M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 536
To appear in IEEE Transactions on Pattern Analysis and Machine Intelligence

Toward Machine Emotional Intelligence: Analysis of Affective Physiological State

Rosalind W. Picard, Elias Vyzas, and Jennifer Healey

MIT Media Laboratory
20 Ames St.
Cambridge, MA 01239

Abstract

The ability to recognize emotion is one of the hallmarks of emotional intelligence, an aspect of human intelligence that has been argued to be even more important than mathematical and verbal intelligences. This paper proposes that machine intelligence needs to include emotional intelligence and demonstrates results toward this goal: developing a machine's ability to recognize human affective state given four physiological signals. We describe difficult issues unique to obtaining reliable affective data, and collect a large set of data from a subject trying to elicit and experience each of eight emotional states, daily, over multiple weeks. This paper presents and compares multiple algorithms for feature-based recognition of emotional state from this data. We analyze four physiological signals that exhibit problematic day-to-day variations: the features of different emotions on the same day tend to cluster more tightly than do the features of the same emotion on different days. To handle the daily variations, we propose new features and algorithms and compare their performance. We find that the technique of seeding a Fisher Projection with the results of Sequential Floating Forward Search improves the performance of the Fisher Projection, and provides the highest recognition rates reported to date for classification of affect from physiology: 81% recognition accuracy on eight classes of emotion, including neutral.

1 Introduction

It is easy to think of emotion as a luxury, something that is unnecessary for basic intelligent functioning and difficult to encode in a computer program; therefore, why bother giving emotional abilities to machines? Recently, a constellation of findings, from neuroscience, psychology, and cognitive science, suggests that emotion plays surprising critical roles in rational and intelligent behavior. Most people already know that too much emotion is bad for rational thinking; much less well known is that neuroscience studies of patients who essentially have their emotions disconnected reveal that those patients have strong impairments in intelligent day-to-day functioning, suggesting that too little emotion can impair rational thinking and behavior [1]. Apparently, emotion interacts with thinking in ways that are non-obvious but important for intelligent functioning. Emotion-processing brain regions have also been found to perform pattern recognition before the incoming signals arrive at the cortex: a rat can be taught to fear a tone even when its auditory cortex is removed, and similar emotion-oriented processing is believed to take place in human vision and audition [2].

Scientists have amassed evidence that emotional skills are a basic component of intelligence, especially for learning preferences and adapting to what is important [3] [4]. With increasing deployment of adaptive computer systems, e.g., software agents and video retrieval systems that learn from users, the ability to sense and respond appropriately to user affective feedback is of growing importance. Emotional intelligence consists of the ability *to recognize, express, and have emotions*, coupled with the ability *to regulate these emotions, harness them for constructive purposes, and skillfully handle the emotions of others.* The skills of emotional intelligence have been argued to be a better predictor than IQ for measuring aspects of success in life [4].

Machines may never need all of the emotional skills that people need; however, there is evidence that machines will require at least some of these skills to appear intelligent in interacting with people. A relevant theory is that of Reeves and Nass at Stanford: human-computer interaction is inherently natural and social, following the basics of human-human interaction [5]. For example, if a piece of technology talks to you but never listens to you, then it is likely to annoy you, analogous to the situation where a human talks to you but never listens to you. Nass and Reeves have conducted dozens of experiments of classical human-human interaction, taking out one of the humans and putting in a computer, and finding that the basic human-human results still hold.

Recognizing affective feedback is important for intelligent human-computer interaction. Consider a machine learning algorithm that has to decide when to interrupt the user. A human learns this by watching how you respond when you are interrupted in different situations: did you receive the interruption showing a neutral, positive, or negative response? Without such regard for your response, the human may be seen as disrespectful, irritating, and unintelligent. One can predict a similar response toward computers that interrupt users oblivious to their positive or negative expressions. The computer could potentially appear more intelligent by recognizing and appropriately adapting to the user's emotion response.

Although not all computers will need emotional skills, those that interact with and adapt to humans in real-time are likely to be perceived as more intelligent if given such skills. The rest of this paper focuses on giving a machine one of the key skills of emotional intelligence: the ability to recognize emotional information expressed by a person.

1.1 Human emotion recognition

Human newborns show signs of recognizing affective expressions such as approval and disapproval long before they acquire language. Affect recognition is believed to play an important role in learning and in developing the ability to attend to what is important, and is likely a key part of the difference between normal child development and development of autistic children, who typically have impaired affect recognition [6]. For example, instead of attending to the parent's speech with exaggerated inflection, the autistic child might tune in to an unimportant sound, missing the guidance provided by the affective cues.

Emotion modulates almost all modes of human communication – word choice, tone of voice, facial expression, gestural behaviors, posture, skin temperature and clamminess, respiration, muscle tension, and more. Emotions can significantly change the message: sometimes it is not *what* was said that was most important, but *how* it was said. Faces tend to be the most visible form of emotion communication, but they are also the most easily controlled in response to different social situations when compared to the voice and other modes of expression. Affect recognition is most likely to be accurate when it combines: (1) multiple kinds of signals from the user with (2) information about the user's context, situation, goals, and preferences. A combination of low-level pattern recognition, high-level reasoning and natural language processing is likely to provide the best emotion inference [7].

How well will a computer have to recognize human emotional state to appear intelligent? Note that no human can perfectly recognize your innermost emotions, and sometimes people cannot even recognize their own emotions. No known mode of affect communication is lossless; some aspects of internal feelings remain private, especially if you wish them to be that way or if you sufficiently disguise them. What is available to an external recognizer is what can be observed and reasoned about, and this always with some uncertainty. Nonetheless, people recognize each other's emotions well enough to communicate useful feedback. Our aim is to give computers recognition abilities similar to those that people have.

1.2 Importance of physiological emotion recognition

When designing intelligent machine interfaces, why not focus on facial and vocal communication – aren't these the modes that people rely upon? There are cases where such modes will be preferable, as well as other behavior-based modes, such as gestural activity or time to complete a task. However, it is a mistake to think of physiology as something that people do not naturally recognize. A stranger shaking your hand can feel its clamminess (related to skin conductivity); a friend leaning next to you may sense your heart pounding; students can hear changes in a professor's respiration that give clues to stress; ultimately it is muscle tension in the face that gives rise to facial expressions. People read many physiological signals of emotion.

Physiological pattern recognition of emotion has important applications in medicine, entertainment, and human-computer interaction. Affective states of depression, anxiety, and chronic anger have been shown to impede the work of the immune system, making people more vulnerable to viral infections, and slowing healing from surgery or disease (See Chap 11 of reference [4].) Physiological pattern recognition can potentially aid in assessing and quantifying stress, anger, and other emotions that influence health. Certain human physiological patterns show characteristic responses to music and other forms of entertainment: e.g., skin conductivity tends to climb as a piece of music "peps you up" and fall as it "calms you down." This principle was utilized in a wearable "affective DJ" to allow more personalized music selections than the one-size-fits-all approach of a disc jockey [8]. Changes in physiological signals can also be examined for signs of stress arising *while users interact* with technology, helping detect where the product causes unnecessary irritation or frustration, without having to interrupt the user or record her appearance. This is a new area for pattern recognition research: detecting when products cause user stress or aggravation, to help developers target areas for re-design and improvement.

Physiological sensing is sometimes considered invasive because it involves physical contact with the person. However, not only is technology improving with conductive rubber and fabric electrodes that are wearable, washable, and able to be incorporated in clothes and accessories [9] but also there are new forms of non-contact physiological sensing on the horizon. In some cases, physiological sensors are perceived as less invasive than alternatives, such as video. Video almost always communicates identity, appearance, and behavior, on top of emotional information. Students engaged in distance learning may wish to communicate to the lecturer that they are furrowing their brow in confusion or puzzlement, but not have the lecturer know their identity. They might not object to having a small amount of muscle tension anonymously transmitted, whereas they may object to having their appearance communicated.

New wearable computers facilitate different forms of sensing than traditional computers. Wearables often afford natural contact with the surface of the skin; however, they do not easily afford having a camera pointed at the user's face. (It can be done with a snugly fitted hat and stiff brim to mount the camera for viewing the wearer's face, a form factor that can be awkward both physically and socially.) In wearable systems, physiological sensing may be set up so that it involves no visible or heavy awkward supporting mechanisms; in this case, the physiological sensors may be less cumbersome than video sensors.

1.3 Related research

The affect recognition problem is a hard one when you look at the few benchmarks which exist. In general, people can recognize an emotional expression in neutral-content speech with about 60% accuracy, choosing from among about six different affective labels [10]. Computer algorithms match or slightly beat this accuracy, e.g., [11], [12]. Note that computer speech recognition that works at about 90% accuracy on neutrally-spoken speech tends to drop to 50-60% accuracy on emotional speech [13]. Improved handling of emotion in speech is important for recognizing what is said as well as how it was said.

Facial expression recognition is easier for people, e.g., 70-98% accurate on six categories of facial expressions exhibited by actors [14], and the rates computers obtain range from 80-98% accuracy when recognizing 5-7 classes of emotional expression on groups of 8-32 people [15], [16]. Other research has focused not so much on recognizing a few categories of emotional expressions but on recognizing specific facial actions—the fundamental muscle movements that comprise Paul Ekman's Facial Action Coding System—which can be combined to describe all facial expressions. Recognizers have already been built for a handful of the facial actions [17] [18] [19] [20] and the automated recognizers have been shown to perform comparably to humans trained in recognizing facial actions [18]. These facial actions are essentially *facial phonemes*, which can be assembled to form facial expressions. There are also recent efforts that indicate that combining audio and video signals for emotion recognition can give improved results [21] [22] [23].

Although the progress in facial, vocal, and combined facial/vocal expression recognition is promising, all of the results above are on presegmented data of a small set of sometimes exaggerated expressions, or on a small subset of hand-marked singly-occurring facial actions. The state of the art in affect recognition is similar to that of speech recognition several decades ago when the computer could classify the carefully articulated digits, "0, 1, 2, ..., 9," spoken with pauses in between, but could not accurately detect these digits in the many ways they are spoken in larger continuous conversations.

Emotion recognition research is also hard because understanding emotion is hard; after over a century of research, emotion theorists still do not agree upon what emotions are and how they are communicated. One of the big questions in emotion theory is whether distinct physiological patterns accompany each emotion [24]. The physiological muscle movements comprising what looks to an outsider to be a facial expression may not always correspond to a real underlying emotional state. Emotion consists of more than its outward physical expression; it also consists of internal feelings and thoughts, as well as other internal processes of which the person having the emotion may not be aware.

The relation between internal bodily feelings and externally observable expression is still an open research area, with a history of controversy. Historically, William James was the major proponent of emotion as an experience of bodily changes, such as your heart pounding or your hands perspiring [25]. This view was challenged by Cannon [26] and again by Schachter and Singer who argued that the experience of physiological changes was not sufficient to discriminate emotions. Schachter and Singer's experiments showed that if a bodily arousal state was induced, then subjects could be put into two distinct moods simply by being put in two different situations. They argued that physiological responses such as sweaty palms and a rapid heart beat inform our brain that we are aroused, and that then the brain must appraise the situation we are in before it can label the state with an emotion such as fear or love [27].

Since the classic work of Schachter and Singer, there has been a debate about whether or not emotions are accompanied by specific physiological changes other than simply arousal level. Ekman *et al.* [28] and Winton *et al.* [29] provided some of the first findings showing significant differences in autonomic nervous system signals according to a small number of emotional categories or dimensions, but there was no exploration of automated classification. Fridlund and Izard [30] appear to have been the first to apply pattern recognition (linear discriminants) to classification of emotion from physiological features, attaining rates of 38%-51% accuracy (via cross-validation) on subject-dependent classification of four different facial expressions (happy, sad, anger, fear) given four facial electromyogram signals. Although there are over a dozen published efforts aimed at finding physiological correlates when examining small sets of emotions (from 2-7 emotions according to a recent overview [31]), most have focused on t-test or analysis of variance comparisons, combining data over many subjects, where each was measured for a relatively small amount of time (seconds or minutes). Relatively few of the studies have included neutral control states where the subject relaxed and passed time feeling no specific emotion and none to our knowledge have collected data from a person repeatedly, over many weeks, where disparate sources of noise enter the data. Few efforts beyond Fridlund's have employed linear discriminants and we know of none that have applied more sophisticated pattern recognition to physiological features.

The work in this paper is novel in trying to classify physiological patterns for a set of eight emotions (including neutral), in applying pattern recognition techniques beyond that of simple discriminants to the problem (we use new features, feature selection, spatial transformations of features, and combinations of these methods), and in focusing on "felt" emotions of a single subject gathered over sessions spanning many weeks. The results we obtain are also independent of psychological debates on the universality of emotion categories [32], focusing instead on user-defined emotion categories.

The contributions of this paper include not only a new means for pattern analysis of affective states from physiology, but also the finding of significant classification rates from physiological patterns corresponding to eight affective states measured from a subject over many weeks of data. Our results also reveal significant discrimination among both most commonly described dimensions of emotion: valence and arousal. We show that the day-to-day variations in physiological signals are large, even when the same emotion is expressed, and that this effect undermines recognition accuracy if it is not appropriately handled. This paper proposes and compares techniques for handling day-to-day variation and presents new results in affect recognition based on physiology. The results lie between the rates obtained for expression recognition from vocal and facial features, and are the highest reported to date for classifying eight emotional states given physiological patterns.

2 Gathering Good Affective Data

In computer vision or speech recognition, it has become easy to gather meaningful data; framegrabbers, microphones, cameras, and digitizers are reliable, easy to use, and the integrity of the data can be seen or heard by non-specialists; however, non-specialists do not usually know what comprises a good physiological signal. Although people recognize emotional information from physiology, it is not natural to do so by looking at 1-D signal waveforms. Not only does it take effort to learn what a good signal is, but the sensing systems (A/D converters and buffering-data capture systems) for physiology do not seem to be as reliable as those for video and audio. Factors such as whether or not the subject just washed her hands, how much gel she applied under an electrode, motion artifacts, and precisely where the sensor was placed, all affect the readings. These are some of the technical factors that contribute to the difficulty in gathering accurate physiological data.

Although dealing with the recording devices can be tricky, a much harder problem is that of obtaining the ground truth of the data, or getting data that genuinely corresponds to a particular emotional state. In vision or speech research the subject matter is often objective: scene depth, words spoken, etc. In those cases, the ground-truth labels for the data are easily obtained. Easy-to-label data is sometimes obtained in emotion research when a singular strong emotion is captured, such as an episode of rage. However, more often the ground truth – which emotion was present – is difficult to establish.

Consider a task where a person uses the computer to retrieve images. Suppose our job is to analyze physiological signals of the user as he or she encounters pleasing or irritating features or content within the system. We wish to label the data according to the emotional state of the user (ground truth). Here, the problem is complicated because there is little way of knowing whether the person was truly pleased or irritated when encountering the stimuli intended to induce these states. We may have tried to please, but the user was irritated because of something she remembered unrelated to the task. We may have tried to irritate, and not succeeded. If we *ask* the person how she felt, her answer can vary according to her awareness of her feelings, her comfort in talking about feelings, her rapport with the administrator(s) of the experiment, and more. *When* you ask is also important – soon and often is likely to be more accurate, but also more irritating, thereby changing the emotional state. Thus, measurement of ground truth disturbs the state of that truth. When it comes to faces or voices, we can see if the person was smiling or hear if her voice sounded cheerful, but that still does not mean that she was happy. With physiology, little is known about *how* emotions make their impact, but the signals are also potentially more sincere expressions of the user's state since they tend to be less mediated by cognitive and social influences.

2.1 Five factors in eliciting emotion

In beginning this research, we thought it would be simple to ask somebody to feel and express an emotion, and to record data during such episodes. Indeed, this is the approach taken in most facial and vocal expression studies to date: turn the camera and microphone on, ask the subject to express joy, anger, etc, record it, and label it by what was requested. However, obtaining high quality physiological data for affect analysis requires attention to experimental design issues not traditionally required in pattern recognition research. Here we outline five (not necessarily independent) factors that influence data collection, to serve as a useful guide for researchers trying to obtain affect data. We summarize the factors by listing their extreme conditions, but there are also in-between conditions:

1. *subject-elicited* vs. *event-elicited*: Does subject purposefully elicit emotion or is it elicited by a stimulus or situation outside the subject's efforts?

2. *lab setting* vs. *real-world*: Is subject in a lab or in a special room that is not their usual environment?

3. *expression* vs. *feeling*: Is the emphasis on external expression or on internal feeling?

4. *open-recording* vs. *hidden-recording*: Does subject know that anything is being recorded?[1]

5. *emotion-purpose* vs. *other-purpose*: Does subject know that the experiment is about emotion?

The most natural set-up for gathering genuine emotions is opportunistic: the subject's emotion occurs as a consequence of personally significant circumstances (*event-elicited*); it occurs while they are in some natural location for them (*real-world*); the subject feels the emotion internally (*feeling*); subject behavior, including expression, is not influenced by knowledge of being in an experiment or being recorded (*hidden-recording, other-purpose*). Such data sets are usually impossible to get because of privacy and ethics concerns, but as recording devices are increasingly prevalent, people may cease to be aware of them, and the data captured by these devices can be as if the devices were hidden. Researchers may try to create such opportunistic situations; for example, showing an emotion-eliciting film in a theater without telling subjects the true purpose of the showing, and without telling them that they have been videotaped until afterward. Even so, an emotion-inducing movie scene will affect some viewers more than others, and some maybe not at all.

The opportunistic situation contrasts with the experiments typically used to gather data for facial or vocal expression recognition systems. The common set-up is one in which data are gathered of a *subject-elicited* external *expression* in a *lab setting*, in front of a visible *open-recording* camera or microphone, with knowledge that the data will be used for analysis of emotion (*emotion-purpose*). Such expressions, made with or without corresponding internal feelings, are "real" and important to recognize, even if not accompanied by true feelings. Social communication involves both unfelt emotions (emphasizing expression) and more genuine ones (emphasizing feeling.) A protocol for gathering data may emphasize external expression or internal feeling; both are important. Moreover, there is considerable overlap since physical expression can help induce the internal feeling, and vice-versa.

In this paper, we gathered real data following a *subject-elicited*, close to *real-world* (subject's comfortable usual workplace), *feel-*

[1] When the presence of the camera or other recording device is hidden, it is ethically necessary (via the guidelines of the MIT Committee on the Use of Humans as Experimental Subjects) to debrief the subject, let them know why the secrecy was necessary, tell them what signals have been recorded, and obtain their permission for data analysis, destroying the data if the subject withholds permission. Unlike with video, it is not yet possible to record a collection of physiological signals without the subject being aware of the sensors in some form (although this is changing with new technology.) However, subjects can be led to believe that their physiology is being sensed for some reason other than emotions, which is often an important deception since a subject who knows you are trying to make them frustrated may therefore not get frustrated. We have found that such deception is acceptable to almost all subjects when properly conducted. Nonetheless, we believe deception should not be used unless it is necessary, and then only in accord with ethical guidelines.

ing, *open-recording*, *emotion-purpose* methodology. The key one of these factors that makes our data unique is that the subject tried to elicit an internal *feeling* of each emotion.

2.2 Single-subject multiple-day data collection

The data we gather is from a single subject over many weeks of time, standing in contrast to efforts that examine many subjects over a short recording interval (usually single session on only one day). Although the scope is limited to one subject, the amount of data for this subject encompasses a larger set than has traditionally been used in affect recognition studies involving multiple subjects. The data are potentially useful for many kinds of analysis, and will be made available for research purposes.

There are many reasons to focus on person-dependent recognition in the early stages of affect recognition, even though some forms of emotion communication are not only person-independent, but have been argued, namely by Paul Ekman, to be basic across different cultures. Ekman and colleagues acknowledge that even simply labeled emotions like "joy" and "anger" can have different interpretations across individuals within the same culture; this complicates the quest to see if subjects elicit similar physiological patterns for the same emotion. When lots of subjects have been examined over a short amount of time, researchers have had difficulty finding significant physiological patterns, which may be in part because physiology can vary subtly with how each individual interprets each emotion. By using one subject, who tried to focus on the same personal interpretation of the emotion in each session, we hoped to maximize the chance of getting consistent interpretations for each emotion. This also means that the expressive data can be expected to differ for another subject, since the way another subject interprets and reacts to the emotions may differ. Hence, a weakness of this approach is that the precise features and recognition results we obtain with this data may not be the same for other subjects. However, the methodology for gathering and analyzing the data in this paper is not dependent on the subject; the approach described in this paper is general.

Emotion recognition is in an early stage of research, similar to early research on speech recognition, where it is valuable to develop person-dependent methods. For personal computing applications, we desire the machine to learn an individual's patterns, and not just some average response formed across a group, which may not apply to the individual.

The subject in our experiments was a healthy graduate student with two years acting experience plus training in visualization, who was willing to devote over six weeks to data collection. The subject sat in her quiet workspace early each day, at roughly the same time of day, and tried to experience eight affective states with the aid of a computer controlled prompting system, the "Sentograph," developed by Manfred Clynes [33], and a set of personally-significant imagery she developed to help elicit the emotional state.

The Clynes protocol for eliciting emotion has three features that contribute to helping the subject feel the emotions and that make it appropriate for physiological data collection: 1. it sequences eight emotions in a way that supposedly makes it easier for many people to transition from emotion to emotion, 2. it engages physical expression – asking the subject to push a finger against a button with a dual axis pressure sensor in an expressive way–but in a way that limits motion artifacts being introduced to the physiological signals. Physical expression gives somatosensory feedback to the subject, a process that can help focus and strengthen the feeling of the emotion [34]. 3. it prompts the subject to repeatedly express the same emotion during an approximately 3 minute interval, at a rate dependent on the emotion in order to try to intensify the emotional experience [33].

The order of expression of the eight states: *no emotion, anger, hate, grief, platonic love, romantic love, joy and reverence* was found by Clynes to help subjects reliably feel each emotion; for example, it would be harder on most subjects to have to transition between the positive and negative states, e.g., *joy, hate, platonic love, grief, etc.* However, because interpretations for each of the emotions may vary with each individual, we do not expect this order to be optimal for everyone.

Descriptive guidelines on the meaning of

Emotion	Imagery	Description	Arousal	Valence
(N)o emotion	blank paper, typewriter	boredom, vacancy	low	neut.
(A)nger	people who arouse rage	desire to fight	very high	very neg.
(H)ate	injustice, cruelty	passive anger	low	neg.
(G)rief	deformed child, loss of mother	loss, sadness	high	neg.
(P)latonic love	family, summer	happiness, peace	low	pos.
Romantic (L)ove	romantic encounters	excitement, lust	very high	pos.
(J)oy	The music "Ode to Joy"	uplifting happiness	med. high	pos.
(R)everence	church, prayer	calm, peace	very low	neut.

Table 1: The subject's descriptions of imagery and emotional character used for each of the eight emotions.

each emotion were developed by the subject before the experiment. The subject reported the images she used to induce each state, the degree to which she found each experience arousing (exciting, distressing, disturbing, tranquil), and the degree to which she felt the emotion was positive or negative (valence) (See Table 1). Daily ratings varied in intensity–both up and down, with no particular trend–but the overall character of each state was consistent over the weeks of data collection.

The eight emotions in this study differ from those typically explored. Although there is no widespread agreement on the definition and existence of "basic" emotions and which names would comprise such a list, researchers on facial expressions tend to focus on *anger, fear, joy, and sadness* with *disgust* and *surprise* often examined as well as *contempt*, *acceptance,* and *anticipation* (e.g., [32]). Theorists still do not agree on what an emotion is, and many of them do not consider *love* and *surprise* to be emotions. Our work is less concerned with finding a set of "basic" emotions, and more concerned with giving computers the ability to recognize whatever affective states might be relevant in a personalized human-computer interaction. The ideal states for a computer to recognize will depend on the application. For example, in a learning-tutor application, detecting expressions of curiosity, boredom, and frustration may be more relevant than detecting emotions on the theorists' "basic" lists.

Clynes' set of eight was motivated by considering emotions that have been communicated through centuries of musical performance on several continents. We started with his set not because we think it is the best for computer-human interaction (such a set is likely to vary with computer applications – entertainment, business, socializing, etc.), but rather because this set together with its method for elicitation had shown an ability to help subjects reliably feel the emotions, and had shown repeatable signs of physical differentiation in how subjects' finger pressure, applied to a finger rest, differs with each emotion [33], [35], a measurable outcome that suggests different states were being achieved. It was important to our investment in long-term data collection that we have a reliable method of helping the user repeatedly generate distinct emotional states.

For the purposes of this research, the specific emotions and their definitions are not as important as the fact that (1) the subject could relate to the named emotion in a consistent, specific, and personal way, and (2) the emotion categories span a range of high and low *arousal* and positive and negative *valence*. These two dimensions are believed to be the most important dimensions for categorizing emotions [36], and continue to be used for describing emotions that arise in many contexts, including recent efforts to categorize emotions arising when people look at imagery [37]. The arousal axis ranges from calm and peaceful to active and excited, while the valence axis ranges from negative (displeasing) to positive (pleasing).

2.3 Experimental method and construction of data sets

Data were gathered from four sensors: a triode electromyogram ($\mathcal{E}$) measuring facial muscle tension along the masseter (with Ag-

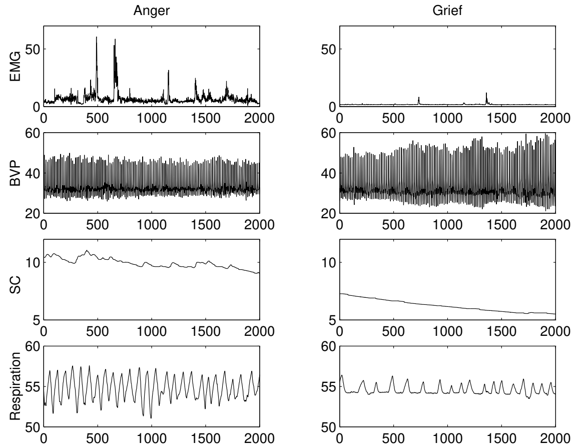


Figure 1: Examples of physiological signals measured from a user while she intentionally expressed anger (left) and grief (right). From top to bottom: electromyogram (microvolts), blood volume pressure (percent reflectance), skin conductivity (microSiemens), and respiration (percent maximum expansion). Each box shows 100 seconds of response. The segments shown here are visibly different for the two emotions, which was not true in general.

AgCl electrodes of size 11mm each, and 10-20 high-conductivity gel), a photoplethysmyograph measuring blood volume pressure ($\mathcal{B}$) placed on the tip of the ring finger of the left hand, a skin conductance ($\mathcal{S}$) sensor measuring electrodermal activity from the middle of the three segments of the index and middle fingers on the palm-side of the left hand (with 11mm Ag-AgCl electrodes and K-Y Jelly used for low-conductivity gel) and a Hall effect respiration sensor ($\mathcal{R}$) placed around the diaphragm. The left hand was held still throughout data collection, and the subject was seated and relatively motionless except for small pressure changes she applied with her right hand to the finger rest. Sensors and sampling were provided by the Thought Technologies ProComp unit, chosen because the unit is small enough to attach to a wearable computer and offers eight optically isolated channels for recording. Signals were sampled at 20 Hz.[2] The ProComp automatically computed the heart rate ($\mathcal{H}$) as a function of the inter-beat intervals of the blood volume pressure, $\mathcal{B}$. More details on this system and on our methodology are available [38].

Each day's session lasted around 25 minutes, resulting in around 28 to 33 thousand samples per physiological signal, with each different emotion segment being around 2 to 5 thousand samples long, due to the variation built into the Clynes method of eliciting the emotional states [33]. Eight signal segments of the raw data (2000 samples each) from Data Set I are shown in Figure 1. On roughly a third of the thirty days for which we collected data, either one or more sensors failed during some portion of the 25-minute experiment because an electrode came loose, or one or more channels failed to sample and save some of the data properly. From the complete or nearly-complete sessions, we constructed two overlapping Data Sets.

Data Set I was assembled before the thirty days were over, and was formed as follows: segments of length 2000 samples (100 seconds) of data were taken from each of the signals $\mathcal{E}, \mathcal{B}, \mathcal{G}$ and $\mathcal{R}$ for each of the eight emotions, on each of nineteen days where there were no failures in these segments of data collection. The 2000 samples were taken from the end of each emotion segment to avoid the transitional onset where the subject was prompted to move to the next emotion. A twentieth day's data set was created out of a combination of partial records in which some of the sensors had failed.

Data Set II is a larger data set comprised of 20 days in which the sensors did not fail during any part of the experiment for the five signals: $\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R}$ and $\mathcal{H}$. These 20 days included 16 of the original days from Data Set I, but for all 20 days we used all of the samples available for each emotion, thereby including the transitional regions. Because different emotions lasted for different lengths of time, the 2000 samples in Data Set I were at times closer or farther away from the beginning of an emotion segment. To avoid this bias, and to maximize data available for training and testing, Data Set II includes all the samples for all the emotions and signals over twenty days, roughly 2000 to 5000 samples

[2] The electromyogram is the only signal for which this sampling rate should have caused aliasing. However, our investigation of the signal showed that it registered a clear response when the jaw was clenched vs. relaxed; thus, it was satisfactory for gathering coarse muscle tension information.

per emotion per signal per day. With an average of twice the number of samples (or minutes of data) as Data Set I, Data Set II resulted in an average 10% gain in performance when we compared it to Data Set I across all the methods.

3 Feature Extraction, Selection, and Transformation

Because the signals involved have different and complex sources, because there are not yet good models to describe them, and because of an interest in seeing how some classical methods perform before an extensive modeling effort is launched, we choose in this paper to explore a feature-based approach to classification.

3.1 Proposed feature sets

The psychophysiology and emotion literature contains several efforts to identify features of bodily changes (facial muscle movements, heart-rate variations, etc.) that might correlate with having an emotion (e.g., [31]). We gather a variety of features, some from the literature and some that we propose. Several that we propose are physically motivated, intended to capture the underlying nature of specific signals, such as the way respiration is quasi-periodic, while others are simple statistics and nonlinear combinations thereof. We do not expect the best classifier to require all the features proposed below, or even to require such a huge number of features. Our effort is to advance the state of the art in pattern recognition of affect from physiology by proposing a large space of reasonable features and systematically evaluating subsets of it and transformations thereof. Below are presented six statistical features, followed by ten more physically-motivated features aimed at compensating for day-to-day variations. Which features were found to be most useful (as a function of each of the classifiers) will be summarized later in Table 9.

The six statistical features can be computed for each of the signals as follows. Let the signal $\{\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R}, \mathcal{H}\}$ from any one of the eight emotion segments be designated by X. The signal is gathered for 8 different emotions each day, for 20 days. Let X_n represent the value of the n^{th} sample of the raw signal, where $n = 1, \ldots, N$, with $N = 2000$ for Data Set I, and with N in the range of 2000 to 5000 for Data Set II. Let $\tilde{X}_n$ refer to the normalized signal (zero mean, unit variance):

$$\tilde{X}_n = \frac{X_n - \mu_X}{\sigma_X} \qquad i = 1, \ldots, 4$$

where μ_X and σ_X are the means and standard deviations of X as explained below. Following are six statistical features we investigated:

1. the means of the raw signals
$$\mu_X = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{1}$$

2. the standard deviations of the raw signals
$$\sigma_X = \left(\frac{1}{N-1}\sum_{n=1}^{N}\left(X_n - \mu_X\right)^2\right)^{1/2} \tag{2}$$

3. the means of the absolute values of the first differences of the raw signals
$$\delta_X = \frac{1}{N-1}\sum_{n=1}^{N-1} \left|X_{n+1} - X_n\right| \tag{3}$$

4. the means of the absolute values of the first differences of the normalized signals
$$\tilde{\delta}_X = \frac{1}{N-1}\sum_{n=1}^{N-1} \left|\tilde{X}_{n+1} - \tilde{X}_n\right| = \frac{\delta_X}{\sigma_X} \tag{4}$$

5. the means of the absolute values of the second differences of the raw signals
$$\gamma_X = \frac{1}{N-2}\sum_{n=1}^{N-2} \left|X_{n+2} - X_n\right| \tag{5}$$

6. the means of the absolute values of the second differences of the normalized signals
$$\tilde{\gamma}_X = \frac{1}{N-2}\sum_{n=1}^{N-2} \left|\tilde{X}_{n+2} - \tilde{X}_n\right| = \frac{\gamma_X}{\sigma_X} \tag{6}$$

The features (1)-(6) were chosen to cover and extend a range of typically measured statistics in the emotion physiology literature[39]. (Means and variances are already commonly computed; the first difference approximates a gradient.) Note that not all the features are independent; in particular, $\tilde{\delta}_X$ and $\tilde{\gamma}_X$ are nonlinear combinations of other features. Also, the heart rate signal, $\mathcal{H}$, is derived from the blood volume pressure signal, $\mathcal{B}$, by a nonlinear transformation performed automatically by the ProComp sensing system. It did not require an additional sensor, and so was dependent upon $\mathcal{B}$. The dependencies are not linear; consequently, they are not obtainable by the linear combination methods used later to reduce the dimensionality of the feature space and can potentially be of use in finding a good transformation for separating the classes. The comparisons below will verify this.

One advantage of the features in (1)-(6) is that they can easily be computed in an online way [40], which makes them advantageous for real-time recognition systems. However, the statistical features do not exploit knowledge we may have about the physical sources of the signals, and provide no special normalization for day-to-day variations in the signals. Factors such as hand washing, gel application, and sensor placement can easily affect the statistics. These influences combine with the subject's daily mood and with other cognitive and bodily influences in presently unknown ways, making them hard to model.

In an effort to compensate for some of the non-emotion-related variations of the signals, and to include more physically-motivated knowledge about the signal (such as underlying periodic excitations), we also compute and evaluate another set of ten physiology-dependent features, $f_1 - f_{10}$, described below:

From the inter-beat intervals of the blood volume pressure waveform, the Procomp computes the heart rate, $\mathcal{H}$, which is approximately 1/the inter-beat intervals. We applied a 500-point (25 sec) Hanning window, h, [41] to form a smoothed heart-beat rate, $b = \mathcal{H} * h$, then took the mean:

$$f_1 = \frac{1}{N}\sum_{n=1}^{N} b_n. \tag{7}$$

The average acceleration or deceleration of the heart beat rate was calculated by taking the mean of the first difference:

$$f_2 = \frac{1}{N-1}\sum_{n=1}^{N-1}(b_{n+1} - b_n) = \frac{1}{N-1}(b_N - b_1). \tag{8}$$

The skin conductivity signal, $\mathcal{S}$ contains high frequency fluctuations that may be noise; these fluctuations are reduced by convolving with h, a 25 sec Hanning window, to form $s = \mathcal{S} * h$. We also use a form of contrast normalization to account for baseline fluctuations; this measure was proposed by Rose in 1966 [42] and found to be valuable over years of psychophysiology [43], [44], where $\max(g)$ and $\min(g)$ are with respect to the whole day's data (for all emotions that day):

$$f_3 = \frac{s \min(s)}{\max(s) - \min(s)}. \tag{9}$$

The mean of the first difference of the smoothed skin conductivity is also proposed:

$$f_4 = \mu_{(s_{n+1} - s_n)} = \frac{1}{N-1}(s_N - s_1). \tag{10}$$

The respiration sensor measured expansion and contraction of the chest cavity using a Hall effect sensor attached around the chest with a velcro band. Let N_d be the number of samples collected that day. To account for variations in the initial tightness of the sensor placement from day to day, we formed the mean of the whole day's respiration data:

$$\mu_{\mathcal{R},\mathrm{day}} = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathcal{R}_n, \tag{11}$$

and then subtracted this to get $r = \mathcal{R} - \mu_{\mathcal{R},\mathrm{day}}$. Two respiration features were then formed as:

$$f_5 = \frac{1}{N}\sum_{n=1}^{N} r_n \tag{12}$$

and

$$f_6 = \frac{1}{N-1}\sum_{n=1}^{N}\left(r_n - \mu_{\mathcal{R},\mathrm{day}}\right)^2. \tag{13}$$

The four features f_7 - f_{10} represent frequency information in the respiration signal. Each

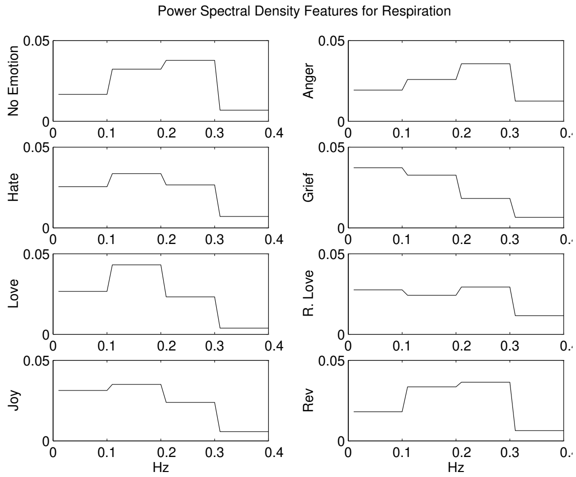


Figure 2: Power Spectral Density from 0.0-0.4 Hz in the respiration signal for eight emotions. The heights of the four bins shown here were used as features $f_7 - f_{10}$.

was computed using the power spectral density function (PSD command) of Matlab, which uses Welch's averaged periodogram. The features, illustrated in Figure 2, represent the average energy in each of the first four 0.1 Hz bands of the power spectral density range 0.0-0.4 Hz.

3.2 Selecting and transforming features

We first compare two techniques that have appeared in the literature to establish a benchmark: Sequential Floating Forward Search (SFFS), and Fisher Projection (FP). Next, we propose and compare a new combination of these, which we label SFFS-FP. This combination is motivated by the dual roles of the two methods: SFFS *selects* from a set of features, while Fisher Projection *linearly transforms* a set of features. Since feature selection is nonlinear, the cascade of these two methods should provide a more powerful combination than either alone. This was confirmed in our experiments (below).

The **Sequential Floating Forward Search (SFFS)** method [45] is chosen because of its consistent success in previous evaluations of feature selection algorithms, where it has been shown to outperform methods such as Sequential Forward Search (SFS), Sequential Backward Search (SBS), Generalized SFS and SBS, and Max-Min in comparison studies [46]. Of course the performance of SFFS is data dependent and the data set here is new and different; SFFS may not be the best method to use. Nonetheless, because of its well documented success in other pattern recognition problems, it will help establish a benchmark for this new application area. The SFFS method takes as input the values of n features. It then does a non-exhaustive search on the feature space by iteratively including and omitting features. It outputs one subset of m features for each m, $2 \leq m \leq n$, together with its classification rate. The algorithm is described in detail in [47].

Fisher projection (FP) [48] is a well-known method of reducing dimensionality by finding a linear projection of the data to a space of fewer dimensions where the classes are well separated. Due to the nature of the Fisher projection method, the data can only be projected down to $c-1$ (or fewer if one wants) dimensions, assuming that originally there are more than $c-1$ dimensions and c is the number of classes. If the amount of training data is inadequate, or if the quality of some of the features is questionable, then some of the dimensions of the Fisher projection may be a result of noise rather than a result of differences among the classes. In this case, Fisher might find a meaningless projection which reduces the error in the training data but performs poorly in the testing data. For this reason, we not only separate training and testing data, but we also evaluate projections down to fewer than $c-1$ dimensions. Note that if the number of features n is smaller than the number of classes c, the Fisher projection is meaningful only up to at most $n-1$ dimensions. Therefore the number of Fisher projection dimensions d is $1 \leq d \leq \min(n, c) - 1$, e.g., when 24 features are used on all 8 classes, all $d = [1, 7]$ are tried, and when 4 features are used on 8 classes, all $d = [1, 3]$ are tried.

A **Hybrid SFFS with Fisher Projection (SFFS-FP)** method is proposed, implemented, and evaluated here for comparison. As mentioned above, the SFFS algorithm proposes one subset of m features for each m, $2 \leq m \leq n$. It *selects*, but does not *transform* the features. Instead of feeding the Fisher algorithm with all possible features, we use the subsets that the

SFFS algorithm proposes as input to the Fisher Algorithm. Note that the SFFS method is used here as a preprocessor for reducing the number of features fed into the Fisher algorithm, and not as a classification method.

4 Classification

This section describes and compares the results of a set of classification experiments leading up to the best results (81%) shown in Tables 6 and 7.

The SFFS software employs a k-nearest-neighbor (k-NN) classifier [49], so that it not only outputs the best set of features according to this classifier, but their classification accuracy as well. We used the k-NN classifier for benchmarking the SFFS method, following the methodology of Jain and Zongker [46]. For FP and SFFS-FP we used a MAP classifier, with details below.

In all three comparisons: SFFS, FP, and SFFS-FP, we used the leave-one-out method for cross validation because of the relatively small amount of data available and the high dimensional feature spaces. In each case, the data point (vector of features for one day's data for one emotion) to be classified was excluded from the data set before the SFFS, FP, or SFFS-FP was run. The best set of features or best transform (determined from the training set) was then applied to the test data point to determine classification accuracy.

For each k, where we varied $1 \leq k \leq 20$, the SFFS algorithm output one set of m features for each $2 \leq m \leq n$. For SFFS-FP, we computed all possible Fisher projections for each of these feature sets. For both the FP and SFFS-FP methods we then fit Gaussians to the data in the reduced d-dimensional feature space and applied Maximum a Posteriori (MAP) classification to the features in the Fisher-transformed space. The specific algorithm was:

1. The data point to be classified (the testing set only includes one point) is excluded from the data set. The remaining data set is used as the training set.

2. The Fisher projection matrix (used in either FP or SFFS-FP) is calculated from only the training set. Then both the training and testing set are projected down to the d dimensions found by Fisher.

3. The data in the d-dimensional space is assumed to be Gaussian. The respective means and covariance matrices of the classes are estimated from the training data.

4. The posterior probability of the data point is calculated.

5. The data point is then classified as coming from the class with the highest posterior probability.

4.1 Initial results - Data Set I

Our first comparison was made on Data Set I using (1)-(6) computed on the four raw physiological signals, yielding 24 features: μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R})$. The results of SFFS, Fisher, and SFFS-FP are shown in Table 2. Table 2 shows the results applied to all eight emotions as well as the results applied to all sets of C emotions, where $C = 3, 4, 5$. The best-recognized emotions were (N)eutral, (A)nger, (G)rief, (J)oy, and (R)everence. The states of (H)ate and the states of love: (P)latonic Love, and Romantic (L)ove, were not well discriminated by any of the classifiers.

How good are these results – are the differences between the classifiers significant? For each pair of results here and through the rest of the paper, we computed the probability that the lower error rate really is lower, treating error rate over the 160 (or fewer) trials as a Bernoulli random variable. For example, (last row of Table 2) we compared the 20 errors made by Fisher to the 10 errors made by SFFS-FP, over the 60 trials for the three emotions (AJR). The confidence that the performance was improved was 98%. Although the performance improvements of SFFS-FP over Fisher in the other rows range from 5 to over 8 percentage points, the confidences range from 77% to 87%; thus the performance improvement of SFFS-FP cannot be said to be statistically significant in these cases, given that 90-95% confidence is usually required for that statement. Nonetheless, all the classifiers in this table provide statistically significant

Number of Emotions	Random Guessing (%)	SFFS (%)	Fisher (%)	SFFS-FP (%)
8	12.5	40.6	40.0	46.3
5 (NAGJR)	20.0	64.0	60.0	65.0
4 (NAGR)	25.0	70.0	61.3	68.7
4 (AGJR)	25.0	72.5	60.0	67.5
3 (AGR)	33.3	83.3	71.7	80.0
3 (AJR)	33.3	88.3	66.7	83.3

Table 2: Classification accuracy for three methods applied to Data Set I, starting with the 24 features μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R})$. The best-recognized emotions are denoted by their first initial: (N)eutral, (A)nger, (G)rief, (J)oy, (R)everence.

improved performance over a random classifier (confidence $> 99.99\%$).

Table 3 shows the number of dimensions and the number of features, respectively, that gave the best results. From looking at these numbers, we can conclude that performance was sometimes improved by projecting down to fewer than $c-1$ dimensions for a c class problem. We can also see that a broad range of numbers of features led to the best results, but in no case were 20 or more features useful for constructing the best classifier.

4.2 The problem of day-dependence

In visually examining several projections into 2-D, we noticed that the features of different emotions from the same day often clustered more closely than did features for the same emotions on different days. To try to quantify this "day dependence" we ran a side experiment: how hard would it be to classify *what day* each emotion came from, as opposed to *which emotion* was being expressed? Because classifying days is not the main interest, we only ran one comparison: Fisher Projection applied to the 24 features μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R})$ computed on Data Set I. Leave-one-out cross validation was applied to every point, and MAP classification was used on the classes (days or emotions), as described above. When the classes were the $c = 20$ days, (vs. the $c = 8$ emotions) then the recognition accuracy jumped to 83%, significantly better than random guessing (5%) and, significantly better than the 40% found for emotion classification using these features.

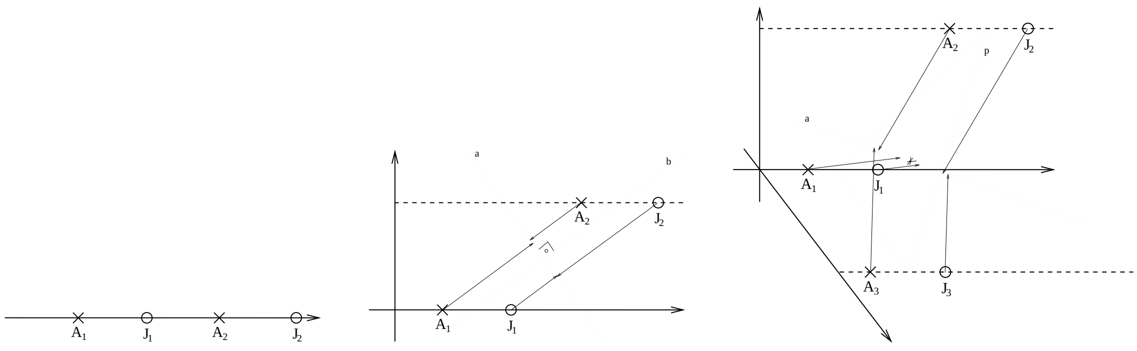
The day dependence is likely due to three factors: (1) skin-sensor interface influences– including hand washing, application of slightly different amounts of gel, and slight changes in positioning of the sensors; (2) variations in physiology that may be caused by caffeine, sugar, sleep, hormones, and other non-emotional factors; (3) variations in physiology that are mood and emotion dependent – such as an inability to build up an intense experience of joy if the subject felt a strong baseline mood of sadness that day. The first factor is slightly more controllable by using disposable electrodes that come from the manufacturer containing a pre-measured gel, and always having the subject wash her hands in the same way at the same time before each session, steps we did not impose. However, we made an effort to place the sensors as similarly from day to day as manually possible and to manually apply the same amount of gel each day. Nonetheless, many of these sources of variation are natural and cannot be controlled in realistic long-term measuring applications. Algorithms and features that can compensate for day-to-day variations are needed.

4.3 Day matrix for handling day-dependence

The 24 statistical features extracted from the signals are dependent on the day the experiment was held. We now augment the 20×24 matrix of the 20 days' 24 features with a 20×19 *day matrix*, which appends a vector of length 19 to each vector of length 24. The vector is the same for all emotions recorded the same day,

Number of Emotions	No. of Dimensions		No. of Features	
	Fisher-24	SFFS-FP	m_{SFFS}	$m_{SFFS-FP}$
8	6/7	4,5/7	13	17
5 (NAGJR)	3/4	3/4	12-17	15
4 (NAGR)	3/3	3/3	9-15,18	19
4 (AGJR)	3/3	2/3	7-8	12
3 (AGR)	2/2	2/2	2-16	12
3 (AJR)	1/2	2/2	6-14	7

Table 3: Number of dimensions and number of features m used in the Fisher projections and feature selection algorithms that gave the results of Table 2. The denominator in the dimensions columns is the maximum number of dimensions that could have been used; half of the time the results were better using a number less than this maximum (the numerator). When a range is shown for the number of features, m, then the performance was the same for the whole range.



(a) (b) (c)

Figure 3: Illustration of a highly day-dependent feature for 2 emotions from 2 different days. (a) The feature values for (A)nger and (J)oy from 2 different days. (b) Addition of an extra dimension allows for a line b to separate Anger from Joy. The data can be projected down to line a, so the addition of the new dimension did not increase the *final* number of features. (c) In the case of data from 3 different days, addition of 2 extra dimensions allows for a plane p to separate Anger from Joy. The data can again be projected down to line a, not increasing the final number of features.

and differs among days. (The vectors were constructed by generating 20 equidistant points in a 19-dim space.) Let us briefly explain the principle with an illustration in Fig. 3.

Consider when the data come from 2 different days and only 1 feature is extracted. (This is the simplest way to visualize, but it trivially extends to more features). Although the feature values of one class are always related to the values of the other classes in the same way, e.g., the mean electromyogram signal for anger may always be higher than the mean electromyogram for joy, the actual values may be highly day-dependent (Fig. 3a). To alleviate this problem an extra dimension can be added before the features are input into the Fisher Algorithm (Fig. 3b). If the data came from 3 different days, 2 extra dimensions are added rather than one (Fig. 3c), etc. In the general case, $D-1$ extra dimensions are needed for data coming from D different days; hence, we use 19 extra dimensions. The above can be also seen as using the minimum number of dimensions so that each of D points can be at equal distance from all others. Therefore the $D-1$ dimensional vector contains the coordinates of one such point for each day.

The effect of the day matrix on classification can be seen in Table 4, where we run Fisher and SFFS-FP with and without the day matrix, and compare it to SFFS, running with the same 24 features as above. Note that it is meaningless to apply SFFS to the day matrix, so that comparison is omitted. The use of the day matrix improves the classification from 3.1% to 9.4%, although only the highest two of these improvements are significant at $> 90\%$ and $> 95\%$ (the other confidences are 71%-78%).

Table 4 also reveals that all the methods perform better on DataSet II than on DataSet I. The improvements range from 5 to 13 percentage points, with confidences 81%, 94%, 97%, 98%, and 99%.

4.4 Baseline matrix for handling day-dependence

We propose and evaluate another approach: use of a *baseline matrix* where the Neutral (no emotion) features of each day are used as a baseline for (subtracted from) the respective features of the remaining 7 emotions of the same day. This gives an additional 20×24 matrix for each of the 7 non-neutral emotions. The resulting classification results on 7 emotions, run on DataSet I, are in Table 5, together with an additional trial of the day matrix for this case. All the results are significantly higher than random guessing (14.3%) with confidence $> 99.99\%$. Comparing among the various ways of handling day-dependence, we find the most significant improvement to be SFFS-FP using the baseline matrix, which at 54.3% is an improvement over 45% (confidence 94%). The performance improvements of SFFS-FP over that of Fisher in the last two rows of the table are also statistically significant (confidence 99%). Combining all the features (original 24 + baseline + day) appears to result in a decrease in performance, suggesting the limitations of Fisher with too many bad dimensions, where it cannot select out features, but only transform them. However, the decreases (from 54.3 to 49.3 and 40.7 to 35.0) have 80% and 84% confidence, so are not considered significant.

4.5 Better features for handling day-dependence

The comparisons above were all conducted with the original six statistical features (1)-(6); we would now like to see how the features $f_1 - f_{10}$ influence classification. We compare four spaces of features from which the algorithms can select and transform subsets: the original six, μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, for $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R})$ for a total of 24 features; these same 24 features plus the same six statistics for $X = \mathcal{H}$ for a total of 30 features; features $f_1 - f_{10}$ plus $\mu_{\mathcal{E}}$ which were shown to be useful in our earlier investigation [50], and the combination of all 40 of these.

The results are in Table 6. First, we consider the case where all 40 features are available (the last row). Comparing each entry in the last row of this table with the same column entry in the first row, we find that all the results are significantly improved (confidence $> 99.7\%$).

Next, we look at the results with the day matrix. It appears to improve upon the performance of the statistical features in the first two rows; however, the improvements are only a few percentage points and are not significant (confidences 63%- 89.5%). The additional features of the day matrix can become a liability in the

Data Set with 24 Features	Without Day Matrix			With Day Matrix	
	SFFS (%)	Fisher (%)	SFFS-FP (%)	Fisher (%)	SFFS-FP (%)
DataSet I	40.6	40.0	46.3	49.4	50.6
DataSet II	49.4	51.3	56.9	54.4	63.8

Table 4: Classification accuracy for all 8 emotions for DataSet I and DataSet II improves with the use of the day matrix. These results were obtained with 24 statistical features, μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R})$ for both data sets. The day matrix adds 19 dimensions to each feature that is input into the Fisher Algorithm.

Feature Space (Dimensions)	SFFS (%)	Fisher (%)	SFFS-FP (%)
Original (24)	42.9	39.3	45.0
Orig.+ Day (43)	N/A	39.3	45.7
Orig.+ Base. (48)	49.3	40.7	54.3
Orig.+ Base.+ Day (67)	N/A	35.0	49.3

Table 5: Data Set I, 7-emotion (all but the Neutral state) classification results, comparing three methods for incorporating the day information.

presence of the $f_1 - f_{10}$ features, which compensate much better for the day-to-day variations (the decrease from 70% to 61.3% is significant at 95% confidence.) Without the day matrix, the $f_1 - f_{10}$ features alone perform significantly better than the statistical features (confidence $>$ 93% in all six comparisons.)

The best performance occurs when the methods have all forty features at their disposal, and the individual best of these occurs with SFFS-FP. The overall best rate of 81.25% is significantly higher ($>$ 95% confidence) than sixteen of the cases shown in Table 6. The three exceptions are in the last row – compared to rates of 78.8% and to 77.5% the confidences are only 71% and 79% that 81.25% is a genuine improvement.

Table 7 affords a closer look at breaking points of the best-performing case: the SFFS-FP algorithm operating on the full set of 40 features. Summing columns, one sees that the greatest number of false classifications lie in the categories of (P)latonic love and (J)oy. It is possible that this reflects an underlying predilection on the part of this subject toward being in these two categories, even when asked to experience a different emotion; however, one should be careful not to read more into these numbers than is justified.

Because of the longstanding debate regarding whether physiological patterns reveal valence differences (pleasing-displeasing aspects) of emotion, or only arousal differences, we consider here an alternate view of the results in Table 7. Table 8 rearranges the rows and columns into groups based on similar arousal rating or similar valence rating. From this we see that both arousal and valence are discriminated at rates significantly higher than random (confidence $>$ 99.99%), and that the rate of discrimination based on valence (87%) vs. the rate obtained based on arousal (84%) do not differ in a statistically significant way.

4.6 Finding Robust Features

A feature-based approach presents not only a virtually unlimited space of possible features that researchers can propose, but an intractable search of all possible subsets of these to find the best features. When people hand-select features, they may do so with an intuitive feel for which are most important; however, hand-

Number of Features: Which formed initial space	Without Day Matrix			With Day Matrix	
	SFFS (%)	Fisher (%)	SFFS-FP (%)	Fisher (%)	SFFS-FP (%)
24: μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R})$	49.4	51.3	56.9	54.4	63.8
30: μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$, $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R}, \mathcal{H})$	52.5	56.9	60.0	58.8	63.8
11: $f_1, f_2, \ldots, f_{10}$, $\mu_\mathcal{E}$	60.6	70.0	70.6	61.3	63.1
40: all of the above	65.0	77.5	81.25	77.5	78.8

Table 6: Comparative classification rates for 8 emotions for Data Set II. The day matrix adds 19 features to the data input to the Fisher algorithm, offering no improvement when $f_1 - f_{10}$ are available.

	N	A	H	G	P	L	J	R	Total
(N)eutral	**17**	0	0	0	3	0	0	0	20
(A)nger	0	**17**	0	0	2	1	0	0	20
(H)atred	0	0	**14**	1	0	0	3	2	20
(G)rief	0	0	1	**15**	0	0	4	0	20
(P)latonic Love	0	0	0	0	**17**	2	1	0	20
Romantic (L)ove	1	1	0	0	3	**14**	1	0	20
(J)oy	0	0	1	2	0	0	**17**	0	20
(R)everence	0	0	0	1	0	0	0	**19**	20
Total	18	18	16	19	25	17	26	21	160

Table 7: Confusion matrix for the method that gave the best performance classifying 8 emotions with Data Set II (81.25%). An entry's row label is the true class, the column label is what it was classified as.

selection is rarely as thorough as machine selection and tends to overlook non-intuitive combinations that may outperform intuitive ones. Machine selection is therefore preferable, even when it is sub-optimal (non-exhaustive). Here we analyze which features the machine searches selected repeatedly. The results of automatic feature selection from twelve experiments involving SFFS (either SFFS as in Jain and Zongker's code, or SFFS-FP, or SFFS-FP with the Day Matrix) are summarized in Table 9.

From Table 9 we see that features such as the means of the heart rate, skin conductivity, and respiration were never selected by any of the classifiers running SFFS, so that they need not have even been computed for these classifiers. At the same time, features such as the mean absolute normalized first difference of the heart rate ($\tilde{\delta}_\mathcal{H}$), the first difference of the smoothed skin conductivity (f_4), and the three higher frequency bands of the respiration signal ($f_8 - f_{10}$), were always found to contribute to the best results. Features that were repeatedly selected by classifiers yielding good classification rates can be considered more robust than those that were rarely selected, but only within the context of these experiments.

There were surprises in that some features physiologists have suggested to be important such as f_3 were not found to contribute to good discrimination. Although the results here will be of interest to psychophysiologists, they must clearly be interpreted in the context of the recognition experiment here, and not as some definitive statement about a physical correlate of emotion.

Valence Rating	Very Neg.	Neg.	Neutral	Pos.	Total
Very Neg. (A)	**17**	0	0	3	20
Neg. (H,G)	0	**31**	2	7	40
Neutral (N,R)	0	1	**36**	3	40
Pos. (P,L,J)	1	3	1	**55**	60

Arousal Rating	Very High	Med. High	High	Low	Very Low	Total
Very High (A,L)	**33**	1	0	6	0	40
Med. High (J)	0	**17**	2	1	0	20
High (G)	0	4	**15**	1	0	20
Low (N,H,P)	2	4	1	**51**	2	60
Very Low (R)	0	0	1	0	**19**	20

Table 8: Rearrangements of Table 7, showing confusion matrices for emotions having similar valence ratings (139/160 = 87%) and similar arousal ratings (135/160 = 84%).

5 Conclusions

This paper has suggested that machine intelligence should include skills of emotional intelligence, based on recent scientific findings about the role of emotional abilities in human intelligence, and on the way human-machine interaction largely imitates human-human interaction. This is a shift in thinking from machine intelligence as one of primarily mathematical, verbal, and perceptual abilities. Emotion is believed to interact with all of these aspects of intelligence in the human brain. Emotions, largely overlooked in early efforts to develop machine intelligence, are increasingly regarded as an area for important research.

One of the key skills of emotional intelligence for adaptive learning systems is the ability to recognize the emotional communication of others. Even dogs can recognize their owner's affective expressions of pleasure or displeasure – an important piece of feedback. One of the difficulties researchers face in this area is the sheer difficulty of getting data corresponding to real emotional states; we have found that efforts in this area are more demanding than traditional efforts to get pattern recognition data. We presented five factors to aid researchers trying to gather good affect data.

The data gathered in this paper contrasts with that gathered for most other efforts at affect pattern recognition not so much in the use of physiology vs. video or audio, but in its focus on having the subject try to generate a *feeling* vs. having the subject try to generate an outward *expression*. One weakness of both data-gathering methods is that the subject elicited the emotion, vs. a situation or stimulus outside the subject eliciting it. Our group at MIT has recently designed and built environments that focus on emotions not generated deliberately by the subject, e.g., collecting affective data from users driving automobiles in city and highway conditions [38] and from users placed in frustrating computer situations [51]; both of these areas aim at data generation in an *event-elicited*, close to *real-world*, *feeling*, *open-recording*, *other-purpose* experiment, with effort to make the open recording so comfortable that it is effectively ignored. Nonetheless, much work remains to be done in gathering and analyzing affect data; attempts to create situations that induce emotion in subjects remain subject to uncertainty.

One facet of emotion recognition is developed here for the first time: classification of emotional state from physiological data gathered from one subject over many weeks of data collection. The corpus of person-dependent affect data is larger than any previously reported, and the methodology we developed for its analysis is subject-independent. In future work, the patterns of many individuals may be clustered into groups, according to similarity of the physiological patterns, and these results leveraged to

	24 Statistical Features				30 Statistical Features				$f_1, f_2, \ldots, f_{10}$ and $\mu_{\mathcal{E}}$				All 40 Features				Grand Total	
	S	SF	SFD	Sum	S	SF	SFD	Sum	S	SF	SFD	Sum	S	SF	SFD	Sum		
$\mu_{\mathcal{E}}$	0	1	1	2/3	0	1	1	2/3	1	1	1	3/3	0	0	1	1/3	8/12	0.67
$\sigma_{\mathcal{E}}$	0	1	1	2/3	0	1	1	2/3					0	1	1	2/3	6/9	0.67
$\delta_{\mathcal{E}}$	1	1	1	3/3	1	1	1	3/3					1	0	1	2/3	8/9	0.89
$\tilde{\delta}_{\mathcal{E}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
$\gamma_{\mathcal{E}}$	1	1	1	3/3	1	1	1	3/3					0	0	1	1/3	7/9	0.78
$\tilde{\gamma}_{\mathcal{E}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
$\mu_{\mathcal{B}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
$\sigma_{\mathcal{B}}$	0	0	1	1/3	0	0	0	0/3					0	0	0	0/3	1/9	0.11
$\delta_{\mathcal{B}}$	0	0	0	0/3	0	1	1	2/3					0	0	0	0/3	2/9	0.22
$\tilde{\delta}_{\mathcal{B}}$	1	0	1	2/3	0	1	1	2/3					0	1	1	2/3	6/9	0.67
$\gamma_{\mathcal{B}}$	0	0	0	0/3	0	0	0	0/3					0	0	0	0/3	0/9	0.00
$\tilde{\gamma}_{\mathcal{B}}$	0	1	1	2/3	0	1	1	2/3					0	0	1	1/3	5/9	0.56
$\mu_{\mathcal{H}}$					0	0	0	0/3					0	0	0	0/3	0/6	0.00
$\sigma_{\mathcal{H}}$					0	0	0	0/3					0	0	1	1/3	1/6	0.17
$\delta_{\mathcal{H}}$					0	1	1	2/3					0	1	1	2/3	4/6	0.67
$\tilde{\delta}_{\mathcal{H}}$					1	1	1	3/3					1	1	1	3/3	6/6	1.00
$\gamma_{\mathcal{H}}$					0	0	0	0/3					0	1	1	2/3	2/6	0.33
$\tilde{\gamma}_{\mathcal{H}}$					1	1	1	3/3					0	1	1	2/3	5/6	0.83
f_1									1	0	0	1/3	0	0	1	1/3	2/6	0.33
f_2									1	1	1	3/3	0	1	1	2/3	5/6	0.83
$\mu_{\mathcal{S}}$	0	0	0	0/3	0	0	0	0/3					0	0	0	0/3	0/9	0.00
$\sigma_{\mathcal{S}}$	0	0	0	0/3	0	0	1	1/3					0	0	1	1/3	2/9	0.22
$\delta_{\mathcal{S}}$	1	1	1	3/3	0	1	1	2/3					1	1	1	3/3	8/9	0.89
$\tilde{\delta}_{\mathcal{S}}$	1	1	1	3/3	0	1	1	2/3					1	1	1	3/3	8/9	0.89
$\gamma_{\mathcal{S}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
$\tilde{\gamma}_{\mathcal{S}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
f_3									1	0	0	1/3	0	0	0	0/3	1/6	0.17
f_4									1	1	1	3/3	1	1	1	3/3	6/6	1.00
$\mu_{\mathcal{R}}$	0	0	0	0/3	0	0	0	0/3					0	0	0	0/3	0/9	0.00
$\sigma_{\mathcal{R}}$	0	0	1	1/3	0	0	0	0/3					0	0	0	0/3	1/9	0.11
$\delta_{\mathcal{R}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
$\tilde{\delta}_{\mathcal{R}}$	1	1	1	3/3	1	1	1	3/3					0	1	1	2/3	8/9	0.89
$\gamma_{\mathcal{R}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
$\tilde{\gamma}_{\mathcal{R}}$	1	1	1	3/3	0	1	1	2/3					0	1	1	2/3	7/9	0.78
f_5									1	0	0	1/3	0	1	1	2/3	3/6	0.50
f_6									1	0	0	1/3	0	1	1	2/3	3/6	0.50
f_7									1	1	1	3/3	0	1	1	2/3	5/6	0.83
f_8									1	1	1	3/3	1	1	1	3/3	6/6	1.00
f_9									1	1	1	3/3	1	1	1	3/3	6/6	1.00
f_{10}									1	1	1	3/3	1	1	1	3/3	6/6	1.00

Table 9: Summary (along rows) of when each feature was chosen in the SFFS-based methods of Table 6: the standard SFFS (S), SFFS followed by Fisher Projection (SF), and SFFS followed by Fisher Projection using the Day Matrix (SFD). The features listed in the left-most column are grouped by signal: electromyogram ($\mathcal{E}$), blood-volume pressure ($\mathcal{B}$), heart rate ($\mathcal{H}$), skin conductivity ($\mathcal{S}$), and respiration ($\mathcal{R}$), for easier physiological interpretation. The totals at the right, when high, suggest that the feature may be a robust one regardless of the classification method.

potentially provide person-independent recognition.

Prior to our efforts, researchers have not reported the way in which physiological features for different emotions from the same day tend to be more similar than features for the same emotions on different days. This side-finding of our work may help explain why so many conflicting results exist on prior attempts to identify emotion-dependent physiological correlates. The day-dependent variations that we found and the ways we developed of handling them may potentially be useful in handling across-subject variations. We proposed methods of normalizing features and baselining, showing that the normalizing features gave the best results, but have the drawback of not being easily implemented in an online way because of the requirement of having session-long summarizing information.

We found that the Fisher Projection applied to a subset of features pre-selected by SFFS always outperformed Fisher Projection applied to the full set of features when assessing percentage of errors made by the classifiers. However, only a few of the improvements of SFFS-FP over FP were significant at $> 95\%$; the rest had confidences ranging from $58 - 87\%$. From the significant improvements, we surmise that the Fisher Projection's ability to transform the feature space may work better when poorer-performing features are omitted up front. The combination is synergistic and may well apply in other domains: feature selection followed by feature transformation.

Forty features were proposed and systematically evaluated by multiple algorithms. The best and worst features have been identified for this subject. These are of interest in the ongoing search for good features for affect recognition, and may aid in trying to understand the differential effects of emotions on physiology.

Although the precise rates found here can only be claimed to apply to one subject, the methodology developed in this paper can be used for any subject. The method is general for finding the features that work best for a given subject and for assessing classification accuracies for that subject.

The results of 81% recognition accuracy on eight categories of emotion are the only results we know of for such a classification, and are better than machine recognition rates of a similar number of categories of affect from speech (around 60%-70%) and almost as good as automated recognition of facial expressions (around 80-98%). Because there were doubts in the literature that physiological information shows any differentiation other than arousal level, these results are exciting. However, these results do not imply that a computer can detect and recognize your emotions with 81% accuracy, because of the limited set of emotions examined, because of the optimistic estimates of error given by the leave-one-out method, because of the fact that this experiment, like others in the literature, has only looked at forced choice among pre-segmented data, because of the use of only one subject's long term data, and because of the nature of true emotion, which consists of more than externally measurable signals. We expect that joint pattern analysis of signals from face, voice, body, and the surrounding situation is likely to give the most interesting emotion recognition results, especially since people read all these signals jointly.

The greater than six-times-chance recognition accuracy achieved in this pattern recognition research is nonetheless a significant finding; it informs long-debated questions by emotion theorists as to whether there is any physiological differentiation among emotions. Additionally, the finding of significant classification rates when the emotion labels are grouped either by arousal or by valence lends evidence against the belief that physiological signals only differentiate with respect to arousal; both valence and arousal were differentiated essentially equally by the method we developed.

A new question might be stated, "What is the precise bodily nature of the differentiation that exists among different emotions and upon what other factors does this differentiation depend?" We expect that physiological patterning, in combination with facial, vocal and other behavioral cues, will lead to significant improvements in machine recognition of user emotion over the coming decade, and that this recognition will be critical for giving machines the intelligence to adapt their behavior to interact more smoothly and respectfully with people. However, this "recognition" will be on measurable signals, which we do not expect to include one's innermost thoughts. Much work remains before emotion interpretation can occur at the level of human abilities.

6 Acknowledgments

The authors would like to thank Rob and Dan Gruhl for help with hardware for data collection, the MIT Shakespeare Ensemble for acting support, Anil Jain and Doug Zongker for providing us their SFFS code, Tom Minka for help with its use, and the anonymous reviewers for helping improve the paper.

7 References

[1] A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain.* New York, NY: Gosset/Putnam Press, 1994.

[2] J. LeDoux, *The Emotional Brain.* New York: Simon & Schuster, 1996.

[3] P. Salovey and J. D. Mayer, "Emotional intelligence," *Imagination, Cognition and Personality*, vol. 9, no. 3, pp. 185–211, 1990.

[4] D. Goleman, *Emotional Intelligence.* New York: Bantam Books, 1995.

[5] B. Reeves and C. Nass, *The Media Equation.* Cambridge University Press; Center for the Study of Language and Information, 1996.

[6] M. Sigman and L. Capps, *Children with Autism: A Developmental Perspective.* The Developing Child, Cambridge, Mass.: Harvard University Press, 1997.

[7] R. W. Picard, *Affective Computing.* Cambridge, MA: The MIT Press, 1997.

[8] J. Healey, R. W. Picard, and F. Dabek, "A new affect-perceiving interface and its application to personalized music selection," in *Proc. Wkshp on Perceptual User Interfaces*, (San Francisco), Nov 1998.

[9] R. W. Picard and J. Healey, "Affective wearables," *Personal Technologies*, vol. 1, no. 4, pp. 231–240, 1997.

[10] K. R. Scherer, "Ch. 10: Speech and emotional states," in *Speech Evaluation in Psychiatry* (J. K. Darby, ed.), pp. 189–220, Grune and Stratton, Inc., 1981.

[11] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[12] T. Polzin, *Detecting Verbal and Non-verbal Cues in the Communication of Emotions.* PhD thesis, School of Computer Science, June 2000.

[13] J. Hansen, 1999. Communication during ICASSP '99 panel on Speech Under Stress.

[14] J. N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality and Social Psychology*, vol. 37, pp. 2049–2058, 1979.

[15] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from log image sequences using optical flow," *IEEE T. Patt. Analy. and Mach. Intell.*, vol. 18, pp. 636–642, June 1996.

[16] I. Essa and A. Gardner, "Prosody analysis for speaker affect determination," in *Workshop on Perceptual User Interfaces '97*, (Banff, Alberta, Canada), pp. 45–46, October 19–21 1997.

[17] I. A. Essa, *Analysis, Interpretation and Synthesis of Facial Expressions.* PhD thesis, MIT Media Lab, Cambridge, MA, Feb. 1995.

[18] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding," *Psychophysiology*, vol. 36, pp. 35–43, 1999.

[19] M. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253–263, 1999.

[20] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE T. Patt. Analy.*

and Mach. Intell., vol. 21, pp. 974–989, October 1999.

[21] L. C. De-Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proc. IEEE Int. Conf. on Info., Comm. and Sig. Proc.*, (Singapore), pp. 397–401, Sept 1997.

[22] T. S. Huang, L. S. Chen, and H. Tao, "Bimodal emotion recognition by man and machine," in *ATR Workshop on Virtual Communication Environments*, (Kyoto, Japan), April 1998.

[23] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proc. 3rd Int. Conf. on Automatic Face and Gesture Recognition*, (Nara, Japan), pp. 366–371, IEEE Computer Soc., April 1998.

[24] J. T. Cacioppo and L. G. Tassinary, "Inferring psychological significance from physiological signals," *American Psychologist*, vol. 45, pp. 16–28, Jan. 1990.

[25] W. James, *William James: Writings 1878-1899*, ch. on Emotion, pp. 350–365. The Library of America, 1992. Originally published in 1890.

[26] W. B. Cannon, "The James-Lange theory of emotions: A critical examination and an alternative theory," *American Journal of Psychology*, vol. 39, pp. 106–124, 1927.

[27] S. Schachter, "The interaction of cognitive and physiological determinants of emotional state," in *Advances in Experimental Psychology* (L. Berkowitz, ed.), vol. 1, pp. 49–80, New York: Academic Press, 1964.

[28] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, pp. 1208–1210, Sep. 1983.

[29] W. M. Winton, L. Putnam, and R. Krauss, "Facial and autonomic manifestations of the dimensional structure of emotion," *Journal of Experimental Social Psychology*, vol. 20, pp. 195–216, 1984.

[30] A. J. Fridlund and C. E. Izard, "Electromyographic studies of facial expressions of emotions and patterns of emotions," in *Social Psychophysiology: A Sourcebook* (J. T. Cacioppo and R. E. Petty, eds.), pp. 243–286, New York: Guilford Press, 1983.

[31] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, and T. A. Ito, "The psychophysiology of emotion," in *Handbook of Emotions* (M. Lewis and J. M. Haviland-Jones, eds.), pp. 173–191, New York: The Guilford Press, 2000.

[32] D. Keltner and P. Ekman, "The psychophysiology of emotion," in *Handbook of Emotions* (M. Lewis and J. M. Haviland-Jones, eds.), pp. 236–249, New York: The Guilford Press, 2000.

[33] D. M. Clynes, *Sentics: The Touch of the Emotions.* Anchor Press/Doubleday, 1977.

[34] C. E. Izard, "Four systems for emotion activation: Cognitive and noncognitive processes," *Psychological Review*, vol. 100, no. 1, pp. 68–90, 1993.

[35] H. Hama and K. Tsuda, "Finger-pressure waveforms measured on clynes' sentograph distinguished among emotions," *Perceptual and Motor Skills*, vol. 70, pp. 371–376, 1990.

[36] H. Schlosberg, "Three dimensions of emotion," *Psychological Review*, vol. 61, pp. 81–88, Mar. 1954.

[37] P. J. Lang, "The emotion probe: Studies of motivation and attention," *American Psychologist*, vol. 50, no. 5, pp. 372–385, 1995.

[38] J. A. Healey, *Wearable and Automotive Systems for Affect Recognition from Physiology.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, May 2000. TR 526.

[39] E. Vyzas and R. W. Picard, "Affective pattern classification," in *AAAI 1998 Fall Symposium, Emotional and Intelligent: The Tangled Knot of Cognition*, (Orlando, FL), Oct. 1998.

[40] E. Vyzas and R. W. Picard, “Offline and online recognition of emotion expression from physiological data,” in *Workshop on Emotion-Based Agent Architectures, Third International Conf. on Autonomous Agents*, (Seattle, WA), pp. 135–142, May 1999.

[41] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1989.

[42] D. T. Lyyken, R. Rose, B. Luther, and M. Maley, “Correcting psychophysiological measures for individual differences in range,” *Psychophysiological Bulletin*, vol. 66, pp. 481–484, 1966.

[43] D. T. Lyyken and P. H. Venables, “Direct measurement of skin conductance: A proposal for standardization,” *Psychophysiology*, vol. 8, no. 5, pp. 656–672, 1971.

[44] M. E. Dawson, A. M. Schell, and D. L. Filion, “The electrodermal system,” in *Principles of Psychophysiology: Physical, social, and inferential elements* (J. T. Cacioppo and L. G. Tassinary, eds.), pp. 295–324, Cambridge: Cambridge Press, 1990.

[45] P. Pudil, J. Novovicova, and J. Kittler, “Floating search methods in feature selection,” *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, Nov. 1994.

[46] A. Jain and D. Zongker, “Feature selection: Evaluation, application, and small sample performance,” *IEEE T. Patt. Analy. and Mach. Intell.*, vol. 19, pp. 153–163, February 1997.

[47] P. Pudil, J. Novovicova, and J. Kittler, “Simultaneous learning of decision rules and important attributes for classification problems in image analysis,” *Image and Vision Comp.*, vol. 12, pp. 193–198, Apr. 1994.

[48] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* Wiley-Interscience, 1973.

[49] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE T. Info. Theory*, vol. IT-13, pp. 21–27, Jan. 1967.

[50] J. Healey and R. W. Picard, “Digital processing of affective signals,” in *IEEE Int. Conf. on Acoust., Sp. and Sig. Proc.*, (Seattle), 1998.

[51] J. Scheirer, J. Klein, R. Fernandez, and R. W. Picard, “Frustrating the user on purpose: A step toward building an affective computer,” *Interaction with Computers*, 2001. To appear.