Symmetric Multimodality Revisited: Unveiling Users' Physiological Activity

Helmut Prendinger and Mitsuru Ishizuka, *Member, IEEE*

***Abstract*—In this paper, we describe our own stance on a research area called "Humatronics," which aims at establishing a (more) symmetric interaction relationship between humans and computer systems. In particular, we will advocate a novel approach to understanding humans that is based on largely involuntary and unconscious physiological information and gaze behavior rather than purposeful and conscious actions or behaviors. "Understanding humans" here refers to users' states related to emotion and affect, attention and interest, and possibly even to their intentions. A key feature of our approach is that it provides insight into a person's cognitive-motivational state without relying on cognitive judgements, such as answers to dedicated queries. Lifelike interface agents are endowed with synthetic bodies and faces and can be considered as prime candidates for outbalancing the asymmetric relationship in current human–computer interaction. As example applications, we will report on two recent studies that utilized lifelike agents as presenters or interaction partners of users. The resulting interactions can be conceived as implementing initial steps toward symmetric multimodality in user interfaces.**

***Index Terms*—User interface human factors.**

I. Introduction and Motivation

The notion of symmetric multimodality has been introduced for dialogue systems that have all the input modes, especially of speech, gesture, and facial expression also available for output [1], and may thus significantly improve the intuitiveness and naturalness of the interaction between humans and computers. Similarly, the field of "Humatronics," to which this volume is dedicated, aims at balancing the asymmetry of the relationship between humans and computers.

A salient feature of a dialogue system with symmetric multimodality is that it requires the representation and processing of both the user's multimodal input and the output of the computer system. On the *input* side, multiple modalities have to be integrated and synchronized, and possibly disambiguated [2]. A person might utter a sentence expressing a joyful experience with a "happy" facial display. In this case, the combined interpretation of speech and facial display leads to a higher probability of the person being in an affective state of "joy" than if both modalities were interpreted individually. On the other hand, a sentence like "That's wonderful" uttered with a facial expression indicating disgust might refer to a sarcastic statement. Keeping in mind that the majority of human utterances is ambiguous (e.g., with respect to their affective meaning), the face modality in this example allows us to disambiguate the valence of the utterance, and hence, contributes to the correct understanding of the user's affective or emotional state.

A different issue is how a computer can display multiple modalities on the *output* (presentation) side. While common computers typically do not provide rich human-style output modalities, recent research in the area of lifelike characters (or embodied conversational agents) advocates the use of virtual interface agents as communication partners of users [3]. Those agents are endowed with virtual (graphical) bodies, faces, and synthetic speech and may thus emulate natural human–human communication. By implementing synchronized conversational gestures, facial expression, and speech [4], lifelike characters allow users to follow social interaction protocols similar to human conversation (for technologies to implement agent-based interactions, see our work on multimodal presentation markup language [5], [6]).

In this paper, we want to focus on interaction modalities that have hitherto received less attention, but seem highly promising for giving computers the capability of *understanding* humans—a theme at the core of Humatronics research. Specifically, we will discuss human physiological activity such as biosignals and eye movements. (A good overview of other modalities such as speech and facial expression can be found in [7].) A salient feature of physiological information consists in its unconscious and nondeliberate nature, which makes it particularly apt to reveal the "true" experience of humans. In daily interactions, humans do not consciously control, e.g., their skin conductance level, heart rate, or pupil size. Certainly, in some situations, such as a biofeedback session [8] or playing the game "Relax-to-win" [9], humans will try to voluntarily influence their autonomous nervous system state, but these situations are rather exceptional in daily life.

Eye gaze, on the other hand, is certainly within deliberate human control, i.e., humans may decide to look into a particular direction or attend to a particular object. However, the activity of the eye and its pupil has been shown to manifest rich information about a person's interpretation of (and attitude to) its environment beyond what is intentionally attended to [10]. A well-known example is visual preference formation when given two visual stimuli [11]. In this setting, humans will

Manuscript received January 7, 2006; revised July 31, 2006. Abstract published on the Internet January 14, 2007. This work was supported in part by a Research Grant (FY1999–FY2003) for the Future Program of the Japan Society for the Promotion of Science (JSPS), in part by a JSPS Encouragement of Young Scientists Grant (FY2005–FY2007), and in part by an NII Joint Research Grant with the University of Tokyo (FY2005).

H. Prendinger is with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: helmut@nii.ac.jp).

M. Ishizuka is with the Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan (e-mail: ishizuka@i.u-tokyo.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2007.891646

eventually direct their focus of attention to the preferred object depending on their attitude and interest.

We want to contrast our approach to understanding humans and their intentions to traditional plan recognition, which refers to the task of inferring the plan or plans of humans from observations of their voluntary, conscious, or purposeful actions [12].

The remainder of this paper is organized as follows. In Section II, we will describe two emotion models and will also address problems in recognizing affective states from physiological information. In Section III, we will briefly discuss the role of eye movements in focus of attention recognition, and then, in Section IV, explain the importance of combining multiple modalities for understanding humans. In Section V, two example applications will be described that illustrate our first steps toward realizing interfaces that are (partly) symmetric with respect to involuntary human expressions. Finally, in Section VI, this paper is rounded off by our conclusions.

II. Recognizing Affective States

Recognizing emotions or affective states from a human's autonomic nervous system (ANS) activity is a hard and challenging problem [13]. Emotions are very short lived (at the scale of some seconds), and ANS activity is always superimposed on humans' ongoing internal nervous system activity. Moreover, humans are typically embedded in complex contexts assuming attention and orientation, or engaged in social interaction.

In the following, we will discuss two emotion models related to ANS activity, *autonomic specificity* of emotions and the *2-D model* of the structure of affect.

A. Autonomic Specificity

A crucial problem for the possibility of emotion recognition from ANS activity relates to the existence of *autonomic specificity* for individual emotions [14], [15]. Research in autonomic specificity investigates whether (some) emotions can be distinguished by their associated pattern of ANS activity, or, in more popular terms, whether (some) emotions have "autonomic signatures."

The seminal early study of Ekman *et al.* tried to relate six emotions to ANS activity [16]. For the investigated emotions (surprise, disgust, sadness, anger, fear, and happiness), four types of physiological measures were taken, namely: 1) heart rate; 2) skin temperature (fingers of left hand and right hand); 3) skin resistance; and 4) muscle tension. Two types of emotion-eliciting conditions have been used, namely: 1) directed facial action and 2) relived emotion. We first report on the two results that were independent of the eliciting conditions.

1) There was a larger increase of heart rate with anger (+8.0 ± 1.8 beats/min) and fear (+8.0 ± 1.6 beats/min), than with happiness (+2.6 ± 1.0 beats/min). The values denote means ± standard errors.
2) The decrease of skin temperature was stronger with anger than with happiness.

A differentiation between emotions based on heart rate change and skin temperature change could be shown for the directed facial action task. The heart rate changes for anger, fear, and sadness were significantly greater than the changes for happiness, surprise, and disgust. Concerning skin temperature, the change (i.e., decrease) related to anger was significantly higher than that of fear, sadness, happiness, surprise, and disgust.

In the relived emotion task, the experiment demonstrated decrease of skin resistance as the discriminating factor. The highest decrease of skin resistance (leading to higher skin conductance) was shown for sadness, whereas the decrease of resistance for fear and anger was small. Another finding of the study was that muscle tension could not be used as a discriminating factor for the emotions under investigation.

While work on the autonomic specificity provides important information regarding the impact of certain emotions on ANS activity, computational models for real-time emotion recognition are often based on simpler theories.

B. Two-Dimensional (2-D) Model of Emotion

The 2-D emotion model advocated in [17] and [18] claims that all emotions can be characterized by two bipolar, but independent, dimensions.

1) *Judged valence*: pleasant or unpleasant (or: positive or negative).
2) *Arousal*: calm or aroused.

Here, named emotions can be conceived as coordinate points in the arousal–valence space. For example, the emotions "sadness," "anger," and "happiness" can be characterized as follows: sadness (low arousal and negative valence), anger (high arousal and negative valence), and happiness (low-medium arousal and positive valence).

Although the exact region of each (named) emotion is hard to define, the relative distance between emotions allows to visualize emotions in a very comprehensible way. Our application described in Section V-A is based on the 2-D model of emotion.

C. Physiological Signals

The relation between the physiological signals and the dimensions of arousal and valence is based work in psychophysiology [19]. By way of example, we will describe the functioning, recognition, and impact of five important signals, namely: 1) galvanic skin response (GSR); 2) electromyography (EMG); 3) blood volume pulse (BVP); 4) pupillary response; and 5) eye blinks (EBs) (see also [13]).

1) GSR: The GSR signal is an indicator of skin conductance. Under certain circumstances, the glands in the skin produce ionic sweat, which changes the electrical resistance. By passing small voltage across two electrodes, the conductance between them can be measured. The electrodes can be attached to two fingers. Skin conductance increases linearly with a person's level of overall arousal. Bechara *et al.* [20] also report on interesting results on the relation between anticipatory skin conductance responses and conscious decision making.

2) EMG: The EMG signal measures muscle activity by detecting surface voltage that occurs when the tiny muscle fibers are contracted by means of electrical impulses (lower arm or masseter muscle). Mean muscle activity has been shown to correlate with negatively valenced emotions.

3) BVP: The BVP signal is processed by a method known as photoplethysmography that shines infrared light onto the skin and measures how much is reflected, which is an indicator of blood flow. Since each heartbeat (or pulse) presses blood through the vessels, BVP can also be used to calculate heart rate and interbeat intervals. Higher heart rate increases with negatively valenced emotions, such as anxiety or fear.

4) Pupillary Response: Naturally, the diameter of the pupil is sensitive to the amount of light falling on the eye. It is also known that the aperture of the pupil constricts and dilates through the control of the ANS activity exerted on the muscle of the iris. There exists extensive literature on how pupil size is affected by human emotional and cognitive processes [21]. Specifically, increase in pupil size is a good indicator of novelty, interest, and positive evaluation, but also cognitive load, whereas decrease in pupil size indicates increased fatigue, and possibly negative stimuli ("perceptual avoidance"). Researchers even observed a continuous decrease in pupil size between the beginning and end of a single experimental session.

While those findings relate pupil dilation to the valence dimension, recent results (for auditory emotional stimulation) demonstrated that pupil size is significantly larger for *both* positive *and* negative stimuli, as opposed to neutral ones [22], which suggests that pupil dilation might also be correlated with the arousal. A recent study also investigated the relation of pupil size to specific affect-related states, such as confusion and surprise [23].

5) EBs: Spontaneous EBs moist the eyeball and, thus, keep the cornea healthy. EBs occur throughout the day, with an average of 15–20 times/min for a relaxed person. From a physiological point of view, only two to four blinks are necessary for an adult. For example, while reading, the blink rate can drop to three blinks per minute.

From a psychological perspective, blink frequency reflects negative affective states, such as nervousness, stress, and fatigue. For example, EB magnitudes were shown to be larger and latencies faster during negative as opposed to positive imagery. Moreover, higher arousal resulted in larger magnitude and shorter latency of EBs.

Other physiological signals, including electrocardiogram, electroencephalogram, as well as event-related brain potentials, are often considered as having too intrusive measurement methods to be of practical use for human–computer interaction. Those and further signals are extensively discussed in [19].

D. Problems in Emotion Recognition From ANS Activity

The psychophysiological literature discusses a number of problems related to the (real-time) assessment of a person's physiological information [14]. Most important among them are the "Baseline Problem," the "Timing of Data Assessment Problem," and the "Intensity of Emotion Problem" that will be discussed in the subsequent sections.

1) Baseline Problem: This problem refers to the difficulty of finding a condition against which physiological change can be compared—commonly referred to as the "baseline." An obvious choice is a "resting" period where the subject can be assumed to experience no particular emotion. However, as Levenson [14, p. 24] notes, emotion "is rarely superimposed upon a prior state of 'rest.' Instead, emotion occurs most typically when the organism is in some prior activation." Consequently, Levenson suggests to adopt a baseline procedure that generates a moderate level of ANS activity. This procedure is a *global baseline* method, i.e., the baseline is taken once and used to normalize biosignal values of whole interaction period. A global baseline guarantees some independence of subjects' individual ANS activity levels as well as independence of situational factors, such as room temperature.

Note, however, that in a more general setting, e.g., when using information about human physiology in a pervasive or ubiquitous computing environment [24], the assumption of a "relaxation" period is impractical as users cannot be expected to provide a baseline measurement before entering the environment. Levenson [14] also pointed out the possibility of methodological problems with global baselines and motivated the recording of *local baseline* as an alternative approach. A local baseline method for the skin conductance signal is described in Healey [25]. She developed an automatic startle detection algorithm that establishes a local baseline at the onset level of the (second) response where the first derivative exceeds a certain threshold (to distinguish high from low arousal), and then finds the local maximum following that point (a peak).

The main (methodological) rationale for a assuming a local baseline is that although biometric signals are "center seeking" (homeostatic), there might be slight shifts in the center point over time (recall also the remark in Section II-C4). In our interactive gaming system study described in Section V-A, results from applying both baseline methods were analyzed.

2) Timing of Data Assessment Problem: This problem refers to the temporal dimension of emotion elicitation, including onset (indicating how fast an emotion is elicited) and duration (apex and offset) of emotions. Levenson [14] suggests 0.5–4 s as an approximation for the duration of emotions, which locates them durationwise between (orienting) reflexes (i.e., an organism's response to novelty) and moods.

The generalization to environments that process an ongoing stream of autonomic activity rather than a specified segment of the interaction may pose additional challenges for determining the occurrence of emotions correctly. As pointed out in [14], when measuring at the wrong time the emotion might be missed or, different emotions might be conflated when too long periods are measured. While the ANS is sometimes considered as a slowly reacting system, latency of onset for autonomic activity related to emotions can be very short, e.g., with surprise. On the other hand, an emotion like "anger" may build up over time and blur the actual "start" of the anger emotion.

3) Intensity of Emotion Problem: This problem concerns the question how the intensity of an emotion is reflected in the physiological data. While at a low level of emotional intensity no informative ANS activity occurs, a very high intensity level may destroy the pattern of ANS activity associated with an emotion [14]. In practice, emotions with little autonomic activity ("relaxed happiness") or moderate intensity levels seem to occur most frequently. It has to be said that to date, issues in emotion intensity remain largely unsolved [14].

III. Recognizing Focus of Attention

Eye movement data have been analyzed for two main purposes [10].

1) *Diagnostic Use*: Eye movement data provide evidence of the user's focus of attention and can be investigated, e.g., to evaluate the usability of interfaces [26].
2) *Interactive Use*: Here, the computer system responds to the observed eye movements and can thus be seen as an input modality, like a pointing device [27].

When eye movements are relatively steady for a short period of time (250–300 ms), they are called *fixations*, whereas rapid shifts from one area to another are called *saccades* [28]. During a saccade, no visual processing takes place. If a cluster of gaze points has less than six entries, it is categorized as part of a saccade [26] (assuming a minimum duration of 100 ms for a fixation at 60 Hz).

Although gaze point and focus of attention are not necessarily always identical, a user's eye movement data provide rich evidence of the user's visual and (overt) attentional processes, since eyes naturally fixate upon visual areas that are surprising, salient, or of interest to a person [29].

Of particular importance to human–computer applications such as educational systems and product presentations is to detect what a person is interested in or what a person's preference is when multiple stimuli are presented. Here, recent work in neuroscience that investigated gaze-based preference decisions shows promising results. In a task involving attractiveness comparisons, the so-called "gaze cascade effect" was observed [11], which refers to a gaze pattern where subjects gradually increase the duration of gaze at one stimulus (the more attractive one), thereby decreasing time to inspect the other (less attractive one).

This finding might also contribute to understanding what a person is (truly) interested in. However, "interest" is a very complex concept that is presumably better handled by combining multiple inputs—the topic of Section IV.

IV. Combining Multiple Modalities

Multiple input modalities are combined for two main purposes, namely: 1) to increase the likelihood of correct classification regarding a single cognitive-motivational state, such as emotion and 2) in order to detect (cognitive-motivational) states that can be considered as emerging from different such states, including interest (confusion and boredom) estimation from emotion recognition and attention detection. We are aware that our distinction is not unequivocal from the viewpoint of measurement. For example, Andreassi [19] discusses both emotion-related ANS activity and eye movements as part of psychophysiology.

A. Emotion Recognition by Multiple Modalities

A natural approach to achieving higher accuracy in recognizing a person's affective state is to combine multiple input modalities. For instance, Huang *et al.* [30] combined speech and facial features to infer a person's emotional state. For the same purpose, Kim *et al.* [2] used speech and biosignals.

While all of the reported combination methods succeed in providing higher recognition accuracy for emotion detection, they are based on offline methods, typically machine-learning techniques. However, in order to implement human–computer interfaces that are sensitive to, e.g., the user's affective state, multiple modalities have to be processed in real time. When interpreted, the system may trigger, e.g., an appropriate empathic response [31]. Obviously, this poses several challenges regarding the synchronization of different input modalities. For one, different signals have different latencies (onset) and durations (apex and offset), as described in Section II-D2, and second, different emotions (let alone moods) span over different periods of time (e.g., surprise versus "growing" frustration).

Once these problems are resolved to a sufficient extent, however, combined modalities may prove useful (at least) along two dimensions.

1) *Increased Accuracy*: The probability of classifying an emotion correctly may be higher if supported by the recognition results of multiple input modalities, rather than by considering the result of each modality in isolation.
2) *Disambiguation*: The consideration of multiple modalities may allow us to handle more complex conversational expressions, e.g., situations where modalities are "conflictive" in that they support incompatible assumptions with respect to a person's affective state (see also [2]).

For an example where disambiguation is required, consider a person that is observed to be highly confused (calculated from ANS activity) while communicating understanding by means of a head-nod upon direct verbal inquiry regarding his or her comprehension by an interlocutor. A disambiguation module might weight each modality and possibly consult other available sources regarding the person's cognitive state.

Besides such cases of "deceptive" behavior, which are conscious, we may also have to consider unconscious forms of conflicts, e.g., when a person *believes* to be in a relaxed state, whereas ANS activity indicates the opposite, such as high arousal or stress.

B. Interest Detection

Our intended concept of *interest* denotes a state of "deep" (and possibly lasting) concern for a visually presented object, by contrast to an accidental glance.

In the setting of a puzzle-solving interaction, Kapoor *et al.* [32] investigated facial expressions and posture data for interest detection. Zeng *et al.* [33] used facial and speech features. Recently, Koshizen *et al.* [34] advocated an approach to estimating user interest (and satisfaction) that is based on both physiological signals and eye movements. In order to achieve high accuracy of estimating interest, gaze duration time is combined with skin potential level-based arousal detection. Here, arousal data allow for more precise segregation into interest and noninterest regions in 2-D space by using a learning scheme called "cross-modal computation."

Although research on interest estimation is still in its infancy, even a straightforward integration of affect and attention can be highly beneficial for improving human–computer interaction.

The recognition of a mental person's state in an *affect–attention space* can reveal to which interface object a user is attending to when experiencing a particular emotion. In this way, a computer system may respond more sensitively to the user's current interaction state.

V. Applications

In this section, we will briefly report on two recent studies that we conducted with the aim of demonstrating the utility and effect of employing multimodal lifelike characters as interface objects. In the *affective gaming* application, a 3-D character capable of facial emotions, "affective noise" (e.g., grumbling), and gestural behavior [35] plays a card game against a user whose emotions are detected through biosignals and taken into account in the agent's response. In the *virtual apartment presentation* application, a 2-D character with deictic face/hand gestures and (synthetic) speech guides the user through an online apartment. Users' focus of attention is recorded in order to estimate the utility of the agent's gestures.

Those applications can be considered as initial steps toward realizing and understanding symmetric multimodality between humans and computers.

A. Interactive Gaming

We measured the user's emotion derived from skin conductance and EMG (based on the model described in Section II-B) in the interactive card game called "Skip-Bo" [36]. As Skip-Bo is a competitive game, the impact of two types of "empathy" on the user was implemented. Empathy can be characterized as an other-oriented emotional response or cognitive act of taking another person's perspective.

- *Negative Empathy*: The agent will display, e.g., gloating joy if the user is recognized to be negatively aroused.
- *Positive Empathy*: The agent displays happiness if the user is detected to be in happy or relaxed affective state.

In both cases, the agent will also display self-centered emotions, such as being happy about its own successful game move. As control conditions, the agent will either display only self-centered emotions or no emotions at all.

In one type of analysis, we focused on game situations (10-s segments) where emotional reactions in the human player were likely to occur, i.e., whenever *either of* the players (user or agent) was able to play at least two pay-off pile cards in a row (which are moves toward winning the game). The results for skin conductance are shown in Fig. 1 (an extensive report on the results is given in [37]).

The results of this study indicated that the absence of negative empathy is conceived as arousing or stressful. For both the "User" and "Agent" winning move situations, we found a significant difference between the negative empathic condition and the positive empathic condition. Further findings included that negative emphatic behavior induces negatively valenced user emotions (derived from EMG), suggesting a certain reciprocity in the user's response [37].

As a complementary analysis, we also investigated the local baseline method discussed in Section II-D1. Fig. 2 depicts a 10-s segment where the agent character expresses "joy" to a subject.

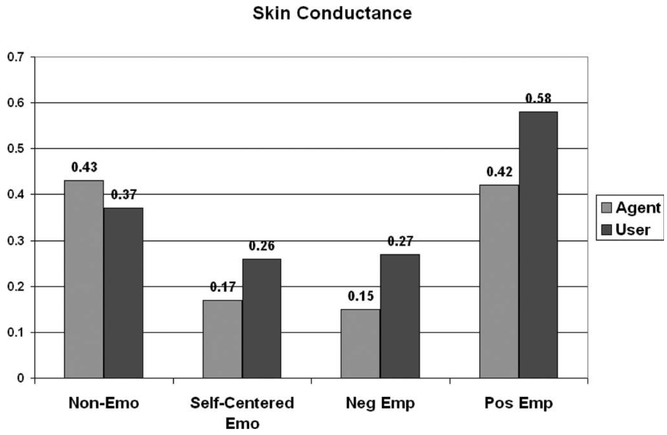

Fig. 1. Average values of normalized skin conductance data within dedicated segments of the interaction in the four conditions, namely: 1) nonemotional; 2) self-centered emotional; 3) negative empathic; and 4) positive empathic. "Agent" refers to situations where the agent performs a winning move. "User" refers to winning move situations of the user.

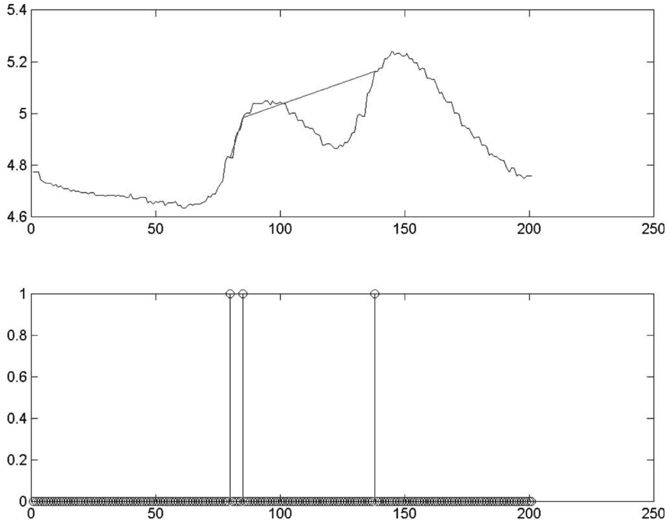

Fig. 2. Agent expresses "joy" to user. (Upper part) Skin conductance signal at 20 samples/s (samples at x-axis); values in microsiemens (y-axis). (Lower part) Starting point and the ending point of crossing startle threshold.

A major finding of the study was that both local and global baselines led to a similar outcome regarding the detection of a user's arousal when the agent displayed some particular (joy, fear, and sadness). This is practically important since, in most real-world applications, a global baseline is hard to obtain.

B. Virtual Apartment Presentation

We tracked users' focus of attention while they watched the presentation of an apartment. Views of each room of the apartment were shown during the presentation, including pictures of some parts of the room and close-up pictures. Besides the 1) *Agent and Speech* version, where a lifelike character presents the apartment (see Fig. 3), we also prepared 2) *Text Box and Speech*, and 3) *Voice (only)* versions in order to compare the effect of the agent to other multimedia presentation methods.

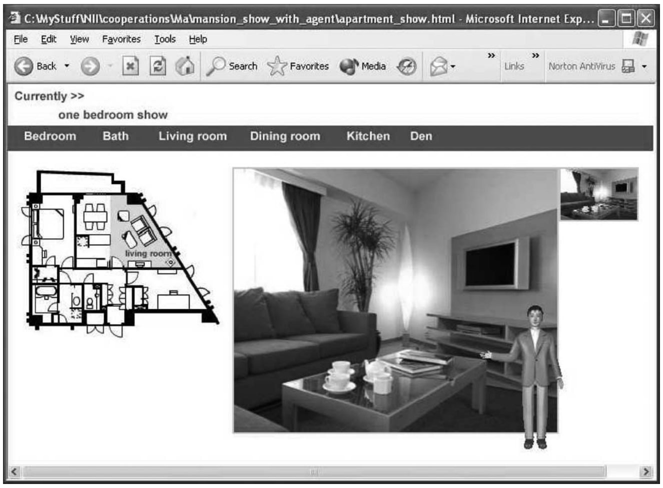
C:\MyStuff\NII\cooperations\Ma\mansion_show_with_agent\apartment_show.html - Microsoft Internet Exp...
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Links
Norton AntiVirus
Currently >>
one bedroom show
Bedroom
Bath
Living room
Dining room
Kitchen
Den
living room

Fig. 3. Lifelike animated agent presents the living room of the apartment by a deictic hand gesture.

Results were distilled from applying both *spatial (cumulative)* and *spatio-temporal* analysis methods [38]. Spatial analysis counts the gaze points that fall within certain screen areas and hypothesizes users' attentional focus. Our findings of the cumulative analysis include, e.g., that users are looking mostly at the character's face, which suggests that users interact socially with agents. While a spatial analysis can indicate where attention is spent, it cannot reveal how users traverse the interface when watching a presentation. In order to address those more complex aspects of character-based interfaces, we also performed a spatio-temporal analysis.

Here, we briefly summarize our experimental findings (see [38] for an extensive discussion).

- The agent's referential (hand or facial) gestures may direct the user's focus of attention to the intended reference object better than a text box or only voice.
- If the uttered sentence contains a trigger word—a word that has a corresponding semantically related visualization—an agent using gestures helps users to locate the (visual) reference object effectively [39]. By contrast, directional support by a text box or voice often shows considerable latency.
- Users often redirect their attention back and forth between the animated agent and the reference object, similar to human–human communication.

The results of this study demonstrate that lifelike characters technology is an effective means of providing navigational aid to a user. A natural extension of this work is to let the interface agents recognize the user's gaze behavior through eye-tracking technology. In this way, the agent may "perceive" the user's interest state and respond to the user appropriately, e.g., by redirecting the user's focus of attention.

VI. Conclusion

This paper has described our approach to Humatronics, a research field that is concerned with finding new ways to achieve a symmetric relationship between humans and computers, with the overall goal of improving the interaction with and accessibility of computational devices. We started out with "revisiting" the symmetric multimodality idea outlined in [1], where all input modes (speech, gesture, and facial expression) are also available for output, as exemplified in the SmartKom system. While input modes in this system are mostly meant to purposefully direct and maintain a mixed-initiative dialogue with an animated agent, we put the emphasis on processing input signals that are (largely) involuntary and unconscious. In this way, we hope to gain better insight into a user's cognitive-motivational state, and hence, be able to better *understand* the user, complementary to what is voluntarily expressed.

Physiological signals and eye movements have been shown to be valuable indexes for a person's cognitive-motivational state related to affect, attention, and interest. An important question is whether those signals also allow us to detect a person's *intention*. The "standard" approach to intention recognition is to infer a person's plan or goal from purposeful actions, e.g., moving objects or pressing buttons [12]. By contrast, a "physiological computing" approach would infer a person's intention from involuntary bodily activity related to emotion and interest. A person's current affective state such as frustration or anger certainly has an impact on what the person is likely to do next. Likewise, a person's way of (eye) scanning a working area (e.g., a computer screen) might be a good indicator of subsequent actions. The development of principled approaches to physiology-based intention recognition seems to be a promising future line of research.

The applications presented in this paper can be considered as first steps toward symmetric multimodal systems. While symmetry on the level of human features (rather than methods) is partly realized for emotions, i.e., recognizing human emotions [13] and expressing agent emotions [3], [6], the area of expressing attention by an agent is largely unexplored. Highly promising work can be found in [40], which proposes a gaze model for an animated agent in order to provide feedback, conversational control, and subtle signaling of interest and engagement. We are currently in the process of developing solutions to the issues sketched in this concluding section.

References

[1] W. Wahlster, "Towards symmetric multimodality: Fusion and fission of speech, gesture and facial expression," in *Proc. 26th German Conf. Artif. Intell.*, 2003, pp. 1–18.

[2] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner, "Integrating information from speech and physiological signals to achieve emotional sensitivity," in *Proc. 9th Eur. Conf. Speech Commun. and Technol.*, 2005, pp. 809–812.

[3] H. Prendinger and M. Ishizuka, Eds., *Life-Like Characters. Tools, Affective Functions, and Applications*, ser. Cognitive Technologies, Berlin, Germany: Springer-Verlag, 2004.

[4] C. Pelachaud and M. Bilvi, "Computational model of believable conversational agents," in *Communication in Multiagent Systems: Background, Current Trends, and Future*, vol. LNCS 2650. New York: Springer-Verlag, 2003, pp. 300–317.

[5] H. Prendinger, S. Descamps, and M. Ishizuka, "MPML: A markup language for controlling the behavior of life-like characters," *J. Vis. Lang. Comput.*, vol. 15, no. 2, pp. 183–203, 2004.

[6] M. Ishizuka and H. Prendinger, "Describing and generating multimodal contents featuring affective lifelike agents with MPML," *New Gener. Comput.*, vol. 24, no. 2, pp. 97–128, Jan. 2006.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer

interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[8] J. Allanson, "Electrophysiologically interactive computer systems," *Computer*, vol. 35, no. 3, pp. 60–65, Mar. 2002.

[9] D. Bersak, G. McDarby, N. Augenblick, P. McDarby, D. McDonnell, B. McDonald, and R. Karkun, "Intelligent biofeedback using an immersive competitive environment," in *Proc. Online, Ubicomp Workshop Designing Ubiquitous Comput. Games*, 2001.

[10] A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*. London, U.K.: Springer-Verlag, 2003.

[11] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier, "Gaze bias both reflects and influences preference," *Nat. Neurosci.*, vol. 6, no. 12, pp. 1317–1322, Dec. 2003.

[12] H. Kautz, "A formal theory of plan recognition," Ph.D. dissertation, Univ. Rochester, Rochester, NY, 1987.

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

[14] R. W. Levenson, "Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity," in *Social Psychophysiology and Emotion: Theory and Clinical Applications*, H. L. Wagner, Ed. Hoboken, NJ: Wiley, 1988, pp. 17–42.

[15] ——, "Autonomic specificity and emotion," in *Handbook of Affective Sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds. Oxford, U.K.: Oxford Univ. Press, 2003, pp. 212–224.

[16] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, Sep. 1983.

[17] P. J. Lang, "The emotion probe: Studies of motivation and attention," *Amer. Psychol.*, vol. 50, no. 5, pp. 372–385, May 1995.

[18] L. Feldman-Barrett and J. A. Russell, "The structure of current affect: Controversies and emerging consensus," *Curr. Dir. Psychol. Sci.*, vol. 8, no. 1, pp. 10–14, 1999.

[19] J. L. Andreassi, *Psychophysiology. Human Behavior & Physiological Response*, 4th ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[20] A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio, "Deciding advantageously before knowing the advantageous strategy," *Science*, vol. 275, no. 5304, pp. 1293–1295, Feb. 1997.

[21] E. H. Hess, "Pupillometrics: A method of studying mental, emotional and sensory processes," in *Handbook of Psychophysiology*, N. Greenfield and R. Sternbach, Eds. New York: Holt, Rinehart and Winston, 1972, pp. 491–531.

[22] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *Int. J. Human-Comput. Stud.*, vol. 59, no. 1/2, pp. 185–198, Jul. 2003.

[23] H. Umemuro and J. Yamashita, "Detection of user's confusion and surprise based on pupil dilation," *Jpn. J. Ergonomics*, vol. 39, no. 4, pp. 153–161, 2003.

[24] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Pers. Commun.*, vol. 8, no. 4, pp. 10–17, Aug. 2001.

[25] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, MIT, Cambridge, MA, 2000.

[26] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *Int. J. Ind. Ergon.*, vol. 24, no. 6, pp. 631–645, Oct. 1999.

[27] R. J. K. Jacob, "The use of eye movements in human-computer interaction techniques: What you look at is what you get," *ACM Trans. Inf. Syst.*, vol. 9, no. 3, pp. 152–169, 1991.

[28] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Eye Tracking Res. and Appl. Symp.*, 2000, pp. 71–78.

[29] G. Loftus and N. Mackworth, "Cognitive determinants of fixation location during picture viewing," *J. Exp. Psychol., Hum. Percept. Perform.*, vol. 4, no. 4, pp. 565–572, Nov. 1978.

[30] T. S. Huang, L. S. Chen, H. Tao, T. Miyasato, and R. Nakatsu, "Bimodal emotion recognition by man and machine," in *Proc. ATR Workshop Virtual Commun. Environments*, 1998.

[31] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users' affective states," *Int. J. Appl. Artif. Intell.*, vol. 19, no. 3, pp. 267–285, Mar./Apr. 2005.

[32] A. Kapoor, R. W. Picard, and Y. Ivanov, "Probabilistic combination of multiple modalities to detect interest," in *Proc. Int. Conf. Pattern Recog.*, 2005, pp. 969–972.

[33] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. S. Huang, D. Roth, and S. Levinson, "Bimodal HCI-related affect recognition," in *Proc. 6th ICMI*, 2004, pp. 137–143.

[34] T. Koshizen, Y. Hasegawa, H. Tusjino, M. Kon, K. Aihara, and H. Prendinger, "A learning system for user modeling by combined cognitive and affective modeling for user interest estimation," in *Proc. 7th ICCM*, 2006, pp. 184–189.

[35] S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth, "Max—A multimodal assistant in virtual reality construction," *KI Zeitschift (German Magazine of Artificial Intelligence)—Special Issue on Embodied Conversational Agents*, vol. 4, no. 3, pp. 11–17, 2003.

[36] C. Becker, H. Prendinger, M. Ishizuka, and I. Wachsmuth, "Evaluating affective feedback of the 3D agent Max in a competitive cards game," in *Proc. 1st Int. Conf. ACII*. Berlin, Germany: Springer-Verlag, 2005, vol. LNCS 3784, pp. 466–473.

[37] H. Prendinger, C. Becker, and M. Ishizuka, "A study in users' physiological response to an empathic interface agent," *Int. J. Humanoid Robotics*, vol. 3, no. 3, pp. 371–391, Sep. 2006.

[38] H. Prendinger, C. Ma, J. Yingzi, A. Nakasone, and M. Ishizuka, "Understanding the effect of life-like interface agents through eye users' eye movements," in *Proc. 7th ICMI*, 2005, pp. 108–115.

[39] R. M. Cooper, "The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing," *Cogn. Psychol.*, vol. 6, no. 1, pp. 84–107, Jan. 1974.

[40] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *Proc. 5th Int. Working Conf. IVA*. New York: Springer-Verlag, 2005, vol. LNAI 3661, pp. 229–240.

Helmut Prendinger received the M.A. and Ph.D. degrees from the University of Salzburg, Salzburg, Austria.

He is an Associate Professor at the National Institute of Informatics, Tokyo, Japan. Previously, he was a Research Associate and a JSPS Postdoctoral Fellow at the University of Tokyo. Earlier, he was a Junior Specialist at the University of California, Irvine. He is a coeditor (with Mitsuru Ishizuka) of a book on lifelike characters that appeared in the Cognitive Technologies series of Springer. His research interests include artificial intelligence, affective computing, and human–computer interaction, in which areas he has published more than 75 papers in international journals and conference proceedings.

Mitsuru Ishizuka (M'78) received the B.S., M.S., and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan.

He is a Professor at the Graduate School of Information Science and Technology, University of Tokyo. Previously, he was with the NTT Yokosuka Laboratory and the Institute of Industrial Science, University of Tokyo. During 1980–1981, he was a Visiting Associate Professor at Purdue University. His research interests include artificial intelligence, multimodal media with lifelike agents, and intelligent WWW information space.

Prof. Ishizuka is a member of the American Association for Artificial Intelligence and the Inter Press Service Japan, and the President of the Japanese Society for Artificial Intelligence.