Towards Emotional Interaction: Using Movies to Automatically Learn Users' Emotional States

Eva Oliveira[1,2], Mitchel Benovoy[3], Nuno Ribeiro[4], and Teresa Chambe[2]

[1] Digital Games Research Centre, Polytechnic Institute of Cávado and Ave,
4750-810 Barcelos, Portugal
eoliveira@ipca.pt
[2] LaSIGE, University of Lisbon FCUL, 1749-016 Lisbon, Portugal
tc@di.fc.ul.pt
[3] Center for Intelligent Machines, McGill University, Montreal, Quebec, Canada
benovoym@cim.mcgill.ca
[4] CEREM – Centro de Estudos e Recursos Multimediáticos, Universidade Fernando Pessoa
nribeiro@ufp.edu.pt

Abstract. The HCI community is actively seeking novel methodologies to gain insight into the user's experience during interaction with both the application and the content. We propose an emotional recognition engine capable of automatically recognizing a set of human emotional states using psychophysiological measures of the autonomous nervous system, including galvanic skin response, respiration, and heart rate. A novel pattern recognition system, based on discriminant analysis and support vector machine classifiers is trained using movies' scenes selected to induce emotions ranging from the positive to the negative valence dimension, including happiness, anger, disgust, sadness, and fear. In this paper we introduce an emotion recognition system and evaluate its accuracy by presenting the results of an experiment conducted with three physiologic sensors.

Keywords: Affective computing, Emotion-aware systems, Human-centered design, Psychophysiological measures, Pattern-recognition, Discriminant analysis, Support vector machine classifiers, Movies classification and recommendation.

1 Introduction

Society's relation with technology is changing in such ways that it is predictable that, in the next years, Human Computer Interaction (HCI) will be dealing with users and computers that can be anywhere, and at anytime, and this changes interaction perspectives for the future. Human body changes, expressions or emotions would constitute factors that became naturally included in the design of human computer interactions [1]. HCI aims to understand the way users experience interactions and strives to stimulate the sense of pleasure and satisfaction [2] by developing systems that focus on new intelligent ways to react to user's' emotions. In fact, our cognition, creativity, health and aesthetic sensibility are highly influenced by our emotions,

P. Campos et al. (Eds.): INTERACT 2011, Part I, LNCS 6946, pp. 152–161, 2011.

playing an important role in our lives. Human Computer Interaction research community has been using physiologic, brain and behavior measures to study possible ways to identify and use emotions in human-machine interactions [3; 4]. However, there are still challenges in the recognition processes, regarding the effectiveness of the mechanisms used to induce emotions. The induction is the process through which people are guided to feel one or more specific emotions, which provokes body reactions. If the induction is not well succeed it would be more difficult to detect or recognize that changes. In this work, we present a novel pattern recognition system, based on discriminant analysis and support vector machine classifiers, which is validated using movies' scenes selected to induce emotions ranging from the positive to the negative valence dimension, including happiness, anger, disgust, sadness, and fear. The recognition engine was designed as a low-cost emotions recognition system and it is part of a system - iFelt - which is an interactive web video system developed to learn user's emotional patterns using movies' scenes selected to induce emotions. In this paper we present our recognition method and the accuracy results by presenting an experiment with three physiologic sensors. Following this introduction, section 2 makes a review of most relevant related work, and section 3 introduces the iFelt system. Section 4 introduces the iFelt experiment, with a focus on the elicitation method. The paper ends in section 5, with conclusions and a discussion of perspectives for future work.

2 Related Work

There are different approaches for the affective evaluation of interactions because emotions are not straightforward to understand and are very complex to model. There are evaluations based on emotional dimensions, such as valence [15] or based in categories [16]. The recognition of emotions through physiological patterns is an unconscious way to assess emotions, which can be even more accurate than self-assessment [5]. However, recognition processes commonly categorize emotions. Thus, the mapping of physiological patterns into emotional labels relies on the application of one or more emotional models. Based on directives from the Humaine Network of Excellence – an European Union initiative devoted to the evaluation of affective systems - there are three commonly used such models: 1) the Categorical model, which defines emotions as discrete states, 2) the Dimensional model, which bases emotion perspectives in a spatial circumplex of emotion properties (the most common refer to arousal and valence), and 3) the Appraisal model, which defends that emotions depend on people's own evaluation of events and its circumstances. The characterization of emotions by physiologic patterns faces some problems, but it has some advantages when compared to other recognition methods.

The characterization of emotions has still some limitations regarding the differentiation of emotions from physiologic signals, namely, in finding the adequate elicitation to target a specific emotion [4]. Moreover, due to the fact that being emotions, time-, space-, context- and individual-based, trying to find a general pattern for emotions, and trying to obtain a "ground truth" can be difficult: these are some of the major problems we currently face [3]. On the other hand, there is a major advantage when compared to facial and vocal recognition, because emotions cannot be intentional

- we cannot trigger the autonomic nervous system (ANS) contrarily to the so called "poker face" where people disguise facial expressions as well as vocal utterances [13]. Another common argument against physiologic measurements for emotion analysis is the fact that sensors are invasive, but new wearable sensors and related technological advances promise to allow for mouses that can incorporate sensors in the same way as, for example, the ones used in wrist band's to measure affective states [6]. Regarding the matter investigated in this work, Picard et al. also claim that cameras for facial recognition can be more invasive than physiologic sensors because they reveal identity or appearance besides emotional information. Films are by excellence the form of art that exploits our affective, perceptual and intellectual activity. In 1996, a research group [8] tested eleven induction methods and concluded that films are the best method to elicit emotions (positive and negative) and mainly when subjects (not studying psychology) are treated individually and are introduced to the purpose of the study. More recently, other researchers used films to induce emotions with different goals. One of the first works is from [8], which tried to find films that induce differential emotional states (dimensional space) while J. Gross et al. [5] tried to find as many films as possible to elicit discrete emotions and find the best films for each discrete emotion. In light of the evidence of distinct physiological responses of emotion, the machine learning and HCI communities have each investigated the automatic recognition of emotions. Picard et al. [6] pioneered this area by showing that some emotional states can be recognized automatically using physiological signals and pattern recognition methods. In [18], the authors made an overview represented in a table (Table 1) that we complete with the elicitation method used, which presents the most relevant studies regarding emotion recognition, using different physiologic signals and different classification methods. Methods that have used movies had the lower results, which in our opinion demands for new ones that can more effectively explore the power of movies to induce emotions given that, according to [8], movie scenes are considered to be one of the best methods to induce emotions.

Table 1. Four studies on automatic physiological-driven classification on affect

Ref	Year	Signals	Participants	Features	Select./Red.	Classifiers	Target	Results	Elicitation
[6]	2001	C,E,R,M	1	40	SFS, Fisher	LDA	8 emotions	81%	images
[3]	2008	C,E,R,M	3	110	SBS	LDA	4 emotions	70%	music
[14]	2008	C,E,R,M	40	5	-	SVM	5 emotions	47%	movies
[7]	2009	C,E,R.EE	10	18	-	LDA, SVM,RVM	3 emotions	51%	movies

Signals: C: cardiovascular activity; E: electrodermal activity; R: respiration; M: electromyogram and;EE: electroencephalographic; Selection: SFS: Sequential Forward Selection; SBS: Sequential Backward Selection; Fisher: Fisher projection; Classifiers: SVM: Support Vector Machine; RVM: Relevance Vector Machines; LDA: Linear Discriminant Analysis;

The collection of physiologic data when users are watching movies was recently developed in the psychology area to test whether films can be efficient emotional inductors, which could help psychologists in specific treatments [10], or in the computer science area to automatically summarize videos according to the emotional impact on their viewers [11,12]. Money & Agius (2009) [12] report an experiment on how user physiological responses vary when elicited by different genres of video content in order to validate the development of personalized video summaries. They also showed specific video segments to a group of viewers monitored with biometric artifacts (electro dermal response, respiration amplitude, respiration rate, blood volume pulse and heart rate) and concluded that there are significant differences between users when watching the same video segments. Money et al. (2009) [12] make use of the dimensional theory to classify the affective results in this study.

In the next section we describe the classification procedure of our system.

3 The iFelt Classification and Recognition Engines

iFelt is an interactive web video system designed to learn users emotional patterns, and explore this information to create emotion based interactions. The system is composed of two components. The "Emotional Recognition and Classification" component which performs emotional recognition and classification of user's emotional states and the "Emotional Movie Access and Exploration" component that explores ways to access and visualize videos based on their emotional properties and users' emotions and profiles. In this paper we are focused on evaluation of the Emotional Recognition and Classification component whereas the Emotional Movie Access and Exploration are thoroughly described in [17]. In this section we will describe in detail the elicitation procedure and how we collect, process, and classify biosignals to learn and later recognize emotional patterns as a low-cost emotions recognition system.

3.1 Emotional Elicitation

Every emotion recognition process needs to address the problem of how to induct emotions. Our emotional recognition and classification component is grounded in the induction of emotional states by having users watch movie scenes. The selection of the movie scenes was based in our own judgment and also based on J. Gross' work [5]. We have selected 3 scenes from their work, and a selection of 13 additional scenes that in our opinion were emotionally intense and represent the set of emotions needed to test our recognition engine, with an average duration of 2 minutes 22 seconds per scene. The system, iFelt, uses the subjects' data obtained while watching movie scenes to create an engine to enhance automatic recognition of users' emotional states. The selected movie scenes induct subjects to feel five basic emotions (happiness, sadness, anger, fear and disgust) and the neutral one, so, every subject should see 16 scenes (four of happiness, four of sadness, four of fear, two of disgust and two of anger) and one neutral scene. Based on their feedback, we associated the captured physiological signals with emotional labels, and trained our engine.

3.2 Biosignal Capture

Biosignal recording uses biosensors for measuring Galvanic Skin Response (GSR), Respiration (Resp) and Electrocardiogram (ECG) and is responsible for users' biosignals recording and signal processing pipeline each sampled at 256 Hz. These sensors were specifically chosen as they record the physiological responses of emotion, as controlled by the autonomous nervous system, and also because they were already proven sufficient for measuring our five basic emotions (happiness, sadness, anger, fear and disgust) [4,12]. From the heart we measured heart rate, heart rate acceleration, and heart rate variability. We also measured the first and second derivative of the heart rate. For respiration, we measured the rate, the amplitude and the first and second derivative of the rate. For GSR, we measured the stats plus the number of Skin Conductance Responses (SCRs).

3.3 Emotional Classification

Emotionally relevant segments of the recordings that are free of motion artifacts are hand-selected and labeled with the help of the video recordings and subjects responses. High-frequency components of the signals are considered to be noise and filtered with a Hanning window [13]. For the GSR signal, a cutoff frequency of 2.0 Hz was used, whereas for the ECG and respiration signals, cutoffs of 128 Hz and 10 Hz were used in the filters. To account for inherent physiological differences between participants, the mean of the 3-minute silent baseline data preceding each stimulus onset was subtracted from the *active* data and the signal range was adjusted to a [0;1] interval. We extract six common statistical features from each type of the filtered biosignals, of size *N* (X_n, n □ [1...*N*]): the filtered signal mean, the standard deviation of the filtered signals, the mean of the absolute values of the first differences of the filtered signals, the mean of the absolute values of the first differences of the normalized signals, the mean of the absolute values of the second differences of the filtered signals and the mean of the absolute values of the second differences of the normalized signals. A total of 6 x 3 = 18 features are computed from the three types of biosignals. These features were chosen to cover the typically measured statistics in physiological recordings. The advantage of using relatively simple statistical feature is that these can be computed efficiently, opening the door for real-time applications. We employed digital signal processing and pattern recognition, inspired by statistical techniques used by Picard [6], in particular in our use of *sequential forward selection* (a variant of sequential floating forward selection). We specifically chose statistical features, as these are computationally easy to produce, which opens the way to future real-time systems, choosing only classifier-optimal features, followed by *Fisher dimensionality reduction*. For the classification engine, however, we implemented linear discriminant analysis (LDA) rather than the maximum *a posteriori* used by Picard, since our previous experiments with physiological data has shown that LDA produces high classification rates. LDA was selected to dimensionally reduced data by building a statistical model for each emotional class and then cataloguing novel data to the model that best fits. We are thus concerned with finding which classification rule (discriminant function) best separates the emotion classes. LDA finds a linear transformation Φ of the x and y axes that yields a new set of values providing an accurate discrimination between the classes. The transformation thus seeks to rotate the axes with parameter v, so that when the data is projected on the new axes, the difference between classes is maximized.

3.4 Emotion Recognition

The pattern recognition module uses discriminant analysis, support vector machine (SVM) and k-Nearest Neighbour (K-NN) classifiers [20] to analyze the physiological data and it was validated by the usage of specific movie scenes selected to induce particular emotions. We used the greedy sequential forward floating selection (SFFS) algorithm to form automatically a subset of the best *n* features from the original large set of *m* ($n < m$). SFFS starts with an empty feature subset and, on each iteration, exactly one feature is added. To determine which feature to insert, the algorithm tentatively adds to the candidate feature subset one that is not already selected and tests the accuracy of a *k*-NN classifier built on this provisional subset. A feature that results in the highest classification accuracy is permanently included in the subset, while a poor feature is deleted. The process stops after an iteration where no feature additions or deletions cause an improvement in accuracy. The resulting feature set is now considered optimal. The SVM classifier generates parallel separating hyperplanes that maximize the margins between the subjects' data, which has the effect of minimizing generalization error. Because SVMs are binary classifiers by nature, we used a one-versus-all decision strategy to perform multiclass classification. This technique divides the single multiclass problem into *c* binary classifiers in which the one with the highest recognition confidence value assigns the final label.

We trained the SVMs with a Radial Basis Functions (RBF) kernel for which the parameters were determined concurrently using an iterative grid selection technique that finds the best combination using the training error of the classifier as a performance metric [20]. The SVM module outputs the identity of the recognized person, along with a classification confidence value based on the distance between the feature vector of the probe and the hyper-margin of the closest subject. The *k*-NN classifier used here classifies a novel object *r* by a majority of “votes” of its neighbors, assigning to *r* the most common class amongst its *k* nearest neighbors, using the Euclidean distance as metric. It was found through experimentation that a value of $k = 5$ resulted in the best possible selected feature subset.

4 Evaluation and Discussion

We used a portable system - Nexus 4 - with 3 inputs channels for ECG, Respiration and HR. It is a wireless 'real-time' data link computer, and can store up to 24 hours of physiological data on its built-in flash memory. The computer software used to process data was Biotrace[1]. Eight participants, averaged 34 year, were submitted to an experiment. Because we presented full length feature films to our subjects, we were able to recruit only eight subjects. However, we deem this number of participants acceptable as a first attempt to classify their emotional reactions. After the subject arrived, the electrodes were attached and the recording system checked. We developed a web interface to easily perform the learning procedure, which began by asking the user to rest for 3 minutes, in a quiet mode. At the beginning of the sessions, a 3-minute silent baseline was recorded, while the participants engaged in focused relaxation by limiting their concentration to their respiration. This pre-stimulus

[1] http://www.mindmedia.nl/english/biotrace.php

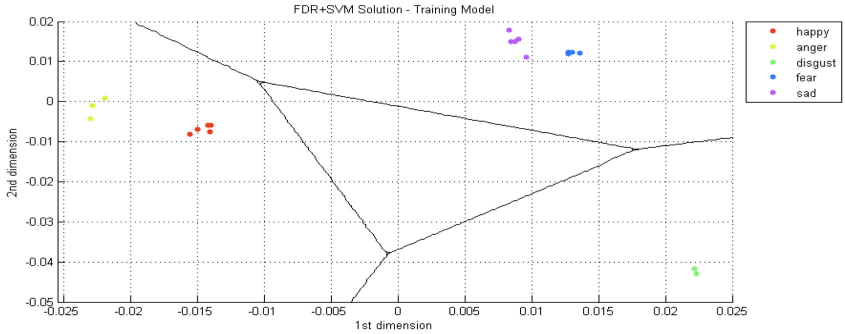

Fig. 1. Class Clustering of Five Emotional States

relaxation time was critical to stabilize the physiology to a homeostatic state, as some participants initially exhibited anxiety at being the focus of attention and being *wired* to the sensors. Then the neutral scene was shown to collect the neutral state right after the baseline. The sequence of 16 movie scenes began with the happiest scenes first alternating with fear related ones. Then, we showed the disgust related scenes alternating with fear related scenes. The last set included the sad scenes, alternating with anger, but chosen in a way that the most intense sad emotions were presented towards the end of the experimental session following recommendations of [5]. We also chose this sequence because it was evident from our previous sessions that when people watched intense sad emotions they became emotionally unavailable to be amused, or even frightened, by any movie scene. In turn, this choice was found to be the most adequate sequence in order to improve recognition accuracy. Upon watching completion, subjects answered a questionnaire describing the emotion from the set which they considered was the dominant, in which intensity and if they enjoyed watching it. After the movies session, the experimenter labeled minute per minute every movie used in this experience with the expected emotion, to compare with the results of the recognition. The learning phase expressed in Figure 1 demonstrates the class clustering of five emotional states: happiness, anger, sadness, fear and disgust projected on the 2D Fisher space along with the SVM class-boundaries of one person, which is satisfactory once we want to learn each user's emotional patterns. It is a positive result because it shows that emotion instances distribution of the same category are near to each other, and apart from the other, which reveals coherence and confidence. To evaluate the pattern recognition process, we compared the expert-labeled movies with the output of the automatically classified data from the iFelt system. The expert labeled the movies in consecutive blocks of one minute length. To compare with the output of the recognition system, which produced classification scores on consecutive five-second windows, the median score was computed over minute long segments of the classified data. At least two subjects watched each of the eight movies, and were classified by the system. With the SVM classifier, the overall average recognition rate is 69% (s.d. 5.0%), which represents a 49% improvement over random choice. Because we are classifying 5 classes, the random probability is 1/5 (20%), which is what the simplest classifier could do. However, our classifier performs at 69%, which is 49% higher than the random choice of 20%. Our k-NN classifier produced an overall average recognition rate of 47% (s.d. 9.3%). The SVM classification score shows promise that the iFelt recognition system can be used to automatically evaluate human emotions. Although it is important to note that further research needs to be conducted to optimize both the classification algorithms for the type of data used here and the scenes used to learn the emotional patterns. However, two key positive aspects of the system emerged which is the use of easily computed statistical features, which can be used to develop real-time classification systems, and a quite reasonable recognition rate, with only three sensors when compared with the works listed in section 2.

5 Conclusion and Future Work

We presented an emotional recognition system capable of automatically recognizing a set of human emotional states using psychophysiological measures and pattern recognition techniques based on discriminant analysis and support vector machine classifiers. Regarding our experiment methodology we concluded that, every time a subject watched an intense sad movie scene, such as atrocities performed in genocides, the person could not feel happiness in any kind of subsequent happy scene. Another conclusion we made was that people, after watching a very disgusting scene, such as watching sputum in a very expository way, normally kind of forgot the past sensations (sadness, happiness, fear) and felt a little indisposed. The third observation we made was that after 16 movie scenes, in a total duration of 40 minutes, the experience became emotionally intense and subjects became tired and lost their good mood. In order to test the performance of our system, a novel emotion elicitation scheme, based on emotions induced by watching selected movie scenes was presented, engendering a moderate degree of confidence in collected, emotionally relevant, biosignals. Discrete state recognition via physiological signal analysis, using pattern recognition and signal processing, was shown to be reasonably accurate. A correct average recognition rate of 69% was achieved using sequential forward selection and Fisher dimensionality reduction, coupled with a Linear Discriminant Analysis classifier. An important conclusion of this work is that it is easily computed statistical features can be implemented in real-time classification systems, which allows moving towards an emotional interaction system, and also reveals that few features can achieve pretty good results. Even though physiologic sensors are invasive, recent technological advances is resulting in the development of wearable sensors less intrusive, which make our recognition engine useful in assessing human emotional states during human-computer interactions and further validates the use of movies as powerful emotional triggers. Our ongoing research also intends to support *real-time* classification of discrete emotional states, adding also arousal/valence mappings from biosignals for multimedia content classification and user interaction mechanisms by developing emotional aware applications that react in accordance to user's' emotions. In the context of our work we are considering using emotion recognition to automatically create emotional scenes, recommend movies based on the emotional state of the user and adjust interfaces according to user's emotions and based on emotional regulation theories. By creating emotional profiles for both movies and users, we are developing new ways of discovery interesting emotional information in unknown or unseen movies, compare reactions to the same movies among other users, compare directors intentions with users effective impact, analyze over time our reactions or directors tendencies. Measuring physiological signals of users is a natural input mechanism that can be used to automatically classify information with emotional semantic which can enhance retrieval systems by adding emotional information to search engines, as can be used in interactive applications that take into account the affective states of users.

Acknowledgements. This work is partially supported by FCT through LASIGE Multiannual Funding, PROTEC 2009 (SFRH/BD/49475/2009) and VIRUS research project (PTDC/EIA-EIA/101012/2008).

References

1. Being Human: Human-Computer Interaction in the Year 2020 (2007), `http://research.microsoft.com/hci2020/`
2. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: Human-Computer Interaction, 3rd edn. Prentice Hall, Englewood Cliffs (December 2003)
3. Kim, J., André, E.: Emotion Recognition Based on Physiological Changes in Listening Music. IEEE Trans. on Pattern Analysis and Machine Intelligence 30(12), 2067–2083 (2008)
4. Maaoui, C., Pruski, A., Abdat, F.: Emotion recognition for human-machine communication. In: 2008 IEEERSJ International Conference on Intelligent Robots and Systems, pp. 1210–1215 (2008)
5. Rottenberg, J., Ray, R., Gross, J.: Emotion elicitation using films. In: The Handbook of Emotion Elicitation and Assessment (2007)
6. Picard, R.W., Vyzas, E., Healey, J.: IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175–1191 (2001g)
7. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. Int. J. Hum.-Comput. Stud. 67, 8 (2009)
8. Westermann, R., Spies, K., Stahl, G., Hesse, F.W.: Relative effectiveness and validity of mood induction procedures: A meta-analysis. European Journal of Social Psychology 26(4), 557–580 (1996j)
9. Philippot, P.: Inducing and assessing differentiated emotion-feeling states in the laboratory. Cognition & Emotion 7(2), 171–193 (1993h); Gross, J.J., Levenson, R. W.: Emotion elicitation using films. Cognition & Emotion 9(1), 87–108 (1995i)
10. Kreibig, S.D., Wilhelm, F.H., Roth, W.T., Gross, J.J.: Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. Psychophysiology 44(5), 787–806 (2007)
11. Soleymani, M.S., Chanel, C.G., Kierkels, J.K., Pun, T.P.: Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes. In: International Symposium on Multimedia, pp. 228–235 (2008)
12. Money, A.G., Agius, H.: Analysing user physiological responses for affective video summarization. Displays 30(2), 59–70 (2009)
13. Cliffs: Discrete-Time signal processing. Prentice-Hall, New Jersey (1989)
14. Lichtenstein, A., Oehme, A., Kupschick, S., Jürgensohn, T.: Comparing Two Emotion Models for Deriving Affective States from Physiological Data. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868, pp. 35–50. Springer, Heidelberg (2008)

15. McQuiggan, S., Lee, S., Lester, J.: Predicting user physiological response for interactive environments: An inductive approach. In: Proceedings of the 2nd Artificial Intelligence for Interactive Digital Entertainment Conference, pp. 60–65 (2006)
16. Isbister, K., Höök, K., Sharp, D., Laaksolahti, J.: The Sensual Evaluation Instrument: Developing an affective evaluation tool. In: Proceedings of ACM CHI (Conference on Human Factors in Computing), Montréal, Québec, Canada (2006)
17. Oliveira, E., Martins, P., Chambel, T.: iFelt: Accessing Movies Through Our Emotions. In: EuroITV 2010, 9th European Conference on Interactive TV and Video, ACM SIGWEB, SIGMM & SIGCHI, Lisboa, Portugal, June 29-July 01, 10pgs (2011)
18. Van der Zwaag, M.D., van den Broek, E.L., Janssen, J.H.: Guidelines for biosignal driven HCI. In: ACM CHI2010 Workshop - Brain, Body, and Bytes: Physiological User Interaction, Atlanta, GA, USA (April 11, 2010)
19. Bishop, C.M.: Pattern Recognition and Machine Learning, 740 pages. Springer, Heidelberg (2006)