Affective computing: problems, reactions and intentions

R.D. Ward\textsuperscript{a,}*\textsuperscript{,} P.H. Marsden\textsuperscript{b}

\textsuperscript{a}Department of Behavioural Sciences, University of Huddersfield, Huddersfield, West Yorkshire, HD1 3DH, UK
\textsuperscript{b}Department of Multimedia and Information Systems, University of Huddersfield, West Yorkshire, Huddersfield HD1 3DH, UK

Available online 10 July 2004

Abstract

Although we share the optimistic vision of affective computing presented in Interacting with Computers 14(2), we question the extent to which affective sensing can support the kinds of applications proposed in the literature. These applications depend upon the detection of affective reactions to HCI situations and events, but it has yet to be shown that such reactions can reliably be detected in subtle and natural situations. We also point out that, in human–human interaction, intentional communicative affect is both easier to recognise and more important than reactive affect. We suggest exploration of this idea may lead to more fruitful applications of affective computing.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Computing; Psychophysiology; Facial expression

Interacting with Computers 14(2) makes an important contribution to the future direction of human–computer interaction. We share the authors’ vision of software able to interact with users in sophisticated, human-like, social and emotional ways, although, as Picard and Klein (2002) demonstrate, the debate about its implications and applications has hardly begun.

In the vein of Picard (1997 pp. 93–94) and Conati (2002) and others, let us begin by imagining a humanoid 3D virtual teacher able to participate in learning dialogues involving both verbal and non-verbal communication. The tutor would be able to make judgments about relevant aspects of learners’ affective states, such as their levels of interest, motivation, understanding, enthusiasm, and other factors accepted as important in

* Corresponding author.

E-mail address: r.d.ward@hud.ac.uk (R.D. Ward).

0953-5438/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
learning. In responding to learners, the tutor would conform to non-verbal conventions such as the maintenance of eye contact, and provide feedback through changes in its own facial expressions. Strategies such as active listening, emotional regulation, acknowledgement of users’ emotions and support for negative emotions, as discussed by Klein et al. (2002), would form part of the tutoring process. The tutor would be able to make use of a wide range of non-verbal information such as eye gaze, pupil size, tone of voice, prosody, posture, body movements and facial expressions, just as human tutors supposedly do. Other sources of information such as skin conductance, muscle potential, cardiovascular measures and brain potential might also be utilised.

From a computational point of view, all these sources are physiological signals with common difficulties of analysis and interpretation. Thus, although the paper by Scheirer et al. (2002) is concerned principally with measures of autonomic arousal, in this case skin conductance and blood volume pressure, the problems it addresses and recommendations it presents are fundamental to this vision of the future of human–computer interaction. On the face of things, some progress seems to have been made. Scheirer et al. adopted a deliberately flawed interface designed to induce user frustration. Experimental participants played a computer game in which they were required to complete a series of visual puzzles as quickly and accurately as possible in order to win a prize of $100. At random points the game was programmed to freeze for several seconds when participants clicked to proceed to the next puzzle, as might happen with a faulty mouse or system problem. Participants’ skin conductance and blood volume pressure were monitored throughout, and analysis of the signals using a pattern recognition technique was able to discriminate between periods when freezing of the game was and was not present. The paper therefore demonstrates software-induced frustration and its automatic detection by means of autonomic measures of arousal.

Our own investigations have demonstrated similar software-induced changes in levels of arousal. In web-based tasks, we have found surprise events such as the sudden appearance of an alert box accompanied by a standard sound, to produce statistically significant increases in users’ skin conductance. We have also found similarly significant responses to content designed to evoke an affective reaction such as amusement (Ward et al., 2002, 2003). These findings provide support for the periodically proposed idea that physiological measurement might be of use in evaluating usability (e.g. Wastell and Newman, 1996) or in driving interaction (e.g. Allanson et al., 1999). In the evaluation of usability, physiological readings might help identify features that users fail to mention in feedback reports, either because they have forgotten or because they thought it insignificant.

However, traditional psychophysicologists tend to be unimpressed by these findings. Nor would they be impressed by the reactions to Scheirer’s faulty mouse. These effects are entirely typical of numerous well-replicated studies of responses to a large variety of stimuli reported in the literature over many decades (e.g. see Andreassi, 2000). All that can be said is that the findings demonstrate that human physiology does respond as might be expected to events that take place during real human–computer interaction.

We should also be concerned that most of our findings relate to relatively strong stimuli such as the surprise appearance of an alert box, and especially the failure of an interface action as common as the clicking of a ‘next’ button when users are working under
pressure. These stimuli represent major interface events and gross failures of usability, and it is hardly surprising that users react strongly to them. If psychophysiology and related sensing techniques are to be of widespread use in evaluating usability or in driving interaction, they need to be able to detect users’ reactions to far more subtle stimuli than these. For example, in evaluating web site usability, designers might be interested in determining which of several alternative designs is most likely to attract or maintain readers’ attention.

The literature contains many experimental findings to suggest that psychophysiology should be of help in such an evaluation. To select an example almost at random, participants in an experiment by Zimny and Weidenfeller (1963) listened to three different pieces of music previously judged to be exciting (the New World Symphony), or calming (Air on the G String), or neutral (Les Sylphides), and showed significantly greater skin conductances in response to the exciting piece.

Unfortunately, in investigating physiological reactions to more subtle aspects of web site design, we have not always been able to find consistent statistically significant effects. In one investigation, we examined participants’ responses to well-designed and ill-designed web sites. Two versions of a web site were developed, both linking to the same information. The ill-designed site was constructed so as deliberately to break a large number of generally accepted web design guidelines by employing devices such as pull-down lists to obscure the structure of the underlying information, impoverished navigational cues necessitating excessive scrolling, gratuitous use of typefaces, colours and animation, and the sudden appearance of advertisements in pop-up windows which had to be closed before the user could proceed. The well-designed site was constructed as far as possible to follow good web design guidelines. In theory, it should have been easy to demonstrate that the two web sites had different physiological effects, but this did not turn out to be the case (Ward and Marsden, 2003).

Similarly in an investigation of physiological reactions to content, participants completed a web-based quiz based upon the UK driving theory test. Using questions from the real theory test, the quiz first presented a series of questions judged to be factual and unexciting, concerned mainly with details of road signs and markings, speed restrictions, and so on. This was then followed by a sequence of questions judged to have high affective content. These questions, also from the real theory test, addressed issues such as how to deal with serious accident injuries (e.g. should you ‘burst the blisters’ caused by burns) and what to do when upset by the bad behaviour of other drivers (e.g. should you ‘stop and take a rest’ or ‘wave your fist and shout abusive language’). Although some interface events in the quiz did tend to produce consistently significant physiological responses, such as delivery of the pass/fail test outcome, the sequence of emotive questions was less reliable in producing responses (Ward et al., 2002, 2003).

It would appear that the stronger the stimulus, the more likely it is to produce a detectable physiological response. The detection of responses to weaker stimuli requires greater experimental control. Responses to relatively weak stimuli are common in the psychophysiological literature, but these have been observed almost entirely in rigorously controlled experimental settings. It is worth considering what this typically involves. Participants enter the experimental room and, after attachment of sensors and receipt of an explanation of the purpose of the experiment and its procedure, are then required to sit
quietly for a settling-down period of perhaps 15 min in order to stabilise their physiological readings to baseline levels. Attachment of sensors sometimes involves abrasion of the skin and the use of conductivity gels. In some cases the temperature and humidity of the experimental room is controlled in order to hold constant the rate at which moisture evaporates from the skin. Experimental procedures are designed so that there are clear experimental stimuli such that any observed changes in physiological readings can be attributable only to those stimuli. Experimental design ensures there are sufficient numbers of participants for meaningful statistical analysis, and as far as possible of course that the experimental results are statistically significant.

Our difficulty in demonstrating statistically significant responses to weak HCI events reflects more of an unwillingness rather than an inability to undertake tightly controlled experimentation. In our own investigations we adopted relatively short settling-in periods of no more than 5 min, used dry electrodes in an ordinary room in which temperature and humidity varied from day to day with the weather, and asked participants to undertake relatively complex tasks that used ordinary software rather than clear distinct stimuli. Tighter experimental control may well have produced statistically significant differences. However, it seems unlikely that real-life usability testing would normally be taken to such lengths. The use of physiology to inform interaction would take place under still less standardised conditions. Everyday human–computer interaction is not easily controlled. There must be some doubt as to whether it can be effectively controlled at all. Software is complex and contains large numbers of potentially confounding elements. To control or remove them would be likely to change the essential look, feel and functionality of the software. There is also the difficulty of standardising the human elements in the interaction. How can an experimenter in comparing the effects of two web sites ensure that all participants are similarly motivated? The reactions of participants who do not really care about the task in hand are likely to be different from those who do. Extrinsic motivation such as the opportunity to win a cash prize, or intrinsic motivation such as the potential embarrassment of failing a driving theory test may be sufficient to motivate participants in certain situations, but the inclusion of motivators is not always possible. If physiological and similar measurements are to be useful in human–computer interaction research, they need not only to be able to identify reactions to subtle events, but also to be able to do so under these kinds of loosely controlled situations.

There are a number of other problems well documented in the literature (Andreassi, 2000; Idzikowski and Baddeley, 1983), which raise difficulties of methodology and analysis. Physiological readings tend to be inconsistent. Different readings may provide contradictory indications. There are large differences between individuals in their degree of physiological response. Even within individuals, the same individual under the same experimental conditions may give very different skin conductance readings on different occasions (much of this being due to changes in temperature and humidity affecting rates at which moisture evaporates from the skin). The distinction between what are and what are not significant features in the data signals can involve somewhat arbitrary assumptions about the magnitude, latency and duration of changes. Scheirer et al. (2002) address these problems by establishing standard signal patterns known as ground truths against which to compare and hence classify patterns in the actual signals.
Then there is the problem of interpretation. Assuming that a reaction of some kind has
definitely been identified in the physiological signals, its actual meaning can still be far
from clear. In usability testing, we might not be too concerned with this, it may be
sufficient to know that something has occurred to alter arousal levels or some other aspect
of a user’s physiology, and steps can then be taken to find out what that something is.
However, if we want to go further and speculate about emotional experiences then we run
into difficulties. Again, even a cursory examination of the psychophysiological literature
illustrates just how difficult this might be. Experimental findings frequently seem to
contradict each other. For example, Turner et al. (1990) observed consistent increases in
heart rate as task difficulty increased across a variety of cognitive tasks. On the other hand,
in verbal learning tasks, Andreassi (1966) observed that heart rates were higher when
participants were learning easier lists of words. These apparently contradictory findings
can be explained by considering the nature of the materials and tasks involved in the
experiments. Andreassi concluded that the higher heart rates in the easier tasks reflected
increased involvement and the motivating effects of good performance in an otherwise
boring task. But taken together these experiments illustrate the problem of attributing
internal experiences from external signals. Does higher heart rate indicate increased or
decreased mental effort? It depends on the circumstances.

Scheirer et al. (2002) do acknowledge this problem and suggest a need for systems
that also make use of visual and auditory information in making judgments about users’
affect. Recently, we have begun to investigate the possible uses in usability testing and
other applications of the automatic detection of users’ facial movements in video
signals (Ward et al., 2003). Facial expression recognition neatly illustrates almost
exactly the same issues and problems as arise in psychophysiology. There are
indications that facial expressions show reactions not apparent in the physiological data
and vice versa. Some users have high facial expressivity, others very little. This does
not mean that users with high expressivity experience stronger emotions, or that those
with low expressivity experience weak emotions. In fact, according to Kappas (2002),
no current facial activity researcher holds that all emotions are always reflected in
facial expressions, and they almost all agree that facial expressions may not be
associated with any emotions at all. Any attempt to determine underlying affective
states from facial activity is therefore affected by oversimplifications of the available
empirical data.

But we should not be as pessimistic as the above may imply. Despite the difficulties one
always returns to the fact that, every day, human beings successfully make judgments
about others’ affective states. We do it from limited information carried in non-verbal cues
such as facial expression, posture and tone of voice. Our perception of these signals is
limited too. In fact, human beings may not always judge the emotions of others very
well—a vast army of counsellors, coaches and therapists has grown up to help us develop
these skills. Amongst the strategies they teach, is to share and to seek verification of our
judgments through dialogue. Without such dialogue some of the things we seem to expect
of psychophysiology would be very difficult for real people. How would we judge by
observation alone whether a user preferred the design of one web site rather than another?
But to the extent that human beings can do these things then it should be possible to get
computers to do them, or at least something like them.
The philosophy of this new approach to human–computer interaction as represented by the papers in Interacting with Computers 14(2) is principally concerned with machine perception of users’ affective reactions. There is an assumption that users respond to content and events occurring during human–computer interaction, and that the machine detects these responses and adapts its behaviour accordingly. To an extent this may be making a similar mistake to that made in the intelligent tutoring systems community some years ago, the mistake of assuming the machine is in charge. It can be argued that this is not how emotion operates in the real world. We need to acknowledge that in human–human interaction, affect has an intentional communicative function. It serves the purpose of managing collaborations with others through the expression of emotions such as assertiveness, affection and anxiety, with its roots in infant attachment behaviour (Oatley, 2000). These affect communications would also seem to be very important in managing adult collaborations such as those between learners and teachers, and could have a role in human–computer dialogues such as use of the kind of virtual teacher imagined at the beginning of this response. It may be that where users work with present everyday software alone and in private, their affective reactions will be limited and difficult to detect. Multiple data sources, and complex techniques may be required for meaningful interpretation, but reliability may still be low. However, if users are given freedom to participate in a form of human–computer interaction in which affect may be used intentionally, purposefully and communicatively, it seems likely that users would emphasise and accentuate the expression of these emotions in order to ensure their intended meaning is clear and understood. These expressions of affect would consequently be easier for a machine to detect and recognise. Emotion then begins to provide a basis for negotiation and transaction, with more equal balance of initiative between machine and user. Creative exploration of this idea may lead to more fruitful applications of affective computing.

Conclusions

If physiological measurement is to be useful in human–computer interaction, in the ways currently envisaged in the literature, it has to be able to identify reactions to subtle events, not just major failures of interaction.

Similarly, physiological measurement has to be able to detect these reactions in loosely controlled naturalistic situations representative of real computer use, rather than tightly controlled laboratory settings.

Psychophysiological data is very noisy, making cause and effect difficult to demonstrate.

Even where there is clear cause and effect, interpretation in terms of users’ internal mental processes and experiences presents serious further problems.

These difficulties are common to a wide range of physiological signals, including those in the visual and auditory domains.

Human beings also find it difficult to recognise others’ affective states, yet often still do so effectively. They are aided by the social and interpersonal context of affective
communication. The implications of this for affective computing require further consideration.

Acknowledgements

Parts of the work referred to above is supported by EPSRC project grant GR/N00586 and the University of Huddersfield.

References


