SIGNAL PROCESSING FOR RECOGNITION OF HUMAN FRUSTRATION

Raul Fernandez and Rosalind W. Picard

M.I.T. Media Lab
20 Ames Street
Cambridge, MA 02139-4307
{galt, picard}@media.mit.edu

ABSTRACT

In this work, inspired by the application of human-machine interaction and the potential use that human-computer interfaces can make of knowledge regarding the affective state of a user, we investigate the problem of sensing and recognizing typical affective experiences that arise when people communicate with computers. In particular, we address the problem of detecting "frustration" in human computer interfaces. By first sensing human biophysiological correlates of internal affective states, we proceed to stochastically model the biological time series with Hidden Markov Models to obtain user-dependent recognition systems that learn affective patterns from a set of training data. Labeling criteria to classify the data are discussed, and generalization of the results to a set of unobserved data is evaluated. Significant recognition results (greater than random) are reported for 21 of 24 subjects.

1. INTRODUCTION

1.1. Motivation

Affective Computing is a newly emerging field which has been defined as "computing that relates to, arises from, or deliberately influences emotions" [1]. There exists a variety of challenging open research problems in this area, especially concerning the issue of emotion detection and classification. The aim of the work presented here is to develop systems that learn typical responses of a user during what psychologists call "a high arousal situation," in particular the kind of frustration that arises while interacting with a computer [2]. The emphasis of this work is on affect detection and classification; it should be noted, however, that this work is motivated by and has implications for other functionalities of a complete affective system. For instance, the ability to evaluate the affective state of a user can help the interface designer build more intelligent adaptive interfaces.

1.2. Experimental Methodology

One of the major obstacles in implementing a system that acts on affective signals resides in collecting suitable data for developing a computational model for the system. To this effect the following experimental situation was conducted in a laboratory setting.

The subjects came to the lab to participate in an experiment that had been advertised on bulletin boards on campus. The subjects were informed that they were to participate in a visual perception experiment, and the real purpose of the experiment was not revealed to them until the debriefing period at the end of the experiment. The experiment consisted of a simple computer game in which there was a monetary reward motivation for superior performance. In order to create a very competitive environment, all subjects were made aware that this was a competitive task and that their reward was dependent on their performance with respect to the rest of the players. The actual task consisted of a simple interaction with a computer in which the user was presented with a series of slides containing multiple items of four different shapes and asked to indicate which shape contained the largest number of items. The only user interface allowed to the subjects was a mouse which they used to click on an icon corresponding to the desired answer. The mouse was designed to enter a delay mode at pre-specified intervals during the execution of the experiment. This was done to try to create the impression that the mouse was failing to work at random points, thus interfering with the subject's goal of finishing in a short time. Subjects were given the choice of repeating the experiment two more times (up to a total of 3 experimental sessions) in order to gather more data for the training phase. The number of stimuli in each session were 5,4, and 7 respectively. During the execution of the experiment, each subject's electrodermal response (GSR) and blood volume pressure (BVP) were sensed, sampled at 20 Hz, and recorded for later processing. Fig. 1 shows a typical response obtained from one of the experimental subjects. The vertical bars indicate the onsets of the stimuli (the times when the mouse failed to work).

2. MODELING

2.1. Defining a Ground Truth

In a classical recognition problem a set of data is used for learning the properties of the model under the different classes to recognize. The classification of this training data is usually fixed, and this knowledge is then used to derive the properties of the separate classes. However, establishing a proper labeling for the training data of this experiment

We would like to thank BT, HP Labs, NEC, and the Media Lab TTT Consortium for their support of this research.

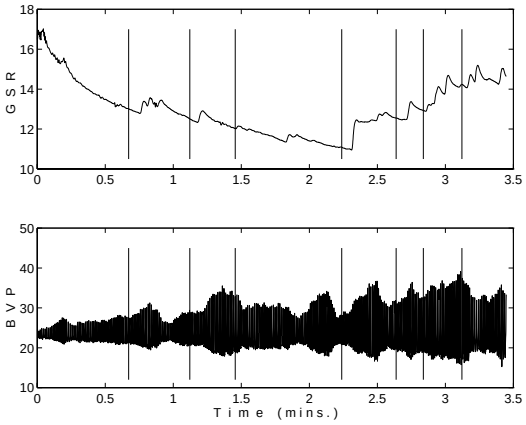

Figure 1: Physiological Signals

is one of the aspects of this problem which deserves careful consideration since the class categorizations we shall use to label the data are based on a presupposed reaction to a stimulus, which is not guaranteed. In other words, there is an uncertainty associated with the class to which the data belongs. The following may be true:

- a stimulus failed to induce a high arousal response
- a subject showed a high arousal response in the absence of the controlled stimulus due to another uncontrolled stimulus

In the discussion that follows, in order to make the problem tractable, we consider the idealized case where there is a response if and only if a stimulus occurs. An intuitive approach to labeling regions consists of using a time window after the onset of the stimulus to designate temporal regions corresponding to *"frustration"* (F) and *"non-frustration"* ($\overline{F}$). Based on this, we have drawn the following labeling rules to assign a classification to the data:

- Allow a 1-second period after the stimulus onset without classification
- Following this, use a 10-second (rectangular) window to designate a F-class
- Allow a 5-second period after this without classification
- Designate the remaining data until the ocurrence of the next stimulus as $\overline{F}$-class

It is not clear how long the time window should be so as to capture physiological changes of interest due to the stimulus. We have used a 10-second window, and in order to reduce the uncertainty between transitions between F and $\overline{F}$, have allowed a latent period of 5 seconds before a new classification is made. The 1-second latency period following the stimulus attempts to model a delayed reaction time between the onset and the response.

2.2. Learning Algorithms

Human physiology behaves like a complex dynamical system in which several factors, both external and internal, shape the outcome. In approximating such a system, we are interested in modeling its dynamical nature and, given that knowledge of all the independent variables that affect the system is limited, we want to approach the problem in a stochastic framework that will help us model the uncertainty and variability that arise over time.

We have approached this problem by implementing a Hidden Markov Model (HMM) based recognition system. We have focused on user-dependent systems not only to account for variability across subjects, but also motivated by the potential applications of an interface that learns physiological affective patterns from an individual.

Estimating the parameters of an HMM from the training data has been discussed at length in the literature. We have used the standard Baum-Welch re-estimation algorithm to obtain initial parameters of the model [3],[4], and then improved initial estimates with the embedded Baum-Welch algorithm [5]. Once the systems have been trained, the performance is assessed by applying a Viterbi decoder to a set of testing data and producing a set of transcription labels [6]. Efficient software implementation of these learning algorithms was done using the Hidden Markov Toolkit (HTK) (version 2.0) (developed at Cambridge University and Entropic Research Laboratories) [5].

How to choose a model structure (i.e. number of states, output distribution types, and model topology) is not clear for the signals we have chosen to measure. In order to investigate the performance of different model types, and also allow model types to be user dependent, we have considered a family of possible HMMs, categorized according to:

- number of states: $N \in \{4, 5, 6, 7\}$
- number of Gaussian components in output distribution: $K \in \{1, 2\}$
- type of covariance matrix Σ: diagonal, full
- transition probability type: left-to-right, ergodic

HMMs for each set of user data were trained over all resulting combinations to select the best performer.

2.3. Feature Extraction

From a set of raw data, such as shownn in Fig. 1, we need to obtain a set of significant feature signals that might have correlates with internal affective states. This is one of the most important research problems that exist in this area: the mappings between affective states and physiological states is still an area which is being investigated at large in the psychophysiology community. In deciding on a feature set, we must account for classical measures of affective states (i.e., level of arousal as registered in a GSR signal, heart acceleration, etc), while bearing in mind that we can also allow the models we are using to exploit more complex dynamic patterns that might not have received so much attention in other studies.

Let $g[n]$ and $b[n]$ represent the discrete time signals obtained by sampling the GSR and BVP signals. It is customary to measure changes in the GSR signal to predict levels of arousal. Motivated by this, we define the following signals:

$$g_\mu[n] \doteq g[n] - \frac{1}{N}\sum_{k=0}^{N-1} g[n-k] \quad (1)$$

$$g_v[n] \doteq \frac{1}{N-1}\sum_{k=0}^{N-1}\left(g[n-k] - \frac{1}{N}\sum_{l=0}^{N-1} g[n-l]\right)^2 \quad (2)$$

Equation (1) is just the GSR signal minus a time varying local sample mean obtained by windowing the GSR signal with an advancing N point rectangular window whereas (2) is a time varying estimate of the local variance of the signal. It is obtained by windowing the GSR with an advancing N point rectangular window and evaluating the unbiased sample variance for every point.

Inspection of the BVP signal reveals that it exhibits a richer structure than the GSR signal.

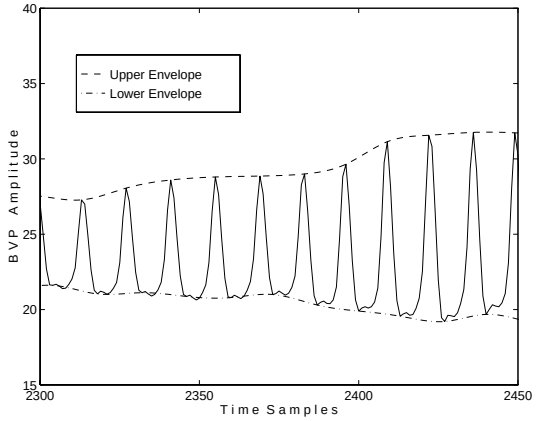

Figure 2: Portion of a BVP signal

As shown in Fig. 2, the BVP signal for instance has a richer harmonic content due to its periodic behavior over time. Its amplitude is also modulated in a way that we might exploit for feature extraction. In particular, the time-varying upper and lower bounds on the amplitude of the signal—let us call them $b_u[n]$ and $b_l[n]$—may be used to define:

$$b_p[n] \doteq b_u[n] - b_l[n] \quad (3)$$

as the "pinch" of the signal.

By finding the peaks of the BVP we can find the peak-to-peak intervals, and by taking this interval as one period of the harmonic oscillations, we estimate local frequency as the reciprocal of the peak-to-peak intervals. Let $T_{p2p}[n]$ denote the number of samples between adjacent peaks (experimentally, we found this simple method to agree with the results of the first harmonic obtained by the more computationally intensive short time Fourier analysis). By registering changes in the value of $T_{p2p}[n]$ (cycle duration), we can obtain an estimate of acceleration and deceleration of the harmonic cycles. This is a signal of interest since BVP is highly correlated with heart rate, and therefore so are changes in the BVP frequency. Define then:

$$b_{\Delta T}[n] \doteq T_{p2p}[n] - T_{p2p}[n-1] \quad (4)$$

Some of the rich structure of the BVP may be described by changes over time as well as frequency. A way of studying this behavior is to observe its evolution in the time-frequency plane; one such approach was hinted at when we mentioned the short time Fourier transform. Another time-frequency approach to have received much attention lately, in particular in the study of non-stationary biosignals for feature extraction, is wavelet analysis [7]. Let us assume that an orthogonal wavelet decomposition of the BVP signal is implemented with J levels of resolution (i.e., via a filter bank decomposition), and let $\{\hat{d}_{jk}\}$ be the wavelet coefficients quantifying the level of "detail" at level j. Similar to (2), we can obtain a local estimate of the variance of the wavelet coefficients by defining:

$$d_v^{(j)}[n] \doteq \frac{1}{M-1}\sum_{k=0}^{M-1}\left(d^{(j)}[n-k] - \frac{1}{M}\sum_{l=0}^{M-1} d^{(j)}[n-l]\right)^2 \quad (5)$$

where $d^{(j)}[n]$ is a time series obtained by interpolating the wavelet coefficients $\{\hat{d}_{jk}\}$ (this is done for the convenience of working with time series that are all of the same length). Using (1), (2), (3), (4), and (5), we then define the following 5-dimensional feature vector:

$$\mathbf{x}[n] = \left[g_\mu[n], \quad g_v[n], \quad b_p[n], \quad b_{\Delta T}[n], \quad d_v^{(j)}[n]\right]^T \quad (6)$$

In the implementation that follows, these are the values of the constants used in obtaining the features: $N = 200$ in (1) and (2) (windowing the GSR with a 10 second window), $M = 30$ in (5) (windowing with a 1.5 second window), $j = 3$ in (5) using Daubechies-4 orthogonal wavelets [8].

Also, in order to avoid numerical errors (especially when estimating covariance matrices of very small values), the extracted features were scaled to exhibit a higher range of amplitudes. For these simulations we used the following scale factors: $[4,\ 10,\ 0.5,\ 500,\ 0.02]^T$. The values were chosen to keep the data within the range ± 2.

3. RESULTS

Recognition rates were evaluated by determining the percentage of instances of F and $\overline{F}$ classes that were classified correctly within the time boundaries determined by the ground truth. Part of the problem consists of finding potential boundaries between the classes, and determining whether the boundaries established in the recognition phase contain the classes assigned in the ground truth. Rather than treating this as a point-by-point classification, we are evaluating it according to the number of labels that were correctly "detected". Fig. 3 shows the distribution of the *overall* recognition rates for individual subjects for the training and testing sets. Furthermore, Fig. 4 shows the recognition rates obtained for each label that was being classified.

Overall recognition rates greater than 50% were obtained for $\frac{7}{8}$ of the subjects in the data set. It should be kept in mind that people are not perfect at recognizing affective expression, and we do not expect perfect recognition from a computer. The goal to beat, at least initially, is better than random or 50%. This was achieved by the method for 21 of 24 subjects.

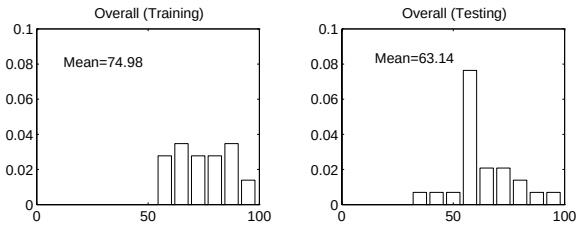

Figure 3: Histogram of Overall Recognition Rates

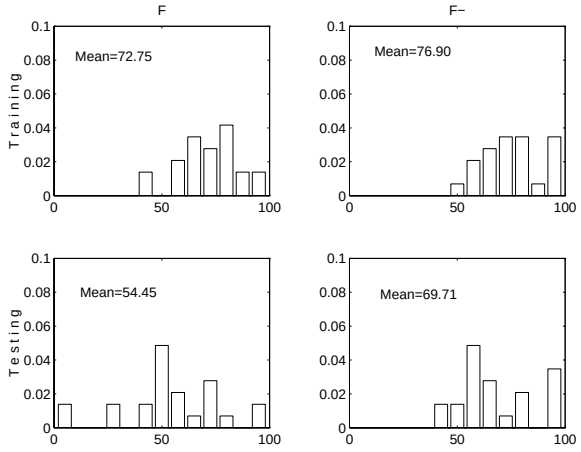

Figure 4: Histogram of Recognition Rates for each Label

No single HMM structure was found to consistently perform better across all subjects, although left-to-right topologies tended to be the best performer for most subjects; as were HMMs with 6 and 5 states predominantly, and unimodal output distributions. Full and diagonal covariance matrices performed comparatively. In addition, we evaluated the performance of the systems under an alternative ground truth. The objective under this ground truth was to model the subjects' habituation to the stimuli and anticipation of where the mouse failures might have ocurred. The system was found to perform better, however, under the standard ground truth.

4. CONCLUSIONS AND FURTHER RESEARCH

This work constitutes one research effort in the area of recognition of human affect for affective computing applications. In particular, this work has approached the domain of human-machine interaction where there is arguably much room for improving the quality of human-computer interfaces. Motivated by the present inability of these interfaces to incorporate much of the affective nature of a human response into their system, we have explored the topic of recognition of human frustration as it arises when humans confront an interface which offers a faulty or inefficient design.

The data analysis was based on a set of biosignals collected in a laboratory setting. Using Hidden Markov Models, a stochastic learning algorithm for time series, and various signal-dependent features we developed subject-dependent systems that learn and predict patterns corresponding to the presence and absence of the affective experience we wanted to model. Using this approach, we have obtained recognition rates over 50% for 21 out of 24 subjects. Further research in this area include redefining the ground truth to account for some of the difficulties explained in Section 2.1, and finding alternative features from the data to improve the results.

5. REFERENCES

[1] R. W. Picard. *Affective Computing.* The MIT Press, Cambridge, Massachusetts, 1997.

[2] Raul Fernandez. Stochastic modeling of physiological signals with hidden markov models: A step toward frustration detection in human-computer interfaces. Master's thesis, Massachusetts Institute of Technology, 1997. Downloadable from: `http://affect.www.media.mit.edu/projects/affect/AC_research/projects/frustration_detection.html`.

[3] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 1986.

[4] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Information Technology Series. Edinburgh University Press, Edinburgh, 1990.

[5] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *HTK - Hidden Markov Model Toolkit.* Entropic Research Laboratory, Inc.

[6] S. J. Young, N. H. Rusell and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical report, Cambridge University Engineering Department, 1989.

[7] M. Akay. Wavelet applications in medicine. *IEEE Spectrum*, pages 50–56, May 1997.

[8] Ingrid Daubechies. *Ten Lectures on Wavelets.* SIAM, 1992.