Affective Computing

Maja Panitc
Delft University of Technology, The Netherlands

INTRODUCTION

We seem to be entering an era of enhanced digital connectivity. Computers and the Internet have become so embedded in the daily fabric of people's lives that they simply cannot live without them (Hoffman et al, 2004). We use this technology to work, to communicate, to shop, to seek out new information, and to entertain ourselves. With this ever-increasing diffusion of computers in society, human-computer interaction (HCI) is becoming increasingly essential to our daily lives.

HCI design was dominated first by direct manipulation and then delegation. The tacit assumption of both styles of interaction has been that the human will be explicit, unambiguous, and fully attentive while controlling the information and command flow. Boredom, preoccupation, and stress are unthinkable, even though they are very human behaviors. This insensitivity of current HCI designs is fine for well-codified tasks. It works for making plane reservations, buying and selling stocks, and, as a matter of fact, almost everything we do with computers today. But this kind of categorical computing is inappropriate for design, debate, and deliberation. In fact, it is the major impediment to having flexible machines capable of adapting to their users and their level of attention, preferences, moods, and intentions.

The ability to detect and understand affective states of a person with whom we are communicating is the core of emotional intelligence. Emotional intelligence (EQ) is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life (Goleman, 1995). When it comes to computers, however, not all of them will need emotional intelligence, and none will need all of the related skills that we need. Yet man-machine interactive systems capable of sensing stress, inattention, and heedfulness, and capable of adapting and responding appropriately to these affective states of the user are likely to be perceived as more natural, more efficacious and more trustworthy. The research area of machine analysis and employment of human affective states to build more natural, flexible HCI goes by a general name of affective computing, introduced first by Picard (1997).

BACKGROUND: RESEARCH MOTIVATION

Besides the research on natural, flexible HCI, various research areas and technologies would benefit from efforts to model human perception of affective feedback computationally. For instance, automatic recognition of human affective states is an important research topic for video surveillance as well. Automatic assessment of boredom, inattention, and stress will be highly valuable in situations where firm attention to a crucial but perhaps tedious task is essential, such as aircraft control, air traffic control, nuclear power plant surveillance, or simply driving a ground vehicle like a truck, train, or car. An automated tool could provide prompts for better performance, based on the sensed user's affective states.

Another area that would benefit from efforts toward computer analysis of human affective feedback is the automatic affect-based indexing of digital visual material. A mechanism for detecting scenes or frames that contain expressions of pain, rage, and fear could provide a valuable tool for violent-content-based indexing of movies, video material, and digital libraries.

Other areas where machine tools for analysis of human affective feedback could expand and enhance research and applications include specialized areas in professional and scientific sectors. Monitoring and interpreting affective behavioral cues are important to lawyers, police, and security agents who are often interested in issues concerning deception and attitude. Machine analysis of human affec-

Table 1. The main problem areas in the research on affective computing

What is an affective state? This question is related to psychological issues pertaining to the nature of affective states and the way affective states are to be described by an automatic analyzer of human affective states.
What kinds of evidence warrant conclusions about affective states? In other words, which human communicative signals convey messages about an affective arousal? This issue shapes the choice of different modalities to be integrated into an automatic analyzer of affective feedback.
How can various kinds of evidence be combined to generate conclusions about affective states? This question is related to neurological issues of human sensory-information fusion, which shape the way multi-sensory data is to be combined within an automatic analyzer of affective states.

tive states could be of considerable value in these situations where only informal interpretations are now used. It would also facilitate research in areas such as behavioral science (in studies on emotion and cognition), anthropology (in studies on cross-cultural perception and production of affective states), neurology (in studies on dependence between emotional abilities impairments and brain lesions), and psychiatry (in studies on schizophrenia) in which reliability, sensitivity, and precision are persisting problems.

BACKGROUND: THE PROBLEM DOMAIN

While all agree that machine sensing and interpretation of human affective information would be quite beneficial for manifold research and application areas, addressing these problems is not an easy task. The main problem areas are listed in Table 1.

On one hand, classic psychological research follows from the work of Darwin and claims the existence of six basic expressions of emotions that are universally displayed and recognized: happiness, anger, sadness, surprise, disgust, and fear (Lewis & Haviland-Jones, 2000). In other words, all non-verbal communicative signals (i.e., facial expression, vocal intonations, and physiological reactions) involved in these basic emotions are displayed and recognized cross-culturally. On the other hand, there is now a growing body of psychological research that strongly challenges the classical theory on emotion. Russell (1994) argues that emotion in general can best be characterized in terms of a multi-dimensional affect space, rather than in terms of a small number of emotion categories. Social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations and that they do not explain the genuine feeling (affect). Also, there is no consensus on how affective displays should be labeled (Wierzbicka, 1993). The main issue here is that of culture dependency; the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent (Matsumoto, 1990). In summary, it is not certain that each of us will express a particular affective state by modulating the same communicative signals in the same way, nor is it certain that a particular modulation of interactive cues will be interpreted always in the same way independent of the situation and the observer. The immediate implication is that pragmatic choices (e.g., application- and user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback.

Affective arousal modulates all verbal and non-verbal communicative signals (Ekman & Friesen, 1969). Hence, one could expect that automated human-affect analyzers should include all human interactive modalities (sight, sound, and touch) and should analyze all non-verbal interactive signals (facial expressions, vocal expressions, body gestures, and physiological reactions). Yet the reported research does not confirm this assumption. The visual channel carrying facial expressions and the auditory channel carrying vocal intonations are widely thought of as most important in the human recognition of affective feedback. According to Mehrabian

Table 2. The characteristics of an ideal automatic human-affect analyzer

multimodal (modalities: facial expressions, vocal intonations)
robust and accurate (despite auditory noise, occlusions, and changes in viewing and lighting conditions)
generic (independent of variability in subjects' physiognomy, sex, age, and ethnicity)
sensitive to the dynamics (time evolution) of displayed affective expressions (performing temporal analysis of the sensed data, previously processed in a joint feature space)
context-sensitive (performing application- and task-dependent data interpretation in terms of user-profiled affect-interpretation labels)

(1968), whether the listener feels liked or disliked depends on 7% of the spoken word, 38% on vocal utterances, and 55% on facial expressions. This indicates that while judging someone's affective state, people rely less on body gestures and physiological reactions displayed by the observed person; they rely mainly on facial expressions and vocal intonations. Hence, automated affect analyzers should at least combine modalities for perceiving facial and vocal expressions of affective states.

Humans simultaneously employ the tightly coupled modalities of sight, sound, and touch. As a result, analysis of the perceived information is highly robust and flexible. Hence, in order to accomplish a multimodal analysis of human interactive signals acquired by multiple sensors, which resembles human processing of such information, input signals cannot be considered mutually independent and cannot be combined only at the end of the intended analysis, as the majority of current studies do. The input data should be processed in a joint feature space and according to a context-dependent model (Pantic & Rothkrantz, 2003).

In summary, an ideal automatic analyzer of human affective information should be able to emulate at least some of the capabilities of the human sensory system (Table 2).

THE STATE OF THE ART

Facial expressions are our primary means of communicating emotion (Lewis & Haviland-Jones, 2000), and it is not surprising, therefore, that the majority of efforts in affective computing concern automatic analysis of facial displays. For an exhaustive survey of studies on machine analysis of facial affect, the readers are referred to Pantic and Rothkrantz (2003). This survey indicates that the capabilities of currently existing facial affect analyzers are rather limited (Table 3). Yet, given that humans detect six basic emotional facial expressions with an accuracy ranging from 70% to 98%, it is rather significant that the automated systems achieve an accuracy of 64% to 98% when detecting three to seven emotions deliberately displayed by five to 40 sub-

Table 3. Characteristics of currently existing automatic facial affect analyzers

handle a small set of posed prototypic facial expressions of six basic emotions from portraits or nearly-frontal views of faces with no facial hair or glasses, recorded under good illumination.
do not perform a task-dependent interpretation of shown facial behavior; yet, a shown facial expression may be misinterpreted if the current task of the user is not taken into account (e.g., a frown may be displayed by the speaker to emphasize the difficulty of the currently discussed problem, and it may be shown by the listener to denote that he did not understand the problem at issue).
do not analyze extracted facial information on different time scales (proposed inter-video-frame analyses are usually used to handle the problem of partial data); consequently, automatic recognition of the expressed mood and attitude (longer time scales) is still not within the range of current facial affect analyzers.

jects. An interesting point, nevertheless, is that we cannot conclude that a system achieving a 92% average recognition rate performs better than a system attaining a 74% average recognition rate when detecting six basic emotions from face images. Namely, in spite of repeated references to the need for a readily accessible reference set of images (image sequences) that could provide a basis for benchmarks for efforts in automatic facial affect analysis, no database of images exists that is shared by all diverse facial-expression-research communities.

If we consider the verbal part (strings of words) only, without regard to the manner in which it was spoken, we might miss important aspects of the pertinent utterance and even misunderstand the spoken message by not attending to the non-verbal aspect of the speech. Yet, in contrast to spoken language processing, which has witnessed significant advances in the last decade, vocal expression analysis has not been widely explored by the auditory research community. For a survey of studies on automatic analysis of vocal affect, the readers are referred to Pantic and Rothkrantz (2003). This survey indicates that the existing automated systems for auditory analysis of human affect are quite limited (Table 4). Yet humans can recognize emotion in a neutral-content speech with an accuracy of 55% to 70% when choosing from among six basic emotions, and automated vocal affect analyzers match this accuracy when recognizing two to eight emotions deliberately expressed by subjects recorded while pronouncing sentences having a length of one to 12 words. Similar to the case of automatic facial affect analysis, no readily accessible reference set of speech material exists that could provide a basis for benchmarks for efforts in automatic vocal affect analysis.

Relatively few of the existing works combine different modalities into a single system for human affective state analysis. Examples are the works of Chen and Huang (2000), De Silva and Ng (2000), Yoshitomi et al. (2000), Go et al. (2003), and Song et al. (2004), who investigated the effects of a combined detection of facial and vocal expressions of affective states. In brief, these studies assume clean audiovisual input (e.g., noise-free recordings, closely-placed microphone, non-occluded portraits) from an actor speaking a single word and displaying exaggerated facial expressions of a basic emotion. Though audio and image processing techniques in these systems are relevant to the discussion on the state of the art in affective computing, the systems themselves have all (as well as some additional) drawbacks of single-modal affect analyzers and, in turn, need many improvements, if they are to be used for a multimodal context-sensitive HCI, where a clean input from a known actor/announcer cannot be expected and a context-independent data interpretation does not suffice.

CRITICAL ISSUES

Probably the most remarkable issue about the state of the art in affective computing is that, although the

Table 4. Characteristics of currently existing automatic vocal affect analyzers

perform singular classification of input audio signals into a few emotion categories such as anger, irony, happiness, sadness/grief, fear, disgust, surprise, and affection.
do not perform a context-sensitive analysis (i.e., application-, user-, and task-dependent analysis) of the input audio signal.
do not analyze extracted vocal expression information on different time scales (proposed inter-audio-frame analyses are used either for the detection of supra-segmental features, such as the pitch and intensity over the duration of a syllable, word, or sentence, or for the detection of phonetic features)—computer-based recognition of moods and attitudes (longer time scales) from input audio signal remains a significant research challenge.
adopt strong assumptions to make the problem of automating vocal-expression analysis more tractable (e.g., the recordings are noise-free, the recorded sentences are short, delimited by pauses, carefully pronounced by non-smoking actors to express the required affective state) and use the test data sets that are small (one or more words or one or more short sentences spoken by few subjects) containing exaggerated vocal

recent advances in video and audio processing make audiovisual analysis of human affective feedback tractable, and although all agreed that solving this problem would be extremely useful, merely a couple of efforts toward the implementation of such a bimodal human-affect analyzer have been reported to date.

Another issue concerns the interpretation of audiovisual cues in terms of affective states. The existing work employs usually singular classification of input data into one of the basic emotion categories. However, pure expressions of basic emotions are seldom elicited; most of the time, people show blends of emotional displays. Hence, the classification of human non-verbal affective feedback into a single basic-emotion category is not realistic. Also, not all non-verbal affective cues can be classified as a combination of the basic emotion categories. Think, for instance, about the frustration, stress, skepticism, or boredom. Furthermore, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state may differ from culture to culture and even from person to person. Hence, the definition of interpretation categories in which any facial and/or vocal affective behavior, displayed at any time scale, can be classified is a key challenge in the design of realistic affect-sensitive monitoring tools. One source of help is machine learning; the system potentially can learn its own expertise by allowing the user to define his or her own interpretation categories (Pantic, 2001).

Accomplishment of a human-like interpretation of sensed human affective feedback requires pragmatic choices (i.e., application-, user- and task-profiled choices). Nonetheless, currently existing methods aimed at the automation of human-affect analysis are not context sensitive. Although machine-context sensing (i.e., answering questions like who is the user, where is the user, and what is the user doing) has witnessed recently a number of significant advances (Pentland, 2000), the complexity of this problem makes context-sensitive human-affect analysis a significant research challenge.

Finally, no readily accessible database of test material that could be used as a basis for benchmarks for efforts in the research area of automated human affect analysis has been established yet. In fact, even in the research on facial affect analysis, which attracted the interest of many researchers, there is a glaring lack of an existing benchmark face database. This lack of common testing resources forms the major impediment to comparing, resolving, and extending the issues concerned with automatic human affect analysis and understanding. It is, therefore, the most critical issue in the research on affective computing.

CONCLUSION

As remarked by scientists like Pentland (2000) and Oviatt (2003), multimodal context-sensitive (user-, task-, and application-profiled and affect-sensitive) HCI is likely to become the singlemost widespread research topic of the AI research community. Breakthroughs in such HCI designs could bring about the most radical change in the computing world; they could change not only how professionals practice computing, but also how mass consumers conceive and interact with the technology. However, many aspects of this new-generation HCI technology, in particular ones concerned with the interpretation of human behavior at a deeper level and the provision of the appropriate response, are not mature yet and need many improvements.

REFERENCES

Chen, L.S., & Huang, T.S. (2000). Emotional expressions in audiovisual human computer interaction. Proceedings of the International Conference on Multimedia and Expo.

De Silva, L.C., & Ng, P.C. (2000). Bimodal emotion recognition. *Proeedings of the. International Conference on Face and Gesture Recognition*.

Ekman, P., & Friesen, W.F. (1969). The repertoire of nonverbal behavioral categories—Origins, usage, and coding. *Semiotica, 1*, 49-98.

Go, H.J., Kwak, K.C., Lee, D.J., & Chun, M.G. (2003). Emotion recognition from facial image and speech signal. *Proceedings of the Conference of the Society of Instrument and Control Engineers*.

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.

Hoffman, D.L., Novak, T.P., & Venkatesh, A. (2004). Has the Internet become indispensable? *Communications of the ACM, 47*(7), 37-42.

Lewis, M., & Haviland-Jones, J.M. (Eds.). (2000). *Handbook of emotions*. New York: Guilford Press.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emotion, 14*, 195-214.

Mehrabian, A. (1968). Communication without words. *Psychology Today, 2*(4), 53-56.

Oviatt, S. (2003). User-centered modeling and evaluation of multimodal interfaces. *Proceedings of the IEEE, 91*(9), 1457-1468.

Pantic, M. (2001). Facial expression analysis by computational intelligence techniques [Ph.D. thesis]. Delft, Netherlands: Delft University of Technology.

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, 9*1(9), 1370-1390.

Pentland, A. (2000). Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 107-119.

Picard, R.W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? *Psychological Bulletin, 115*(1), 102-141.

Song, M., Bu, J., Chen, C., & Li, N. (2004). Audio-visual based emotion recognition—A new approach. *Proceedings of the International Conference Computer Vision and Pattern Recognition*.

Wierzbicka, A. (1993). Reading human faces. *Pragmatics and Cognition, 1*(1), 1-23.

Yoshitomi, Y., Kim, S., Kawano, T., & Kitazoe, T. (2000). Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. *Proceedings of the International Workshop on Robot-Human Interaction*, (178-183).

KEY TERMS

Affective Computing: The research area concerned with computing that relates to, arises from, or deliberately influences emotion. Affective computing expands HCI by including emotional communication, together with the appropriate means of handling affective information.

Benchmark Audiovisual Affect Database: A readily accessible centralized repository for retrieval and exchange of audio and/or visual training and testing material and for maintaining various test results obtained for a reference audio/visual data set in the research on automatic human affect analysis.

Context-Sensitive HCI: HCI in which the computer's context with respect to nearby humans (i.e., who the current user is, where the user is, what the user's current task is, and how the user feels) is automatically sensed, interpreted, and used to enable the computer to act or respond appropriately.

Emotional Intelligence: A facet of human intelligence that includes the ability to have, express, recognize, and regulate affective states, employ them for constructive purposes, and skillfully handle the affective arousal of others. The skills of emotional intelligence have been argued to be a better predictor than IQ for measuring aspects of success in life.

Human-Computer Interaction (HCI): The command and information flow that streams between the user and the computer. It is usually characterized in terms of speed, reliability, consistency, portability, naturalness, and users' subjective satisfaction.

Human-Computer Interface: A software application, a system that realizes human-computer interaction.

Multimodal (Natural) HCI: HCI in which command and information flow exchanges via multiple natural sensory modes of sight, sound, and touch. The user commands are issued by means of speech, hand gestures, gaze direction, facial expressions, and so forth, and the requested information or the computer's feedback is provided by means of animated characters and appropriate media.