M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 509
To appears in special issue of IWC

Frustrating the User On Purpose: A Step Toward Building an Affective Computer

Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, Rosalind W. Picard
Massachusetts Institute of Technology
Media Laboratory
20 Ames St., Cambridge, MA 02139
{rise,galt,phaedra,picard}@media.mit.edu

Abstract

Using social science methods to induce a state of frustration in users, we collected physiological, video and behavioral data, and developed a strategy for coupling these data with real-world events. The effectiveness of the proposed strategy was tested in a study with thirty-six subjects, where the system was shown to reliably synchronize and gather data for affect analysis. Hidden Markov Models were applied to each subject's physiological signals of skin conductivity and blood volume pressure in an effort to see if regimes of likely frustration could be automatically discriminated from regimes when all was proceeding smoothly. This pattern recognition approach correctly classified these two regimes 67.4% of the time. Mouse-clicking behavior was also synchronized to frustration-eliciting events, and analyzed, revealing NN distinct patterns of clicking responses

Keywords: Affect, affective computing, user interface, pattern recognition, human-computer interaction, biosensing, emotion physiology.

1 Introduction

Affective computing has been described as "ccomputing that relates to, arises from, or deliberately influences emotions [1]. Affect synthesis and recognition are beginning to drive the ways that researchers think about, and build, interactive computer systems. However, the construction of computational systems that recognize, or in other ways understand, a user's emotional state, is a multidisciplinary undertaking. Researchers from a variety of areas, including psychology, physiology, human-computer interaction, signal processing, and pattern recognition, must work together to unravel the complex questions which arise from inquiry into this new area.

Why build an affective computer? At present, computer systems interact with users in ways that do not allow for the complexities of naturalistic social interaction. Yet recent evidence demonstrates that humans have an inherent tendency to respond to media in ways that are natural and social, mirroring ways that humans respond to one another in social situations [2]. Current systems are impoverished in the options they have for both understanding communication from the user, and communicating to the user. A computer that could decode and produce affective responses would appear significantly improved in its interactive capabilities. This has widespread implications for HCI, ranging from better educational software to computer-mediated communication.

1.1 Our Approach

This paper describes an initial attempt at addressing issues involved with building an affective computing system that uses multiple sensors as input in an effort to infer the user's emotional state. We describe an experimental paradigm that broadly addresses the complex variables of data gathering, signal processing, synchronization and context dependency , which are pertinent to the design of any affectively intelligent system. Specifically, we tried to induce and measure user frustration, by creating a computer game in which the mouse, at random intervals, "failed" to work properly.

It should be mentioned at the outset that this paper places a primary focus on the methodological treatment of these issues, in addition to describing the outcome variables. While we are pleased to report encouraging initial data and analyses, another of our main objectives is to describe what we learned during the process of collecting and making sense of several channels of data. These two concepts –method of data collection and results in recognizing affect from the data– are coupled, but it is important to remember that successful data synchronization and collection does not imply successful affect pattern recognition. The latter is a notoriously difficult problem, highlighted by a longstanding debate in the emotion theory literature about whether or not emotions can even be differentiated by physical responses. Consequently, the results presented here go beyond describing a methodology for gathering data about emotional expression; they also begin to address a larger debate about which physical signals manifest differentiation with emotional state.

One of our most significant contributions is the recommendation of a model of data gathering that can help HCI researchers explore the potential of using multiple sensing technologies. This model should be robust enough to work with various subsets of sensors, be they physiological, non-physiological, or a combination of both.

1.2 Measurement of Emotional Expression

What clues about a user's affective state could one give to a computer? Facial expressions, gestures, and voice are some of the first things that come to mind, especially since they are readily communicated over a distance. Other indices of affect, such as physiological response, tend to be harder to understand, and may require physical contact for sensing. However, in many cases people are in physical contact with computers, sometimes more than they are in physical contact with other people.

Skin-surface sensing may at first seem undesirably obtrusive. Most current-day physiological sensors have a rather clunky interface and dangling wires that can get in the way. However, physiological sensing is gradually moving into devices that people are naturally in physical contact with. Although the sensors used in our initial experiments described below were standard medical sensors placed on the hand, these same sensors have also been comfortably built into jewelry, shoes, clothing, and a mouse [3], [4]. New wearable computing designs are expanding the opportunities for users to be in natural physical contact with the computer.

It is interesting to consider some of the pros and cons of sensing with "highly public" means such as cameras and microphones, vs. with relatively "intimate" means such as skin-surface sensors. Although the former involves no physical contact, and certainly provides an easy-to-understand means of communication, it can also be viewed as an invasion of privacy that is hard for the user to control. The user may want her emotion communicated, but not want her appearance transmitted or her voice recorded. Furthermore, it may be hard for a single user to disable a computer vision or voice recognition system that is built into a "smart room." By building physiological sensors into wearable systems, or embedding them in traditional input devices such as the mouse or keyboard, the user retains primary control. He has the choice of physically removing or disabling the sensors easily and whenever he wants, and he can be assured that these signals do not provide identifying information as would video face recognition or audio speaker-identification systems.

There is mounting evidence suggesting that physiological signals may have characteristic patterns for specific emotional states (see, for example [1], [5]). However, emotion researchers still argue about the definition of emotion and what constitutes an emotional state, so that it is still very hard to compare results of efforts to recognize emotions from physiology. Many researchers eschew the use of categorical labels for emotional states and instead describe emotion by a set of two or more dimensions. The most common two dimensions for describing emotion are arousal (the intensity of feeling), and valence (positive or negative). Using a multi-dimensional description of emotion, Lang and his team have achieved some success with eliciting predictable physiology patterns by exposing subjects to photographs of varying emotional tone [6].

2 Relevant Background: Frustration Theory and Psychophysiology

2.1 Frustration Theory

Frustration theory, studied in the psychology community since the 1930's, has been historically difficult to define. Since frustration was originally conceptualized during the rise of the behavior theorists, much of the work on frustration has involved animal behavior. Not surprisingly, conceptual discussion has therefore focused on the following question: is frustration really a behavior, or is it an emotional response such as anxiety? Lawson describes Rosenzweig's theory of frustration [7] as "the occurrence of an obstacle that prevented the satisfaction of a need." Others have paired frustration with aggression, suggesting that there is an action-and-reaction behavioral loop [7]. In this formulation, the occurrence of frustration always increases the tendency for an organism to respond aggressively, i.e., a rat will increase its vigor when an obstacle is placed between it and its reward.

For our purposes, it makes sense to define frustration as an increase in negative arousal when something uncontrollable impedes the subject's progress toward a goal. This kind of frustration is referred to as unconditioned or primary frustration, in which there is a hypothetical unconditioned reaction to the frustrating event. The immediate consequence of this is a short-term increment in generalized, energizing drive or arousal [7]. Primary frustration, in this view, has an affective or emotional component.

One of the principal independent variables (causes) of frustration has been defined as the delayed reinforcement (reward) of a conditioned response [8]. In a traditional behavioral paradigm, this might be implemented as a delay in delivery of food (reward) after a trained animal presses the correct lever (response). In our experiment, the lever-pressing is analogous to clicking the mouse to advance the screen, and the delivery of food corresponds to screen advancement. Therefore, if we introduce a delay in the game's response to the user's actions, we would expect that the result will be similar to the animal's frustration response. The above concepts are also familiar to the HCI community as issues of immediate feedback and user control [9]. These user-interface guidelines are long-established in the field of HCI, and are part of what are known as principles of Direct Manipulation [10]. Our experimental paradigm exploits purposefully the violation of these guidelines. In a companion paper [Klein et al], it was verified that inserting unwanted delays into the user's task led to significantly more frustration in users compared to a control group performing the same task without the delays. If it is true that users consistently achieve a state of high arousal and negative valence in direct, repeated response to such flouted rules of immediate feedback and control, an added value of work such as ours is to provide yet further confirmation of the theory that these design guidelines are valid and necessary.

2.2 Psychophysiology

Physiological signals such as skin conductivity, heart rate, and muscle tension may provide key information regarding the intensity and quality of an individual's internal experience. These kinds of signals are easily converted to digital format and may eventually be unobtrusively monitored, making them very accessible to pattern recognition techniques. Although debate exists regarding the specificity of signals to particular emotional states, we suggest that psychophysiological data may at least provide information regarding the valence and intensity of the user's internal state, and may be helpful by acting in tandem with computer vision, hearing, and natural language processing to make computers more aware of user affect. Attention to methodological detail is necessary in order to address the complexity and high individual variability in physiological reaction to external and internal events.

Applied psychophysiological research has been defined as "the scientific study of social psychological and behavioral phenomena as related to and revealed through physiological principles and events" [11]. Cacioppo and Tassinary [11], [12] explore the nature of psychophysiological relationships, considering several possibilities of physiological-to-social and physiological-to-behavioral connections: one-to-one, many-to-many, one-to-many, and many-to-one. In the case of our frustration experiment, we have allowed for the many-to-one case, assuming that multiple features of a series of signals will provide the most information about an elicited reaction.

Two physiological signals were chosen for the current experiment, although we do not claim that the two we chose are optimal for measuring frustration. These two signals are the galvanic skin response (GSR) and the blood volume pressure (BVP). We will focus on these two in the rest of this paper for concreteness, but we stress that the methodological principles described here are independent of the specific signals measured.

GSR, also sometimes called galvanic skin conductivity or electrodermal response, has been closely linked to emotion and attention. It is measured by passing a small current through a pair of electrodes placed on the surface of the skin and measuring the conductivity level. Increased arousal potentiates the signal. GSR is highly influenced by frustrative nonreward situations, and has often been used to measure subjects' reactions to a situation or discrete stimulus that elicits anxiety [13].

BVP, also known as peripheral blood flow measurement, and blood volume pulse uses the light absorption characteristics of blood to measure the blood flow through skin capillary beds in the finger (a technique known as photoplethysmography.) Small capillaries such as these tend to contract upon subjects' contact with an anxiety-provoking stimulus, causing the envelope of the signal to "pinch" inwards. The periodic component of this signal can also provide heart rate, which if measured precisely enough, can be used to extract heart-rate variability, which may give clues to valence [14].

3 Methodological Recommendations

One might think that it is easy to build a system that frustrates the user. However, we found that it was quite difficult to build a system that frustrates users in a way that is reliable, repeatable, controllable, and characteristic over a series of individuals. In order to create stimuli that effectively elicited an emotional response of frustration in the user, we looked at a number of possible scenarios, but quickly settled on flouting several established user-interface design guidelines described by Mayhew [9]. Specifically, we built a system that impeded the user's goal to score well in a time-limited visual perception "game," by causing unprovoked delays at seemingly random points.

The following section details the methodological issues we encountered in the process of creating an experiment to elicit frustration. We describe experiment-specific solutions as well as a recommended general principle for each design point:

3.1 Supporting the deception

Director 5.0 features a GUI builder that offers easy-to-build widgets with built-in visual feedback mechanisms. In particular, a button-builder yields a button that, when clicked, provides immediate reverse-video flashing of the button. This experiment, however, required that feature to be disabled. Since we wished to otherwise support direct manipulation in the interface, we chose to change the immediate feedback on button clicks from this standard flashing to simply showing the next puzzle. If the buttons provided reverse-video flashing upon release of the mouse button, users might not believe the deception that the mouse/system were malfunctioning. Removing this feedback, users would have no other clue that the system was not frozen, or that the mouse was not stuck.

Recommendation: Eliciting emotion in the laboratory often involves deception. Interface design should support this goal, although it may include the reversal of established HCI guidelines.

3.2 Adding delays to manage delays

When we first inserted system-freeze delays and tested the system on representative users (college and graduate students who did not know the system was rigged to pause). We found that testers were properly at a loss to account for the system failure, and often responded by repeated, rapid-fire clicking of the mouse on the same or on different buttons. Once the pre-programmed delay ended, however, this rapid-fire clicking would catapult users unintentionally past several subsequent puzzles, until the user realized s/he had regained control of the system. We didn't want users to skip an unknown quantity of puzzles, since it would skew many critical aspects of the experiment. All testers regained control of the system within 600ms. We therefore implemented a one-second delay on the puzzle that immediately followed each freeze-delaying puzzle. Since users invariably took over a second to complete each puzzle and move on to the next, this "echo" delay had the effect of mitigating this catapult behavior, while remaining invisible to the user. Subsequent user testing revealed that this fix was completely effective.

Recommendation: Observe natural user interaction. Predisposed behaviors may require complex and counterintuitive redesigning in order to elicit desired emotional reactions.

3.3 Randomizing delays

We needed to support the deception that the mouse / system was malfunctioning as a matter of random chance. We therefore dealt with the issue of occurrence of the delays by varying delay times, randomizing the occurrence of delays within games, and varying the amount of delays over the three possible games a subject would play (see Figure 1).

Recommendation: Duplicate the variety in real life scenarios as much as possible.

3.4 Synchronization and Context

An important aspect of this study was the realization that sufficiently sensitive instruments can be used in tandem with sophisticated computational media to create the foundation for systems that are able to sense affect in the user. A critical requirement for such a system, though, is timing and contextual knowledge. The system needs to be furnished with much more than physiological signals: It needs detailed, highly accurate information on when those signals were created, and under what circumstances.

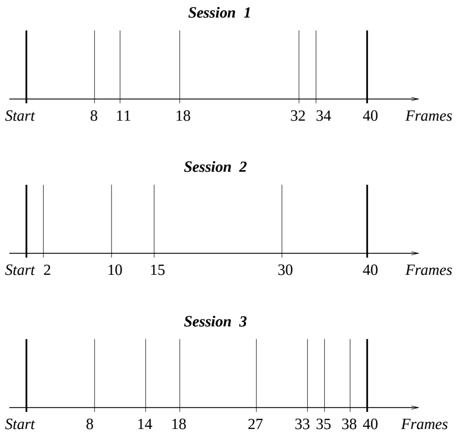

Figure 1: Delay schedule for the three game sessions

To support integrated, millisecond synchronization, a digital clock was hand-built into the game software in Director, and used both to display the time elapsed for the current game to the user, and as a gross index for synchronization with the sensing system. The time was displayed in small type (24-pt., see Figure 4) on the main monitor, and in large (124-pt.) type on the smaller monitor, which faced the video camera. Both displays showed minutes and seconds since the start of the current game. This served to reinforce the time pressure to the subject, as well as to capture the exact time on video for its synchronization.

Macromedia Director 5.0 enables one to write messages to the Message Window in a logfile, which may be shown or hidden at runtime. In this experiment, concurrently with each mouse click, messages were written to a message window, which was hidden from the subject's view. Once a game was completed, the administrator would debrief and excuse the subject, reveal the Message Window, and paste the contents to a text file.

In the logfile, this same timing scheme, in minutes and seconds, was recorded, as was the computer's own clock time at the start of the experiment. To further refine this measurement scheme, however, the logfile also recorded the current number of the system's "ticks" at each mouse click and at other strategic points in the experiment. Ticks are Director's ingrained time-measurement scheme. These ticks occur every 8ms, and are counted from the moment that Director was most recently started. These measures, in tandem, provided the high degree of timing accuracy that is needed to synchronize time-sensitive physiological data with real-world stimuli.

The sole input device with which subjects interacted with the game system was a standard Macintosh mouse that had been modified so that it included a second cable that plugged into the physiological sensing system (described below), and yielded a pulse on each mouse click. Every time that the modified mouse was clicked, it was recorded both as a timed event in the logfile and as a pulse in the sensing system. By modifying the mouse hardware to "talk directly to" the physiological sensing system, behavioral mouse clicks and physiological responses were accurately synchronized.

Since the logfile generated by the game application also recorded contextual information about mouse clicks: correct/incorrect game answer, puzzle number, and occurrence and status of the system delays, this altered mouse yielded a mouse-click record that served as a critical, high-precision synchronization data between stimulus and user response. (See Figure 2).

Recommendation: Multiple data inputs must be very precisely synchronized, which may require creating overlapping events to facilitate their alignment. This may require customized means such as novel hardware modification.

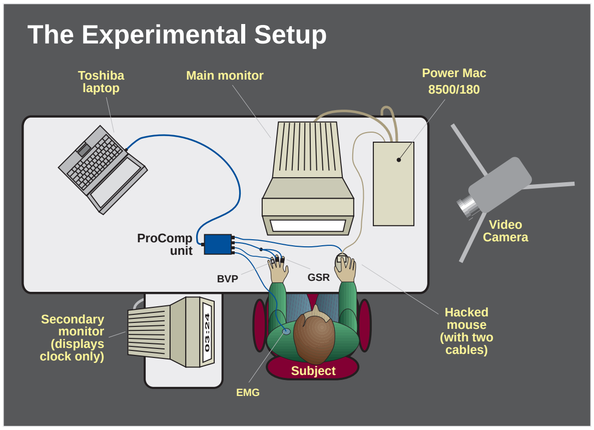

Figure 2: Layout of the experimental setup

4 The Pilot Study

This study was executed with prior approval of MIT's Committee On the Use of Human Experimental Subjects, in accordance with their ethical guidelines on privacy, deception, and subject rights.

4.1 Subjects

Thirty-six undergraduate and graduate students participated in this experiment. They were recruited through flyers posted in various buildings around the MIT campus. They were told that the experiment would last for one hour and they would receive ten dollars for their participation. Subjects were led to believe that their task would be "participation in a visual cognition game", a believable story, given the fact that the experiment took place in the Vision and Modeling group at the MIT Media Laboratory. If subjects were told up front that the goal was to try to frustrate them, then most of them probably would not have gotten frustrated. Consequently, it was necessary to initially deceive the subjects in order to elicit the desired emotional reaction in ways that closely resembled a real-life situation. All subjects were debriefed afterwards as to the true nature of the experiment, and reminded of their rights to have their data withdrawn if they wished.

4.2 Materials

Psychophysiology Sensing System

The sensing system consisted of GSR and BVP sensors attached to the first three fingers of the subject's non-dominant hand. Subjects used their dominant hand for the mouse.

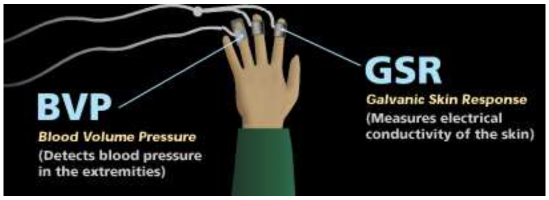
BVP
Blood Volume Pressure
(Detects blood pressure in the extremities)
GSR
Galvanic Skin Response
(Measures electrical conductivity of the skin)

Figure 3: Detail of biosensor placement on the subject. BVP and GSR sensors are placed on the subject's non-dominant hand

The sensors attached via wires to a ProComp Plus analog-to-digital unit. The ProComp Plus, manufactured by Thought Technology, is a multimodality, 8-channel, medically-approved, safe system for monitoring of biosignals, and converts the analog signals into digital form. The ProComp unit was connected through fiber-optic cable and adapter to a Toshiba 110CS Satellite laptop PC computer with a 10-inch color display that was hidden from the subject's view, although in the same room. The laptop computer recorded the signals from the ProComp Plus unit at 20 samples/second, using software designed by Thought Technology, running under DOS.

Game System Hardware and Software

The game system (see Figure 2) consisted of a Power Macintosh 8500/180 with one large, 21" color monitor that displayed the experimental game, and a second 13" color monitor that displayed a large (124 pt.) digital clock.

We designed, built and tested an interactive software game specifically for this experiment using Macromedia Director 5.0 for the Macintosh. We selected Director for its rapid multimedia prototyping capability, and our ability to quickly build and use a game application in the actual experiment. The system development underwent six iterations of design, prototyping, user testing and redesign, over a six-week period. The game consisted of a series of 40 similar visual puzzles, each on a separate screen in modal succession.

Other Equipment

A video camera recorded the subject's upper torso and hands, as well as the elapsed time of the experiment on the smaller monitor, which faced the camera.

4.3 Pilot Study Procedures

Upon responding to the flyers or requests, subjects were scheduled for a one-hour time slot. They were then told the "cover story": that the purpose of the experiment was our interest in how their physiology would react to a series of brightly colored graphics as they interacted with the game. After subjects arrived at the lab, they were asked to read and sign MIT's standard subject's rights forms, and then were ushered into a conference room where the experiment took place. They were then given the game instructions.

The game consisted of a series of puzzles, and the task was to click the mouse on the correct box at the bottom of the screen which corresponded to the items of which there were "the most" on the above array. This advanced the screen to the next puzzle. Subjects received ten dollars for their participation, but the game was also a competition; the individual who received the best overall score and speed at the end of the data collection was told s/he would receive a one hundred dollar prize. This incentive was set up as a way to mimic a real-life situation where users would be racing toward a goal (e.g. meeting a deadline, getting a paper printed out on time, etc.).

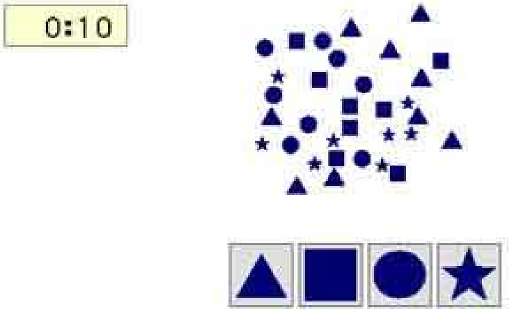

Figure 4: A typical puzzle with clock showing elapsed time

At irregular intervals, a delay occurred during which the mouse appeared not to work properly. If questioned, the experimenter nonchalantly answered "Oh, it sticks sometimes. Please keep going."

4.4 Design Results

The results of the methodology above are that the experiment, which illustrates the four principles above, ran smoothly on 36 subjects, successfully deceiving them, and successfully producing tightly synchronized streams of mouse click behavior, video, physiological signals, and events in the "game". It is worth emphasizing that the specific signals collected are not the emphasis; the current sensing system accommodates 8 channels of data, and we could have easily plugged in sensors besides GSR and BVP, such as respiration, skin temperature, electromyogram, and so forth.

In summary, the design methodology described above was found to be successful for eliciting two episodes: 1. "All is going smoothly," and 2. "The system is not advancing, impeding the user's goal." We now turn to the pattern recognition section of this paper, which examines whether the physiology and behavior of the user showed any distinctive differences during these two episodes.

5 Pattern Recognition

Data analysis of human physiology and behavior is a complex problem. Several factors, both external and internal, shape the output of the sensors. The goal here was to use the physiological data to see if the computer could be taught to identify and discriminate differences between how a user responded when "all was going smoothly" vs. how a user responded when "the system wasn't working properly." We also analyzed the behavioral data from the mouse for cues to different patterns people use in responding to perceived system delays. The video data was not used in the pattern recognition analysis below, but is available for future work.

5.1 Physiological Data Modeling

In choosing a model that adequately captures the behavior of the physiological signals, we need to consider the dynamic or time-evolving nature of the signals. Also, in order to make these models robust to variations that are hard to predict, or too complex to model, it behooves us to consider models with an underlying probabilistic framework. One of the most successful techniques, which has received much attention in the probabilistic literature of dynamic systems, is that of Hidden Markov Models (HMMs). HMMs have been successfully used to model time series like speech, and are currently in use in speaker-dependent and independent speech recognition systems.

An HMM is a finite-state model with a fixed number of internal states including an initial and a final state. An HMM is fully defined by a set of probability density functions which describe the possible outcomes associated with each state, and a set of transition probabilities which describe the likelihood of transitioning between any two given states of the HMM. As one traverses an HMM from its initial to its final state, a time series is generated according to the states visited and the probability density functions sampled from at each time step. This description of an HMM treats it as a generative tool; that is to say, if we know the structure of the HMM, we can use it to create sample time series. In practice, however, we shall use the HMM model to make inferences. From a set of observables (extracted from physiological time series) we will attempt to reconstruct the structure of an HMM that could have generated these time series with very high likelihood. Once the structure of the HMM is known, we can use the model to further classify more data; this may be done by fitting an HMM to each of the classes of interest, and subsequently using these as competing experts trying to classify an unlabeled time series. The classification is then chosen from the expert that assigns the highest likelihood to the data. There are well known algorithms for estimating the parameters (consisting of the inter-state transition probabilities and the state probability density functions) of an HMM (Baum-Welch estimation algorithm) [15]; this is the core of the training stage. In the testing phase, or decoding, the goal is to assign a label to different portions of the time series. Since each label is associated with a different HMM, this problem consists of finding the points in the time series where there's a transition from the final state of one HMM to the initial state of another. Viterbi decoding [15] is a dynamic programming algorithm that allows recovery of the state sequence of a series of HMM, and hence provides the desired parsing of the time series.

The Baum-Welch algorithm is a maximum-likelihood procedure to estimate the parameters of an HMM with a given number of states; one must therefore fix a-priori the number of states of an HMM before doing the training. The form of the probability density functions must also be established prior to training. One of the most common parameterized forms for density estimation consists of expressing a probability density function as a finite mixture of Gaussian components; furthermore, one may consider the Gaussians in the mixture to have diagonal or full covariance matrices. Finally, one may specify a-priori the topology of an HMM by constraining some of the transition probabilities between states. As a special case of this, we obtain the causal or left-to-right topology; that is, an HMM in which one may only visit states that have not yet been visited. This structure can sometimes be useful if the data follows a non-recurrent sequential pattern.

Finding an optimal structure can sometimes be a difficult problem, so we have opted for a simple approach to selecting a structure, namely, to select a subset of structures, train for each, and then evaluate the performance for each one of them for each subject in the data set. We considered a subset of HMM structures by varying the number of states (between 4 and 7 states); the number of Gaussians in the mixtures of the densities (1 or 2 mixtures); the form of the covariance matrices (diagonal or full); and the topology of the HMM (left-to-right or fully connected). Since the objective is to find a possibly user-dependent structure, we have to treat the 32 possible combinations that result by varying the parameters above for each one of the subjects.

One of most important issues is how to obtain a set of features from the raw data (GSR and BVP) that might have correlates with internal affective states. This is still an open research question: the mappings between affective and physiological states is being investigated at large in the psychophysiology community. In deciding on a feature set, we should account for classical measures 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 much attention in other studies. We have proposed the following set of five features from the raw data (for more details see [16]:
i. the GSR signal detrended by subtracting a time-varying sample mean (found with a moving 10-sec window).
ii. a local time-varying unbiased sample variance of the signal in (i) (found with a moving 10-sec. window).
iii. the "pinch" of the BVP signal (or difference between the upper and lower envelope of the signal) (see Figure 5).
iv. the variation (first difference) of the peak-to-peak interval of the BVP signal.
v. the local variance of the detail coefficients in a 3-level wavelet expansion of the BVP signal.

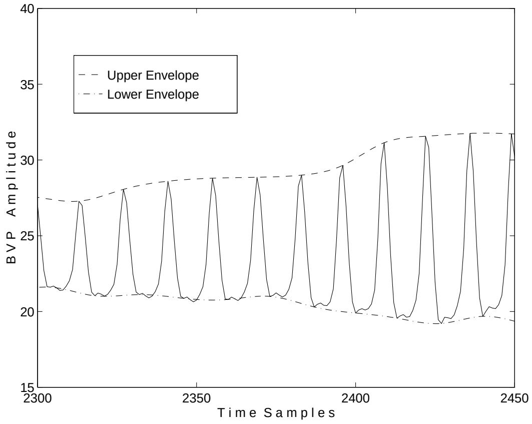

Figure 5: Example of a BVP signal

The GSR signal is a non-stationary signal which varies its baseline unpredictably across an experimental session. For this reason, the GSR features extracted on (i) and (ii) remove this time variation and examine the local amplitude and variance of the signal. The BVP signal is a much richer signal that allows us to exploit its harmonic structure for feature extraction. In (iii) we extract how much the amplitude of the signal constricts or expands. Because the BVP is correlated with heart rate, we have extracted a measure of heart-rate variability by measuring the variation in peak-to-peak intervals. Finally, we have included a different measure of frequency variation over time by doing a wavelet expansion of the signal and analyzing the local variance of the wavelet coefficients (over a 1.5 second window). Because the time series obtained in (iv) and (v) are sparser than the original time series, these values have been interpolated to obtain time series of equal length, which we can then stack in a 5-dimensional feature vector.

5.2 Establishing a Ground Truth

We wish to treat the data analysis as a classification problem and determine whether we can characterize and predict possible instants of frustration from a set of observed physiological readings. Before proceeding, a ground truth needs to be established in order to test the classifications. This is a non-trivial problem which deserves careful consideration since the class categorizations we shall use to label the data have only been induced, not firmly established. In other words, there is an uncertainty associated with the class to which the data belongs. There is, for instance, a possibility that a stimulus failed to induce a frustration response, and conversely, that a subject showed a frustration response in the absence of the controlled stimulus due to another uncontrolled stimulus, such as a cognitive event. It was not possible to stop and ask the subject for confirmation at each instant, as that would have disturbed the experiment. Furthermore, self-report data on negative emotions is notoriously variable, depending on many factors unrelated to the present feelings of the subject. Consequently, we cannot claim that the two episodes we distinguish truly correspond to frustration and to non-frustration; all we can say is that things were going smoothly or not, and that there was or wasn't a difference in the person's physiology or behavior as detected by the applied models.

In the 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. We do not wish to abandon this framework entirely and will adopt a deterministic rule to label the training examples. However, establishing a proper labeling for the training data is one of the aspects of this problem that should be adaptive and subject to further discussion.

Our only degree of belief about what class the data belongs to is given by the onset of the pre-controlled stimuli during the course of the experiment. A rather intuitive approach to define the classes is to consider the response following a stimulus as representative of a frustration episode. How we establish the temporal segmentation following a stimulus deserves some attention. The time window we use to capture this response has to be wide enough to allow a latency period, as well as the true physiological response due to the stimulus. The latency period consists of the time lag that elapses between the onset of the stimulus and the start of the physiological change due to the stimulus. Some authors have established that for galvanic skin response this delay can be as much as 3 seconds [17]. The following diagram illustrates the principle used to label the data portion between any two stimuli:

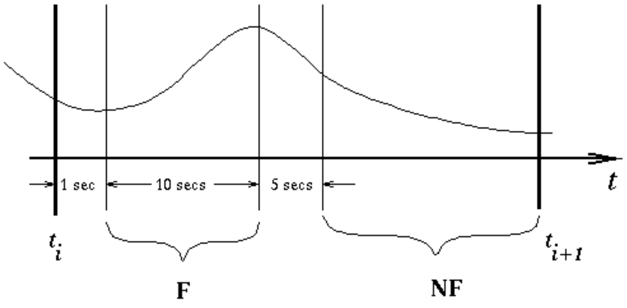

Figure 6: Ground truth labelling

This figure shows a portion of a mythical signal between two stimuli (corresponding to instances when the mouse failed to work) represented by the bold vertical bars. Following the onset of one stimulus, we allow a dormant period of 1 second to pass before we start assigning the labels; then we window the following 10 seconds of data as representative samples of the class we want to model as frustration (F). In order to transition out of this class, since the model boundaries are not known with precision, we allow another dormant period (of 5 seconds) without any classification, and then consider the rest of the signal up until the next stimulus to correspond to the class of non-frustration (NF). If the remaining set of samples is less than a minimum number of samples required to assign a label (3 seconds in these simulations), then a label is not assigned to this region. If the time windows used on two adjacent stimuli overlapped (the stimuli were spaced out by less than 11 seconds,) then the two resulting segments of data labeled as F would be merged together.

The chosen labels may be viewed as corresponding to positive and negative examples of the phenomenon we want to model. The reader should bear in mind, however, that this is a simplified mnemonic and modeling device and not an argument for what the true state of the physiology is since we can safely assume that human physiology exhibits widely complex modes of behavior. The labeled regions roughly correspond to areas in which we have a higher degree of confidence about the class induced, whereas the unlabeled regions represent "don't-care" regions where our knowledge of the transition between affective states is too poor to include in the ground truth.

5.3 Evaluation and Discussion

We divided the experimental sessions for each subject into a training and a testing set. The results reported in this paper apply to 24 subjects that had sufficient experimental data (corresponding to 2 or 3 sessions); it was found that 12 subjects with only 1 session did not have enough experimental data to yield significant results. For the 11 subjects with 2 sessions, one session was randomly selected for training and the other for testing. For the remaining 13 subjects who had 3 sessions, two were used as training data and the remaining one as testing (the testing session was selected randomly as the second or third session).

After training each HMM structure reported above for a total of 24 subjects, the training and testing data were parsed (segmented into regimes labeled as F or NF) using Viterbi decoding. To evaluate the performance of the system, we calculated the percentage of data samples that had been correctly classified (this evaluation criterion, of course, only applies to labeled samples; the "don't-care" regions described above are left out of the evaluation), and the HMM that performed the best on the testing set for each subject was chosen. The percentage of properly classified data samples is a measure of how much the two parsings (the one determined by the ground truth labeling rules, and the one outputted by the system) agree, and is therefore a measure of the performance of the system.

The histograms below show the distribution of the overall recognition rates for all subjects, as well as the distribution of the recognition rates for the individual categories (F and NF). The height of each bar is proportional to the number of subjects for whom the system attained the accuracy shown on the horizontal axis. As might be expected for the task at hand, these histograms clearly show that performance is subject dependent.

The performance to beat was that of a random classifier which outputs a decision (F or NF) on every data point with equal chance (random guessing, therefore, is 50%). It should be noted, however, that a fairer assessment of the performance of a system of this kind would take into account prior knowledge about the likelihood of occurrence and duration of each label, which is likely to change as a function of personality, time-pressure, etc. Because of the nature of the experiment, each subject spent a variable amount of time on each experimental session. However, in the ground truth, the duration of each frustration episode was held constant, in accord with the labeling rules described above. Consequently, the number of frustration episodes and the time spent in each could vary across subjects. This perhaps offers suggestions to design alternative ground truth labeling for future re-modeling work in this area by taking into account the length of time that each subject invested in the experiment and adapting the length of the frustration episodes accordingly.

The overall (F and NF combined) performance for the training set was significantly better than random for all 24 subjects (the mean value of the recognition rate was 81.87%). For the testing set, performance was significantly better than random for 22 of the 24 subjects (the mean value of the overall recognition rate was 67.40% and was 70.93% for the 22 subjects who achieved rates better than random). The histogram in Figure 8 shows the disparity between the recognition rates for the F and NF labels of the test set. This may reflect the uncertainty we have in the ground truth of these data.

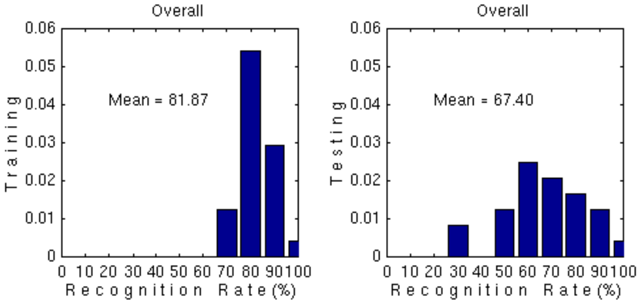

Figure 7: Histogram of Overall Recognition Rates (Training and Testing Sets)

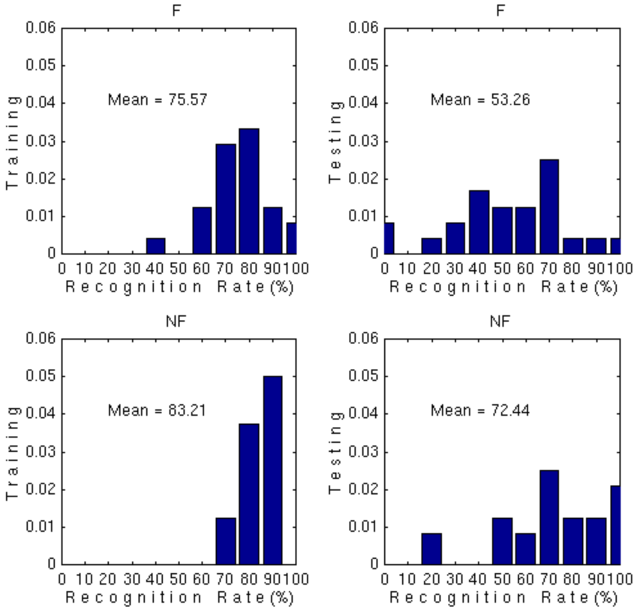

Figure 8: Histogram of Recognition Rates for F and NF labels (Training and Testing Sets)

5.4 Characterizing mouse-clicking behavior

The methodology used in this experiment's design also allowed us to look at a behavior variable. We examined the mouse-clicking behavior of the user each time there was a potential "frustration-elicitor", i.e., when the computer went into a delay mode and the user could not advance in the game. Specifically, we computed the number of mouse clicks following each such event, and plotted the fit distributions to these data (shown in Figure 9). We expected that some subjects would be very "passive" showing few or no extra clicks, whereas some subjects would show a large number of clicks in response to the delay stimuli. High-density click patterns did not always occur in response to a problem, but they never occurred unless there was a perceived problem. (When there was no problem with the system, the mouse click always advanced the game properly.)

We clustered the data sets of click behavior obtained from the 24 subjects to examine whether similar patterns of behavior emerged. Assuming an underlying Poisson distribution governing each cluster, we fit a number of clusters, ranging from 3 to 5, using an iterative K-means algorithm. Using this approach, we obtained 4 distinct clusters for the entire data set (we found that increasing the number of clusters beyond 4 only yielded empty clusters). The Poisson distributions and the number of members fit to each cluster are shown in Figure 9 below. The horizontal axis of the cluster distributions represent the number of clicks, and the vertical axis represents the probability of that number of clicks being made by a user.

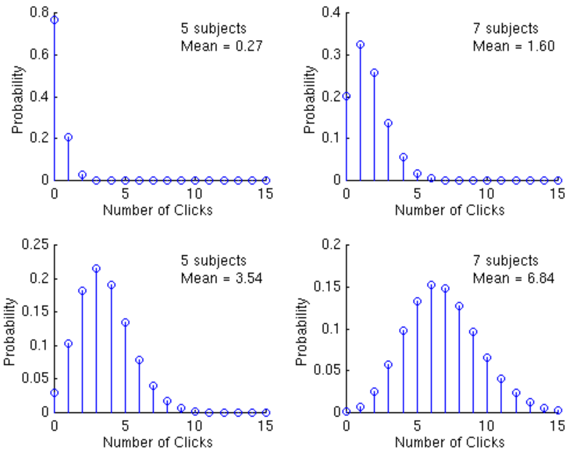

Figure 9: Poisson distributions for each cluster, illustrating four distinct patterns of mouse-clicking behavior when the system appeared to be stuck

5.5 Discussion

The four patterns found by clustering reveal four types of behavioral responses to the problem. The upper-left panel, for instance, indicates a type of person who usually just waited without clicking, occasionally clicked one extra time, and rarely clicked more than that. As we move to the upper-right and lower-left panel, we see this behavior shifting to a higher number of clicks. Finally the lower-left panel represents a cluster of users who always clicked, and usually clicked a large number of times.

The results obtained from the behavioral measure extracted in this study suggest that for 19 of our 24 subjects (i.e. all but the subjects in the first cluster), a system that can discern extraneous mouse clicks may use this data to draw inferences based not just on physiology, but on behavioral cues as well. This approach may require, however, that the system have precisely-timed awareness of its own behavior, a kind of rudimentary "self-awareness," so that it can sense things like delays followed by "catapulting forward" and other patterns, e.g., repeated typing of something erroneous, which may indicate that it is causing a problem for the user. Also, the current system set-up could be augmented to measure other forms of physical interaction, including the intensity and direction of pressure exerted by the user on the mouse. The mouse-clicking patterns found here are just one example of a potentially useful behavioral variable that may give clues to a user's affective state.

6 Conclusions and Future Directions

This paper has described an experimental methodology for eliciting events likely to lead to user frustration, and for successfully gathering and synchronizing precise physiological, behavioral, visual and operational data. Four general methodological principles were proposed and illustrated with a specific experimental design. This design was then used successfully to gather accurately synchronized data from twenty-four subjects. We analyzed the physiological and behavioral data gathered, proposing new features for extraction from the physiological portion of this data, and developing an automatic technique for classifying the features using HMM's. The resulting classification was significantly better than random for 22 out of 24 subjects, suggesting that there is some important discriminating information in the two physiological signals of GSR and BVP, although this discrimination is far from perfect. We also found four classes of mouse-clicking patterns exhibited by users when the system did not advance to the next screen on the first click. Both the physiological and mouse-clicking patterns pointed to user-dependent behaviors, but ones which a machine could nonetheless begin to model and learn to recognize.

In future experiments or applications, the specific kinds of signals collected can be expected to differ for different goals. Here we used skin conductivity and blood volume pressure, while another implementation might use heart rate variability and muscle tension. We may in fact discover at some point that other sensors are more ideal in the current experiment than the ones we actually used. The key guiding principles presented in this work, however, are invariant to the specific physiological signals measured. Means of precise synchronization and links to external and behavioral context are the key contributions of the methodology presented here for making use of physiological information.

Even in an ideal affective computing system, we envision that user responses will not always be unambiguous, and that in some cases recognition system may need to prompt the user for subjective input, for continual reevaluation of ground truth. This prompting will also need to be carefully orchestrated, so that it is sensitively conducted in a way that does not increase the user's frustration. Over time, we expect that a system could "get to know" an individual's patterns of frustration, and correlate these with things that it is doing, which might be responsible for the frustration. Although it would not necessarily be able to deduce causation on its own, with a little more input on the part of the user such deductions might be possible. It might occasionally be proactive and ask the user something like: "Would you prefer that this system's behavior X go away?" Overall, information regarding which system functions are most correlated with episodes of user frustration-like responses could be extremely valuable for human-computer interaction designers, providing them with a "continuous human factors" analysis, not just before a product is released, but while it is out in the field being used.

Until the correct combination of physiological and behavioral signals becomes apparent for recognizing a state such as user frustration, there needs to be a lot more focus on specific pattern recognition techniques. A logical next step for us or other researchers would be to repeat this experiment, using the same methodology, but varying the situations to induce a broader range of emotional responses. For example, we could run the same game, but instead of injecting likely frustration-eliciting stimuli, we could inject likely joy-eliciting stimuli, such as the computer game adding extra points to the user's score or the computer presenting the user with sincere-sounding praise for something the user did. The system described in this paper is designed to be extensible to such future inquiries.

Although we collected up to three different data sets from each subject, a second goal is to take a more detailed look at individual response in a longitudinal design; gathering a larger amount of signals from single subjects over a series of repeated observations, especially over many days. In related work on recognizing emotional expression in physiology, it has been observed that there can be more difference in how the same emotion is expressed on different days, then there is in how different emotions are expressed on the same day [18].

Ideally, an affect-recognizing computer should be able to use the information it gains from the user to enhance the computer-human interaction. If a system recognizes that the user is experiencing distress, it might act to ameliorate that stress, or simply monitor it and make an internal note associating one of the system behaviors with a probability that that behavior is frustrating. In a companion submission, Klein et al describe alternate responses that a computer agent might use to try to help a user in reducing frustration that arises in a human-computer interaction [Klein]. Whatever the strategy, the system will probably work best once it learns the individual preferences of its user, including possibly characteristics of the user's personality.

Eventually, we hope to address complex affective data sets collected from the naturalistic situations occurring outside the laboratory. This can be done by porting the paradigm presented here to wearable computing systems, equipped not just with sensors for a person's emotional expression, but also equipped with sensors to discern information about the situation the person is in. In sum, we suggest that the methodology presented here has many applications outside the specific experiment described in this paper. It addresses the design issues involved in the simultaneous monitoring of several input devices, while also providing data for subsequent pattern analysis, all within the context of trying to learn more about characterizing a user's affective response.

Our broader goal still echoes Winograd's [19] view that we must perform experiments in which we pay close attention to the entire "user experience." We have emphasized that a critical part of this experience involves emotion, and that an affective computer would respect this by trying to recognize and respond appropriately to the emotion. Although there is still a lot to be investigated, including real-time accurate recognition of user signals, improvement of sensor selection, exploratory analyses of more behavioral variables, and improvement of machine awareness of situations, we think that the approach presented here offers a significant first step toward the development of computers that not only pay close attention to user experience, but begin to recognize and respond to the affective qualities that people naturally bring to a human-computer interaction.

References

[1] R.W. Picard. *Affective Computing.* M.I.T. Press, Cambridge, MA, 1997.

[2] B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television, and New Media LIke Real People and Places.* Cambridge University Press, New York, 1996.

[3] R.W. Picard and J. Healey. Affective wearables. *Personal Technologies*, 1(4):231–240, 1997.

[4] T. Marrin and R.W. Picard. The ”conductor’s jacket”: A device for recording expressive musical gestures. In *Proc. Intnl. Computer Music Conf.*, December 1998.

[5] R.W. Levenson P. Ekman and W.V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221:1208–1209, 1983.

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

[7] R. Lawson. *Frustration: The Development of a Scientific Concept.* The MacMillan Company, New York, 1965.

[8] A. Amsel. *Frustration Theory.* The Cambridge University Press, Canada, 1992.

[9] D.J. Mayhew. *Principles and Guidelines in Software User Interface Design.* Englewood Cliffs, New Jersey, 1992.

[10] B. Schneiderman. *Designing the User Interface: Strategies for Effective Human Computer Interaction.* Adsison-Wesley, Reading, MA, 1986.

[11] J. Cacioppo and L. Tassinary. *Principles of Psychophysiology: Physical, Social and Inferential Elements*, chapter Psychophysiology and Psychophysiological Inference. Cambridge University Press, Cambridge, England, 1990.

[12] J. Cacioppo and L. Tassinary. *Principles of Psychophysiology: Physical, Social and Inferential Elements*, chapter The Skeletomotor System. Cambridge University Press, Cambridge, England, 1990.

[13] A. Schell M. Dawson and D. Fillion. *Principles of Psychophysiology: Physical, Social and Inferential Elements*, chapter The Electrodermal System. Cambridge University Press, Cambridge, England, 1990.

[14] J. Papillo and D. Shapiro. *Principles of Psychophysiology: Physical, Social and Inferential Elements*, chapter The Cardiovascular System. Cambridge University Press, Cambridge, England, 1990.

[15] L.R. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–16, January 1986.

[16] R. 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, EECS, 1997.

[17] M. Helander. Applicability of driver's electrodermal response to the design of the traffic environment. *Journal of Applied Psychology*, 63(4):481–488, 1978.

[18] R.W. Picard E. Vyzas. Affective pattern classification. In *AAAI Fall Symposium Series. Emotional and Intelligent: The Tangled Knot of Cognition*, Orlando, FL, October 23-25 98.

[19] L. De Young T. Winograd, J. Bennet and B. Hartfield, editors. *Bringing Design to Software*. Addison-Wesley, Reading, MA, 1996.