	7.8

2) Pre-processing the subjects' ratings for classification purposes: We aim at training a system that is able to recognize low/high engagement level of viewers, and thus design a binary classification task for the purpose. Being this study inherently viewer-centered, we dichotomize ratings on a subject basis: each clip watched and rated by subject x is given a *high/low* engagement label when it is rated above/below the mean of all ratings provided by x. Thus, the obtained labels reasonably reflect whether the viewer was *more/less* engaged in watching the clip in comparison to the other clips.

3) Agreement between raters: We employed Fleiss kappa [4] to analyze the agreement between all viewers over the ratings. In this case, we have 15 raters (participants), 11 items (clips) and 2 labels (*more/less engaging*). The outcome of Fleiss kappa is 0.23 (p-value=0), which according to the rules of thumb provided by Ladis and Koch [11] represents a fair agreement between raters. This suggests that the clip sequence used as stimuli was perceived in a similar way by the participants.

C. Modalities and features

We collected affective cues of the viewers in response to the stimuli through three different modalities: *(i)* facial motion tracking, *(ii)* a commercial headphone-like one-channel EEG sensor, and *(iii)* galvanic skin response.

To extract facial motion features, we exploited an effective, realtime facial expression recognition system provided by Joho *et al.* [8]. The system tracks 12 facial interest points and provides features representing their motion at each video frame.

Koelstra et al. [10] reported the mean correlations (over a population of 32 participants) of emotional dimensions (arousal, valence and, and general ratings) with the power in the broad frequency bands of theta (4-7 Hz), alpha (8-13 Hz), beta (14-29 Hz) and gamma (30-47 Hz) in ElectroEncephaloGraphy (EEG). According to [10], frontal EEG channels are significantly correlated with emotions evoked by dynamic stimuli (Music Videos). In this study we used a bluetooth single-dry EEG sensor. The sensor is commercially available[3] and captures the frontal EEG activities. The EEG processing module provides 8 features corresponding to the values of low/high alpha, low/high beta, mid/low gamma, delta, and theta EEG bands, along with 2 aggregate features representing the level of attention and meditation of the users. We recorded the module output features while the users were watching the clips. These features were sampled at each second.

Galvanic skin response (GSR) measures the electrical resistance of the skin between two electrodes that are positioned on the medial phalanges of the middle and the index fingers. The GSR sensor passes a negligible current through the body and the measured electrical resistance changes due to variations in perspiration rate. Therefore, GSR signal involves information of sweat glands that are controlled by the sympathetic nervous system, and hence the signal contains information about the emotion of the users. Particularly, GSR values decrease due to an increase of perspiration, which usually occurs when the user is experiencing stress or surprise. Lang et al. [12] showed that the mean value of the GSR signal is correlated with the level of arousal. We recorded the GSR signal at a 100Hz sample rate using a commercially available[4] bluetooth GSR sensor.

IV. DATA ANALYSIS

This section describes *(a)* the steps taken to pre-process the data and extract the features used in our experiments *(b)* the procedure followed to assign each clip with a label representing engagement, and the classification strategies adopted.

A. Data pre-processing and feature extraction

1) Facial Motion features: The facial tracking module provides 12 Motion-Units (MUs) [18] features for each frame of the facial video captured by the webcam. Over each clip and for each user we report some statistical measures in Table II, describing the overall temporal distribution of the MUs.

2) EEG features: The EEG signal analyzer module provides 10 features/s, corresponding to the values of low/high alpha, low/high beta, mid/low gamma, delta, theta EEG bands, and the level of attention and meditation[5]. Similar to what we did for MU features for face, over each clip and for each user we calculated statistics (see Table II) that describe best the overall temporal distribution of the individual features.

3) GSR features: The pre-processing of the GSR signal and the feature extraction procedure in this study follows the state of the art techniques adopted by Kim and Andre [9], Soleymani et al. [21], and Koelstra et al. [10]. We removed the trend of the GSR signal by subtracting the temporal low frequency drift computed by smoothing the signal with a 100 points moving average. The features extracted from the GSR signal are presented in table II.

[2]Due to the small number of clips, the median is used to estimate more reliably the actual mean value.

[3]http://www.neurosky.com

[4]http://www.shimmer-research.com

[5]The algorithms used for computing attention and meditation levels are closed-source and patented by the hardware manufacturer.

TABLE II. FEATURES EXTRACTED FROM EACH MODALITY.

Modality	Extracted Features
Facial Motion (n=72)	mean, skewness, variance, std of each MU over time, as well as the percentage of times each MU had a value above/below its mean $\pm$ std.
Frontal EEG (n=60)	mean, skewness, variance, std of each raw feature over time, as well as the percentage of times each entry had a value above/below its mean $\pm$ std.
GSR (n=31)	average skin resistance, average of derivative, average of derivative for negative values only (average decrease rate during decay time), proportion of negative samples in the derivative vs. all samples, number of local minima in the GSR signal, average rising time of the GSR signal, 10 spectral power in the [0-2.4]Hz bands, zero crossing rate of Skin conductance slow response (SCSR) [0-0.2]Hz, zero crossing rate of Skin conductance very slow response (SCVSR) [0-0.08]Hz, SCSR and SCVSR mean of peaks magnitude

B. Classification strategies

Several approaches have been adopted in previous research efforts to model human affective responses; many rely on multimodal setups (i.e., collecting affective responses through several channels) [2], [9], [10], [15], [21], [22]. Advantages of choosing a multimodal strategy include *(i)* the information coming from the channels can be fused at different stages of the experimental pipeline, providing researchers with insights on the interaction between different modalities, *(ii)* in controlled scenarios it is possible to compare the information content provided by each channel, and *(iii)* the model built can benefit from some degree of fault-tolerance: if a channel sensory infrastructure fails for some reason, the overall system might still be able to work thanks to information derived from other channels.

Approaches to fuse the information content of different modalities can be divided into two broad categories, namely *(i)* feature fusion (or early integration), and *(ii)* decision fusion (or late integration) [14]:

- in feature fusion, the features extracted from different modalities are concatenated into a single feature vector which serves as input to the recognition system; hence, the synchronous characteristics of the involved modalities are fed into the system;
- in decision fusion, the feature vectors coming from each channel serve as input for independent classifiers, whose outputs are then combined to obtain the final assessment. A decision fusion-based system can be thus thought as a mixture of expert classifiers (uni-modal classifiers). The decision fusion technique allows to model asynchronous characteristics of the modalities.

An important advantage of decision fusion over feature fusion is that, since the information content of each modality is processed independently, it is relatively easy to employ an optimal weighting method to adjust the relative contribution of each modality to the final decision, in accordance with the reliability of each channel [10].

In our study, we built a multimodal recognition system using two different experimental settings: *(a)* a clip-based setup, for which we generated 10 random folds from the complete set of samples (i.e., video-clips); and *(b)* a subject-based setup, for which we adopted a leave-one-subject out cross-validation approach. For the former, the generation of the random folds was repeated 100 times and the results were averaged in order to properly assess the performance of the system.

In both settings, in order to better exploit the information content of the modalities in different levels (feature/decision fusion), we trained and evaluated 4 classifiers (all linear kernel SVMs) independently as follows: *(i)* the EMO classifier that gets the 72 features related to facial motion; *(ii)* the EEG classifier using the 60 features obtained by the EEG sensor; *(iii)* the GSR classifier using the 31 GSR features; and *(iv)* the early-integration classifier using feature-level fusion of all available features as input. At feature-fusion level, in order to reduce the dimensionality of the feature vector, Principal Component Analysis (PCA) projections of the EEG and EMO feature vectors were concatenated to the GSR feature vector; from the experimental point of view, at each train/test fold the PCA projection matrices were calculated over the training samples and used at testing stage.

Furthermore, we report the performance of the system when using two different decision-fusion techniques: *(i)* a simple majority vote over the three classifiers' decisions (EMO/EEG/GSR), and *(ii)* a late-fusion strategy for finding the optimum weights to be given to the three classifiers for obtaining the best performance. We observe that *(i)* the best performance is obtained when using the majority-voting approach, and *(ii)* the contributions of EEG and GSR based classifiers are comparable. The next section elaborates and describes the results in more depth.

V. EXPERIMENTAL RESULTS

In section III, we described how we derived clip labels from the viewers' self-assessments. Upon assigning the *low/high engagement* labels to the affective responses, we provide an analysis of their distribution. The proportion of the number of samples belonging to the *high engagement* class is calculated for all the subjects; the mean and STD of this value over all the users are respectively 0.57 and 0.07, which gives a sense of how much the classes are imbalanced. Thus, in order to reliably assess performance, we report the F1-score (average of F1-scores for each class) along with accuracy.

We performed a weighted sum decision fusion over the output of the uni-modal classifiers. Assuming f_{GSR}, f_{EEG}, and f_{EMO} represent the output of the uni-modal classifiers, the decision output is calculated as:

$$f_{out} = w_{GSR} \times f_{GSR} + w_{EEG} \times f_{EEG} + w_{EMO} \times f_{EMO} \quad (1)$$

where w_{GSR}, w_{EEG}, and w_{EMO} represent the weights given to the contribution of each uni-modal classifier and are normalized by:

$$w_{GSR} + w_{EEG} + w_{EMO} = 1 \quad (2)$$

The optimal values for the weights are estimated by a exhaustive search in the regular grid space, where each weight is incremented from 0 to 1 by 0.01 steps: at each train/test fold, the values producing the best recognition performance over the training data are selected and used in the classifying the test samples (see table III).

Table III shows measured classification accuracy and F1-scores of engagement classification for both the clip-based

TABLE III. AVERAGE ACCURACIES (ACC) AND F1-SCORES OBTAINED. * INDICATE WHETHER THE F1-SCORE DISTRIBUTION OVER SUBJECTS IS SIGNIFICANTLY HIGHER THAN THE BASELINE ($^* = p < 0.05, ^{**} = p < 0.01, ^{***} = p < 0.001$).

	Clips-based		Subjects-based	
Feature Type/ Fusion Technique	**ACC**	**F1**	**ACC**	**F1**
Facial Motion	0.57	0.59***	0.62	0.60*
Frontal EEG	0.57	0.58***	0.65	0.62**
GSR	0.57	0.63***	0.71	0.68**
Feature-Fusion	0.55	0.62***	0.71	0.69***
Decision-Fusion (weighted sum)	0.68	0.66***	**0.75**	**0.73*** **
Majority Vote - equally weighted sum	0.69	0.67***	**0.76**	**0.74*** **
Best possible performance	0.70	0.68	0.76	0.75

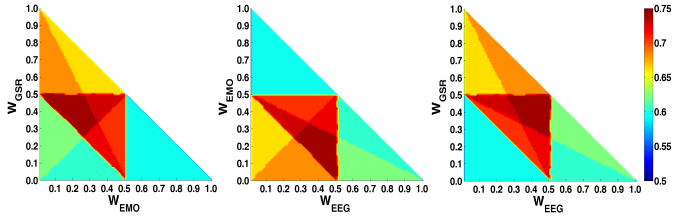


Fig. 2. The outcome of weighted decision fusion for the subject-based setup. The plots show the best possible performance for the different combinations of weights, and suggests that the EEG and GSR channels provide similar contributions. Image best viewed in color.

and the subject-based cross-validation setups. To test their significance, the F1-distribution over all the runs/participants is compared to the 0.5 baseline using an independent one-sample t-test. As shown in the table, uni-modal performance for the GSR modality is better than the other uni-modal approaches. Moreover, fusing techniques are found to significantly improve the recognition rates: in particular, majority-voting decision fusion works better than the other employed fusion techniques.

Classification performance for the subject-based cross-validation setup is found to be significantly better in comparison with the clip-based setup: this suggests that affective responses have higher inter-subject variability, while showing consistent intra-subject patterns. The best possible performances reported in Table III are calculated by assigning fixed weights to the uni-modal classifiers according to Eq. 1, and choosing the ones providing best performance.

For the sake of comparison, Figure 2 displays the mean performance (F1-score) over all participants/folds when different weights are used for the weighted sum decision-fusion approach.

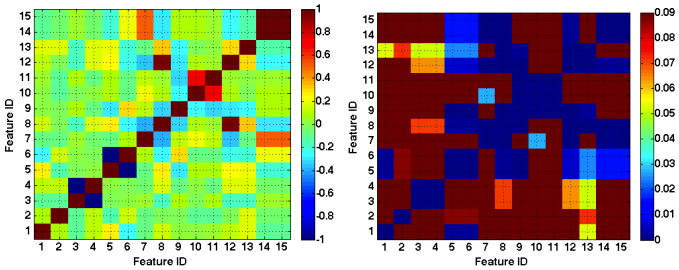
Furthermore, table IV reports the correlations between the users' self assessment ratings and the extracted features (only for $p \leq 0.05$). Likewise the standard features extracted from the GSR signal, interestingly, the percentage of time the level of meditation was above/below the relative $\mu \pm \sigma$ value turned out to be correlated with the users' level of engagement. Moreover, the movement of lip corner is significantly correlated with the level of engagement which is consistent with the findings reported in [23], [24]. Figure 3 displays results of a correlation analysis between the top-15 features listed in Table IV), showing the degree of redundancy between them.

Based on the presented results, we summarize our main observations as follows:

1) as shown in Table IV the facial motion features significantly correlated with the users' ratings are the ones that are very likely to capture micro-expressions [3];
2) psycho-physiological indications of a state of engagement seem to be highly subjective, as indicated by results in Table III;
3) although the best modality for predicting the level of engagement in our scenario was found to be the GSR channel, the best performance is obtained using a decision-level fusion strategy. Thus, the employed modalities seem to complementary contribute information for prediction of engagement level.
4) finally, according to the results reported in Table III, the performance of the majority-voting decision-fusion (that is, using equal weights for each classifier) is very close to the best possible performance.

TABLE IV. FEATURES SHOWING SIGNIFICANT ($p < .05$) CORRELATION WITH USERS' RATINGS. WE REPORT THE MEAN OF SUBJECT-WISE CORRELATIONS ($\bar{R}$), THE MOST NEGATIVE (R^-) CORRELATION, THE MOST POSITIVE CORRELATION (R^+), AND RELATED p-VALUES.

id	**Feature**	$\bar{R}$	R^-	R^+	p
1	Horizontal movement of the left lip corner (Mean)	-0.18	-0.69	0.25	0.014
2	Left cheek (Skewness)	0.24	-0.31	0.62	0.020
3	Level of meditation ($N_{samples} > \mu+\sigma$)	0.22	-0.11	0.43	0.007
4	Level of meditation ($N_{samples} < \mu-\sigma$)	0.22	-0.43	0.11	0.007
5	Power of delta band ($N_{samples} > \mu+\sigma$)	-0.19	-0.83	0.24	0.033
6	Power of delta band ($N_{samples} < \mu-\sigma$)	-0.19	-0.24	0.83	0.033
7	Power of high beta band (Skewness)	-0.13	-0.55	0.32	0.037
8	Zero crossing rate of skin conductance in slow response (SCSR) [0-0.2]Hz	-0.28	-0.54	-0.06	0.001
9	Proportion of negative samples in the derivative vs. all samples	0.27	-0.21	0.70	0.001
10	Average of absolute values of the derivative of skin conductance	-0.28	-0.75	0.51	0.003
11	Average of absolute values of the second derivative of skin conductance	-0.35	-0.74	0.07	0.004
12	Zero crossing rate of skin conductance in very slow response (SCVSR) [0-0.08]Hz	-0.24	-0.47	0.07	0.005
13	Average number of peaks in the GSR signal	-0.25	-0.69	0.71	0.009
14	SCSR mean of peaks magnitude	-0.21	-0.42	0.29	0.010
15	SCVSR mean of peaks magnitude	-0.21	-0.42	0.29	0.010



(a) (b)

Fig. 3. Pearson coefficients (a) and p-values (b) of a correlation analysis between the top-15 features in Table IV.

VI. CONCLUSIONS AND FUTURE WORKS

We have described a multimodal approach to engagement classification, framing it in the context of Affective Cinema. We envision novel experiences led by advances in the Affective Computing research, on the one hand, and Movie Studies, supported by the foreseeable wide deployment of consumer-level entertainment infrastructure (e.g., Smart-TVs) on the other.

Focusing on the relation between psycho-physiologic channels and affective responses to visual contents, this work contributes in understanding which channels and combinations thereof are effective for detecting a viewer's level of engagement. The use of cheap and commercially available hardware allows the deployment of similar Affective Cinema systems for both research and artistic purposes. Limitations of this work can be identified in *(i)* the single-viewer experience, and *(ii)* the relatively small size of the sample used.

Our findings show that the fusion of multiple modalities is beneficial to the classification performance, and provide insights on the cross-modal dynamics of the different affective channels investigated: in particular, EEG and GSR responses seem to contribute similarly to the engagement classification task under study; moreover, Facial Motion features seem to provide complementary information; finally, the psycho-physiologic features employed to asses the viewers' state of engagement seem to indicate high inter-subject variability.

Besides collecting data from more subjects under the same laboratory setting, for the purpose of further investigating their psycho-physiologic responses, future developments of this study will tackle the issue of moving beyond the single-viewer experience, towards the assessment of collective engagement in distributed scenarios. Leveraging the aforementioned (expected) wide deployment of Smart-TVs in conjunction with the ever-increasing availability of broad-band connectivity, the latter research line will build upon techniques from Social Network Analysis (SNA) and Natural Language Processing (NLP) communities.

Behavior-based technologies can, in our vision, foster new ways to produce and experience movies: not only Affective Computing researchers might benefit from a rising interest in real-word applications which would give access to large collections of real-usage data, but they could also have a major role in innovating the movie industry.

Acknowledgment. This work was partially supported by the FIRB 2008 project S-PATTERNS.

REFERENCES

[1] G. Chanel, K. Konstantina, and T. Pun. Gamemo: how physiological signals show your emotions and enhance your game experience. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 297–298. ACM, 2012.

[2] G. Chanel, C. Rebetez, M. Betrancourt, and T. Pun. Emotion assessment from physiological signals for adaptation of game difficulty. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 41(6):1052–1063, 2011.

[3] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[4] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[5] M. Greenwald, E. Cook, and P. Lang. Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *J. Psychophysiology*, 3:51–64, 1989.

[6] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, 2005.

[7] B. Jochems, M. Larson, R. Ordelman, R. Poppe, and K. P. Truong. Towards affective state modeling in narrative and conversational settings. In *INTERSPEECH*, pages 490–493, 2010.

[8] H. Joho, J. Staiano, N. Sebe, and J. M. Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools Appl.*, 51(2):505–523, 2011.

[9] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, 2008.

[10] S. Koelstra, C. Mühl, 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):18–31, 2012.

[11] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, Mar. 1977.

[12] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30:261–273, 1993.

[13] C. Latulipe, E. A. Carroll, and D. Lottridge. Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1845–1854, New York, NY, USA, 2011. ACM.

[14] J.-S. Lee and C.-H. Park. Robust audio-visual speech recognition based on late integration. *Multimedia, IEEE Transactions on*, 10(5):767–779, 2008.

[15] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Adv. Sig. Proc.*, 2004(11):1672–1687, 2004.

[16] B. O'Neill. Toward a computational model of affective responses to stories for augmenting narrative generation. In *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, ACII'11, pages 256–263, 2011.

[17] B. O'Neill and M. Riedl. Toward a computational framework of suspense and dramatic arc. In *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part I*, ACII'11, pages 246–255, 2011.

[18] N. Sebe and M. S. Lew. *Robust Computer Vision: Theory and Applications*. Kluwer Academic, 2003.

[19] N. B. Shaul. *Hyper-Narrative Interactive Cinema: Problems and Solutions*. Consciousness, literature & the arts. Rodopi, 2008.

[20] R. Sinha and O. A. Parsons. Multivariate response patterning of fear and anger. *Cognition and Emotion*, 10(2):173–198, 1996.

[21] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *T. Affective Computing*, 3(1):42–55, 2012.

[22] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, 3(2):211–223, 2012.

[23] J. Staiano, M. Menéndez, A. Battocchi, A. De Angeli, and N. Sebe. UX_Mate: from Facial Expressions to UX Evaluation. In *ACM DIS*, pages 741–750, 2012.

[24] M. Yuki, W. W. Maddux, and T. Masuda. Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, 43(2):303–311, Mar. 2007.

[25] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 126–133. ACM, 2007.