Valence-arousal evaluation using physiological signals in an emotion recall paradigm

Chanel, Guillaume; Ansari Asl, Karim; Pun, Thierry

How to cite

CHANEL, Guillaume, ANSARI ASL, Karim, PUN, Thierry. Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: 2007 IEEE International Conference on Systems, Man and Cybernetics, SMC 2007 : proceedings. Montreal (Canada). [s.l.] : Institute of Electrical and Electronics Engineers (IEEE), 2007. doi: 10.1109/ICSMC.2007.4413638638

This publication URL: https://archive-ouverte.unige.ch/unige:47741

Publication DOI: 10.1109/ICSMC.2007.4413638638

Valence-Arousal Representation of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses

Mohammad Soleymani, Guillaume Chanel, Joep Kierkels, Thierry Pun
Computer Vision and Multimedia Laboratory, Computer Science Department
University of Geneva,
Battelle Campus, Building A, Rte. de Drize 7,
CH - 1227 Carouge, Geneva, Switzerland
Phone: +41 (22) 379 0185
Mohammad.soleymani@cui.unige.ch

Keywords: Multimodal indexing, affective personalization and characterization, emotion recognition and assessment, affective computing, physiological signals analysis.

1 Introduction

In this paper, we propose an approach for affective representation of movie scenes based on the emotions that are actually felt by spectators. Affect, to be used as a multimodal indexing feature, is estimated from both physiological signals and multimedia content analysis. Emotions are not discrete phenomena but rather continuous ones. Psychologists therefore represent emotions or feelings in an n-dimensional space (generally 2- or 3-dimensional). The most famous such space, which is used in the present study, is the 2D valence/arousal space [1]. Valence in the range [0-1] represents the way one judges a situation, from unpleasant (negative emotion) to pleasant (positive emotion). Arousal in the range [0-1] expresses the degree of excitement, from calm to exciting. The main goal of this work is therefore to automatically estimate the arousal and valence created by a movie scene as affect indicators for emotional characterization of these scenes.

In order to represent affect of movie scenes, arousal and valence can be represented by grades; therefore, we compared self-assessment arousal and valence grades of the emotional content of scenes, with affective grades automatically estimated from physiological responses and from multimedia content analysis. There are only a limited number of studies on multimedia content-based affective representation/understanding of movies, and these mostly rely on self-assessments or population averages to obtain the emotional content of movies [2; 3]. While affect characterizations is usually based on self assessment, we propose here to use physiological signals as a modality for emotion assessment. In emotion assessment, physiological responses are valued for not interrupting users for self reporting phases; furthermore self reports are unable to represent dynamic changes. Physiological measurements give the ability of measuring the user responses dynamically.

With the advancement of wearable systems for recording peripheral physiological signals, it is becoming more practically feasible to employ these signals in an easy-to-use human computer interface and use them for emotion assessment [4; 5]. As opposed to signals from the central nervous system (electro-encephalograms) we therefore concentrated on the use of peripheral physiological signals for assessing emotion, namely: galvanic skin resistance (GSR), blood pressure which provided heart rate, respiration pattern, skin temperature, and electromyograms (EMG). In order to record facial muscles activity we used EMG from the Zygomaticus major and Frontalis muscles.

2 Material and Methods

A video dataset of 64 movie scenes of different genres, each one to two minutes long, was created from which content-based low-level multimodal features were determined. The scenes were extracted from the following movies: Saving Private Ryan (action, drama), Kill Bill, Vol. 1 (action), Hotel Rwanda (drama), The Pianist (drama), Mr. Bean's Holiday (comedy), Love Actually (comedy), The Ring, Japanese version (horror) and 28 Days Later (horror). Experiments were conducted during which physiological signals were recorded from spectators watching the clips. In order to obtain their self-assessed emotions, participants were asked to characterize each movie scene by arousal and valence grades using self-assessment Manikins (SAM) [6]. To avoid fatigue of the participants, the protocol divided the show into two approximately 2 hours sessions of 32 movie scenes in random order. Before each scene, a 30 seconds long emotionally neutral clip was shown as baseline. Eight healthy participants (three female and five male, from 22 to 40 years old) participated in the experiment, during which peripheral physiological signals and facial expression EMG signals were recorded for emotion assessment. Examples of recorded physiological signals in a surprising scene are given in Figure 1.

After finishing the experiment three types of affective information about each movie clip were available from different modalities:

- multimedia content-based information extracted from audio and video signals;
- physiological responses from spectators' bodily reactions (due to the autonomous nervous system) and from muscular activity originating from facial expressions;
- self-assessed arousal and valence, used as 'ground truth' for the true feelings of the spectator.

Next, we aim at demonstrating how those true feelings about the movie scenes can be estimated in terms of valence and arousal from the information that is either extracted from audio and video signals or contained within the recorded physiological signals. Corresponding to each of the 64 video clips, a set of multimedia feature vectors each composed of 64 elements was extracted. Each feature vector highlights a single characteristic (e.g., average sound energy) of the 64 movie scenes. These features, like audio energy, motion component in video, color variance, were chosen based on existing literature, for example [1] and [2]. Regarding the use of physiological signals, the following affect representative features were extracted: energy of EMG signals, average and standard deviation of GSR, temperature, the main frequency of respiration pattern acquired from respiration belt, heart rate, heart rate variability, and average blood pressure [3].

In order to select features that are the most relevant for estimating affect, the correlation between the single-feature vectors and the self-assessed arousal/valence vector was determined. Only the features with high absolute correlation coefficient between features and self assessments were chosen for affect estimation.

A linear regression of the features from physiological signals was used to estimate degrees of arousal and valence. The same regression was applied on multimedia features to estimate those degrees. In order to determine the optimum estimated points in the arousal/valence space (each corresponding to one particular emotion), the linear regression was computed by means of a linear relevance vector machine (RVM) from the Tipping RVM toolbox [6] This procedure was performed four times on the user self assessed arousal/valence, and on the feature-estimated arousal/valence, for optimizing the weights corresponding to: (1) physiological features when estimating valence; (2) physiological features when estimating arousal; (3) multimedia features when estimating valence; (4) multimedia features when estimating arousal.

2 Results and Conclusion

Physiological responses of participants were recorded while watching baseline clips as well as the movie scenes. The key features were extracted from these responses after removing the average baseline of physiological signals. After choosing the features having high correlation coefficients with self assessments, arousal/valence grades were estimated. Facial EMG signals and GSR were found to be highly correlated with valence and arousal grades for all subjects. Concerning multimedia content features, audio energy and visual features like color variance were found to be the most relevant for affect characterization.

Table 1. Average of Euclidean distances between estimated points and self-assessed points in the valence-arousal space, for participants 1 to 8.

	Average distances with multimedia features	Average distances with physiological features
1	0.25	0.22
2	0.20	0.18
3	0.19	0.21
4	0.20	0.21
5	0.23	0.25
6	0.24	0.28
7	0.19	0.18
8	0.17	0.19

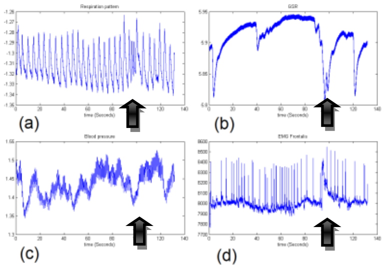

Figure 1. Physiological response (participant 2) to a surprising action scene. The following raw physiological signals are shown: respiration pattern (a), GSR (b), blood pressure (c), and Frontalis EMG (d). The surprise moment is indicated by an arrow.

A leave-one-out cross validation was done to evaluate the regression accuracy, using one sample (one scene out of 64) as test and the rest of the dataset as training set. The accuracy of the estimated arousal/valence grades was evaluated by computing the Euclidean distance in valence-arousal space between the estimated points and the self assessments (ground truth). Valence, arousal and the estimation error are expressed in normalized ranges [0-1]; Table 1 shows the average distance, which is the estimation error between self assessed points and estimated points.

The error distance mostly lies in the range [0-0.25], which is sufficiently small to allow the use of our estimations for emotional annotation or tagging. The weakness of affect estimation for some participants was often caused by inconsistent self assessments (such as declaring high valence values for negative emotions). It was observed that the error on valence estimation was lower than the error on arousal estimation. It is now planned to use prior information about movies such as genre, ratings, etc. from movie databases to enhance estimation accuracy.

Previous work in the field of emotion assessment was mostly focused on discrete classification of basic emotions; in contrast, the method presented here allows for a continuous emotion determination and representation. Quite promising results were obtained regarding the accuracy of multimodal affect estimation. This shows the possibility of using this method to automatically annotate and index multimedia data, for instance for designing personal emotion-based content delivery systems. Having a precise affect estimation method, a similar strategy could be applied to neuro-marketing where consumers' reactions to marketing stimuli could be predicted.

References

1. J. A. Russell and A. Mehrabian, "Evidence for A 3-Factor Theory of Emotions", Journal of Research in Personality, vol. 11, no. 3, pp. 273--294 (1977).
2. A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling", IEEE Transactions on Multimedia, vol. 7, no. 1, pp. 143--154 (2005).
3. H. L. Wang and L. F. Cheong, "Affective understanding in film", IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 6, pp. 689--704 (June 2006).
4. G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm", Trans. IEEE Conf. SMC, Systems, Man and Cybernetics, Montreal (October 2007)
5. J.A. Healey, "Wearable and Automotive Systems for Affect Recognition from Physiology", Ph.D. Massachusetts Institute of Technology, (May 2000).
6. J. D. Morris, "Observations: SAM: The self-assessment manikin - An efficient cross-cultural measurement of emotional response", Journal. of Advertising Research, vol. 35, no. 6, pp. 63--68, (1995).
7. M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine", Journal of Machine Learning Research, vol. 1, no. 3, pp. 211--244, (2001).