Affective Multimodal Human-Computer Interaction

Maja Pantic[1], Nicu Sebe[2], Jeffrey F. Cohn[3] and Thomas Huang[4]
[1] Faculty of EEMCS, Delft University of Technology, The Netherlands
[2] Faculty of Science, University of Amsterdam, The Netherlands
[3] Psychology and Psychiatry, University of Pittsburgh, USA
[4] Beckman Institute, University of Illinois at Urbana-Champaign, USA
mpantic@ieee.org, nicu@science.uva.nl, jeffcohn@pitt.edu, huang@ifp.uiuc.edu

ABSTRACT
Social and emotional intelligence are aspects of human intelligence that have been argued to be better predictors than IQ for measuring aspects of success in life, especially in social interactions, learning, and adapting to what is important. When it comes to machines, not all of them will need such skills. Yet to have machines like computers, broadcast systems, and cars, capable of adapting to their users and of anticipating their wishes, endowing them with the ability to recognize user's affective states is necessary. This article discusses the components of human affect, how they might be integrated into computers, and how far are we from realizing affective multimodal human-computer interaction.

Categories and Subject Descriptors
A.1 [**Introductory and Survey**]
H1.2 [**User/Machine Systems**]: Human information processing
H.5.1 [**Multimedia Information Systems**]: Audiovisual input
I.5.4 [**Pattern Recognition Applications**]: Models, Learning

General Terms
Algorithms, Theory, Performance

Keywords
Affective computing, Multimodal Human-Computer Interaction

1. INTRODUCTION
We have entered an era of enhanced digital connectivity. Computers and the Internet have become so embedded in the daily fabric of people's lives that we can no longer live without them [20]. We use this technology to work, communicate, shop, seek out new information, and entertain ourselves. With the ever-increasing diffusion of computers into society, human-computer interaction (HCI) is becoming increasingly essential to our daily lives.

HCI design was first dominated by direct manipulation and then delegation. The tacit assumption of both approaches to interaction has been that the human will be explicit, unambiguous and fully attentive while controlling information and command flow. Boredom, preoccupation, and stress are unthinkable even though they are "very human" behaviors. The insensitivity of current HCI designs is acceptable 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' level of attention, preferences, moods, and intentions.

The ability to detect and understand affective states and other social signals of someone with whom we are communicating is the core of social and emotional intelligence. This kind of intelligence is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life [18]. When it comes to computers, however, they are socially ignorant [35]. Current HCI technology does not account for the fact that human-human communication is always socially situated and that discussions are not just facts but part of a larger social interplay. Not all computers will need social and emotional intelligence and none will need all of the related skills humans have. Yet, human-machine interactive systems capable of sensing stress, inattention, and heedfulness, and capable of adapting and responding to these affective states of users are likely to be perceived as more natural, efficacious, and trustworthy. For example, in education, pupils' affective signals inform the teacher of the need to adjust the instructional message. Successful human teachers acknowledge this and work with it; digital conversational embodied agents must begin to do the same by employing tools that can accurately sense and interpret affective signals and social context of the pupil, learn successful context-dependent social behavior, and use a proper affective presentation language (e.g. [33]) to drive the animation of the agent. The research area of machine analysis and employment of human affective states to build more natural, flexible HCI goes by the general name of affective computing as introduced by Picard [36].

2. THE APPLICATION DOMAIN
In addition to 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 [19]. Automatic assessment of boredom, inattention, and stress would be highly valuable in situations in which firm

MM'05, November 6–11, 2005, Singapore.

attention to a crucial but perhaps tedious task is essential. Examples include air traffic control, nuclear power plant surveillance, and operating a motor vehicle. An automated tool could provide prompts for better performance informed by assessment of the user's affective state.

Other domain areas in which machine tools for analysis of human affective feedback could expand and enhance scientific understanding and practical applications include specialized areas in professional and scientific sectors. In the security sector, affective behavioral cues play a crucial role in establishing or detracting from credibility. In the medical sector, affective behavioral cues are a direct means to identify when specific mental processes are occurring. Machine analysis of human affective states could be of considerable value in these situations in which only informal, subjective 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 emotion dysfunction or impairment and brain lesions) and psychiatry (in studies on schizophrenia and mood disorders) in which reliability, sensitivity, and precision of measurement of affective behavior are persisting problems [30].

3. THE PROBLEM DOMAIN

While all agree that machine sensing and interpretation of human affective information would be widely beneficial, addressing these problems is not an easy task. The main problem areas can be defined as follows.

What is an affective state? This question is related to psychological issues pertaining to the nature of affective states and the best way to represent them.

Which human communicative signals convey information about affective state? This issue shapes the choice of different modalities to be integrated into an automatic analyzer of human affective feedback.

How are various kinds of evidence to be combined to optimize inferences about affective states? This question is related to how best to integrate information across modalities for emotion recognition.

What is an affective state? Traditionally, the terms "affect" and "emotion" have been used synonymously. Following Darwin, discrete emotion theorists propose the existence of six or more basic emotions that are universally displayed and recognized [11], [23]. These include happiness, anger, sadness, surprise, disgust, and fear. Data from both Western and traditional societies suggests that non-verbal communicative signals (especially facial and vocal expression) involved in these basic emotions are displayed and recognized cross-culturally. In opposition to this view, Russell [38] among others argues that emotion is best characterized in terms of a small number of latent dimensions, rather than in terms of a small number of discrete emotion categories. Russell proposes bipolar dimensions of arousal and valence (pleasant versus unpleasant). Watson and Tellegen propose unipolar dimensions of positive and negative affect while Watson and Clark proposed a hierarchical model that integrates discrete emotions and dimensional views [24], [43], [44]. Social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations. They argue further that emotion is culturally constructed and no universals exist. From their perspective, subjective experience and whether or not emotion is better conceptualized categorically or dimensionally is culture specific. Then there is lack of consensus on how affective displays should be labeled. For example, Fridlund argues that human facial expressions should not be labeled in terms of emotions but in terms of Behavioral Ecology interpretations, which explain the influence a certain expression has in a particular context. Thus, an "angry" face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. Yet, people still tend to use *anger* as the interpretation rather than *readiness-to-attack* interpretation. Another issue is that of culture dependency: the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent [25], [45]. In summary, previous research literature pertaining to the nature and suitable representation of affective states provides no firm conclusions that could be safely presumed and adopted in studies on machine analysis of human affective states and affective computing. Also, it is not only discrete emotional states like surprise or anger that are of importance for the realization of proactive human-machine interactive systems. Sensing and responding to behavioral cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are essential. Hence, in contrast to traditional approach, we treat affective states as being correlated not only to discrete emotions but to other, aforementioned social signals as well. Furthermore, since 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 independently of the situation and the observer, we advocate 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.

Which human communicative signals convey information about affective state? Affective arousal modulates all human communicative signals [11]. However, the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioral cues [1]. Human judges seem to be most accurate in their judgment when they are able to observe the face and the body. Ratings that were based on the face and the body were 35% more accurate than the ratings that were based on the face alone. Yet, ratings that were based on the face alone were 30% more accurate than ratings that were based on the body alone and 35% more accurate than ratings that were based on the tone of voice alone [1]. These findings indicate that to interpret someone's behavioral cues, people rely on shown facial expressions and to a lesser degree on shown body gestures and vocal expressions. However, although basic researchers have been unable to identify a set of voice cues that reliably discriminate among emotions, listeners seem to be accurate in decoding emotions from voice cues [21]. Thus, automated human affect analyzers should at least include facial expression modality and preferably they should also include (one or both) modalities for perceiving body gestures and tone of the voice. Finally, while too much information from different channels seem to be confusing to human judges, resulting in less accurate judgments of shown behavior when 3 or more observation channels are available (e.g. face, body, and speech) [1], combining those multiple modalities

(including speech and physiology) may prove appropriate for realization of automatic human affect analysis.

How are various kinds of evidence to be combined to optimize inferences about 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. Thus, in order to accomplish a multimodal analysis of human interactive signals acquired by multiple sensors, which resembles human processing of such information, input signals should not be considered mutually independent and should not 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 [30]. The latter refers to the fact that one must know the context in which the observed interactive signals have been displayed (who the expresser is and what his current environment and task are) in order to interpret the perceived multi-sensory information correctly.

Hence, an "ideal" automatic analyzer of human affective information should be able to emulate at least some of the capabilities of the human sensory system as summarized below.

An "ideal" automatic human-affect analyzer would be:
- multimodal (handling the face, body, and tone of voice)
- robust and accurate (despite occlusions, changes in viewing and lighting conditions, and ambient noise)
- generic (independent of physiognomy, sex, age, and ethnicity of the subject)
- sensitive to the dynamics of displayed affective expressions (performing temporal analysis of the sensed data, previously processed in a joint feature space)
- context-sensitive (realizing environment- and task-dependent data interpretation in terms of user-profiled affect-descriptive labels)

4. THE STATE OF THE ART

To interpret someone's behavioral cues, including emotional states, people rely mainly on shown facial expressions [1], [23], 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, readers are referred to [30]. The survey indicates that the capabilities of currently existing facial affect analyzers are rather limited. Nevertheless, the automated systems achieve an accuracy of 64% to 98% when detecting 3-7 emotions deliberately displayed by 5-40 subjects.

Limitations of existing facial-affect analyzers are [30]:
- handle only 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 constant illumination
- do not perform a context-dependent interpretation of shown facial behavior
- do not analyze extracted facial information on different time scales (short videos are handled only); consequently, inferences about the expressed mood and attitude (larger time scales) cannot be made by current facial affect analyzers

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 unless both systems are tested on the same dataset. The main problem is that no database of images exists that is shared by all diverse facial-expression-research communities. In general, small databases of facial-expression images are made and exploited by each research community. The databases shared currently by several research communities are the Cohn-Kanade Facial Expression Database [22], the JAFFE Database [47] and the MMI Facial Expression Database [32].

Although scientists conducted a large number of studies of vocal expression in an attempt to specify what aspects of the voice are predictive of expressed or portrayed emotion, they have been unable to identify a set of voice cues that reliably discriminate among emotions [21]. In the light of this problem it is not surprising that computer science and related fields produced rather disappointing results on automated vocal affect expression analysis. For a survey of studies on automatic analysis of vocal affect, the readers are referred to [30]. The survey indicates that the existing automated systems for auditory analysis of human affect are quite limited. Similarly to the case of automatic facial affect analysis, it is still not possible to compare different vocal affect analyzers since isolated, small databases of speech material are made and exploited by each research community.

Limitations of existing vocal-affect analyzers are [30]:
- 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 (environment-, 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 or word, or for the detection of phonetic features) – inferences about moods and attitudes (longer time scales) cannot be made by current vocal-affect analyzers
- adopt strong assumptions (e.g., the recordings are noise free, the recorded sentences are short, delimited by pauses, carefully pronounced by non-smoking actors) 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 expressions of affective states

Relatively few of the existing works combine different modalities into a single system for human affective state analysis. Although the studies in psychology on the accuracy of predictions from observations of expressive behavior suggest that the combined face and body are the most informative [1], except of a tentative attempt of Balomenos et al. [2], there is virtually no other effort reported on automatic human affect analysis from combined face and body gestures. Examples of existing works combining different modalities into a single system for human affective state analysis are those of Chen & Huang [2], Yoshitomi et al. [46], De Silva & Ng [6], Go et al. [17], and Song et al. [42], who investigated the effects of a combined detection of facial and vocal expressions of affective states. In brief, these works achieve an accuracy of 72% to 85% when detecting one or more basic

emotions from clean audiovisual input (e.g., noise-free recordings, closely-placed microphone, non-occluded portraits) from an actor speaking a single word and showing exaggerated facial displays of a basic emotion. Although 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 (and some additional) drawbacks of single-modal affect analyzers. In turn, many improvements are needed if those systems 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 separate processing and interpretation of audio and visual data do not suffice.

5. CHALLENGES

Probably the most remarkable issue about the state of the art in the research on affective multimodal HCI is that only a few efforts toward the implementation of audiovisual human-affect analyzer (combining the facial and the vocal affect analysis) have been reported so far. Although the studies in psychology suggest that the combined face and body are very informative when analyzing human expressive behavior, a single effort toward the realization of such a bi-modal affect analyzer has been reported up to date. No effort toward the integration of more than two modalities into an automated human-affect analyzer has been reported so far.

Another issue concerns the interpretation of behavioral cues in terms of affective states. The existing work usually employs singular classification of input data into one of the "basic" emotion categories. Yet pure expressions of "basic" emotions are less frequently elicited; much of the time people show blends of emotional displays. Hence, the classification of human non-verbal affective feedback into a single "basic"-emotion category may not be 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.

Virtually all the existing human-affect analyzers assume that the input data are isolated or pre-segmented expressions showing a single temporal pattern (onset → apex → offset) of an affective state that begins and ends with a neutral state. In reality, such segmentation is an exception. Human expressive behavior is more complex. Transitions from one affective state to another may include multiple apexes and may be direct, without an intermediate neutral state. For this reason, existing human-affect analyzers have difficulties with handling spontaneously occurring expressions of emotion. In addition, eliciting spontaneous affective behavior, which could be used to train human-affect analyzers, represents a research challenge on its own right. Hence, while answering the question of how to parse the stream of spontaneous affective behavior is essential for the realization of affective multimodal HCI, we also recognize the likelihood that such a goal is still in the relatively distant future.

Realization of a human-like interpretation of sensed affective behavior requires context-dependent choices (i.e., environment-, user- and task-profiled choices). Nonetheless, currently existing methods aimed at the automation of human-affect analysis are context insensitive. Although machine-context sensing, that is, answering questions like who is the user, where is (s)he, and what is (s)he doing, has witnessed recently a number of significant advances [34], 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 multimodal human-affect analysis has been established yet. This lack of common testing resource forms a major impediment to comparing, resolving and extending the issues concerned with automatic, multimodal human affect analysis and understanding. It is one of the most critical issues confronting affective multimodal HCI.

6. RECOMMENDATIONS

The remarkable aspect of human expressive behavior is its communicative power: even fleeting glimpses ("thin slices") of expressive behavior communicate a great deal of information. This suggestion is confirmed by findings indicating that judgments about the meaning of expressive behavior are quite accurate even when they are based on brief observations [1]. This is especially true for emotions; judgments about emotions are fairly accurate even from exposures to nonverbal behavior lasting only 375 ms [1]. Much of this expressive behavior is unintended and unconscious (and yet extremely effective). In fact, these expressive nonverbal cues are so subtle that they are neither encoded nor decoded at an intentional, conscious level of awareness [1]. This suggests the following:

- While continuous analysis of human expressive behavior would be ideal, automated human-affect analyzers can be useful even if they are able to analyze only short observations of expressive behavior.
- Biologically inspired classification techniques, like Artificial Neural Networks, may prove more suitable for tackling the problem of human affect recognition than methods like Expert Systems, which consider classification problems from a logical rather than biological perspective. The former are motivated by human unconscious problem solving processes while the latter are inspired by human conscious problem solving processes.

As noted above and remarked already by Pantic and Rothkrantz [30], a typical issue of multimodal data processing proposed so far is that multisensory data are processed separately and only combined at the end. This practice may follow from experimental studies that have shown that a late integration (decision-level data fusion) provides higher recognition scores than an early integration approach [39]. The differences in time scale of the features from different channels and the lack of a common metric across the modalities add and abet the underlying inference that the features from different channels are not sufficiently correlated to be fused at the feature level. Yet, people display audiovisual expressive cues in a complementary and redundant manner. In order to accomplish a human-like multimodal analysis of multiple input signals acquired by different sensors, the signals cannot be considered mutually independent and cannot be combined in a context-free manner at the end of the intended analysis. The input data should be processed in a joint feature space and according to a context-dependent model. In practice, however, besides the

problems of context sensing and developing context-dependent models for combining multisensory data, one should cope with the size of the required joint feature space, which can suffer from large dimensionality, different feature formats, and timing [30]. A potential way to achieve the target temporal fusion of multisensory data and context is to use learned probabilistic models like Dynamic Bayesian Networks (DBN) [29], [15]. Note, however, that classical DBN learning methods can fail when the data exhibits complex behavior, as is the case with spontaneously occurring expressive behaviors. Iteratively learning sets of DBN models in a supervised manner, in which learning is optimized for classification performance [16], may prove successful in that case.

If we consider the state of the art in audio and visual signal processing, noisy and partial input data should also be expected. A multimodal system should be able to deal with imperfect data and generate its conclusion so that the certainty associated with it varies in accordance to the input data. A way of achieving this is to consider the time-instance versus time-scale dimension of human nonverbal communicative signals as suggested by Pantic and Rothkrantz [30]. By considering previously observed data (time scale) with respect to the current data carried by functioning observation channels (time instance), a statistical prediction and its probability might be derived about both the information that have been lost due to malfunctioning/inaccuracy of a particular sensor and the currently displayed action/reaction. Probabilistic graphical models, such as Hidden Markov Models (HMM) and DBN are well suited for accomplishing this. These models can handle noisy features, temporal information, and partial data all by probabilistic inference. Hierarchical HMM-based systems [6] have proven successful for facial expression recognition. DBN and HMM variants [15] seem to perform well for user intent recognition, office activity recognition, and event detection from realistic audiovisual stream [14]. This suggests that probabilistic graphical models are a promising approach to fusing realistic (noisy) audio and video for context-dependent detection of behavioral events such as affective states.

An issue that makes the problem of human affect recognition even more difficult to solve in a general case is the dependency of a person's behavior on his/her personality, cultural, social network, and the context in which the observed behavioral cues are encountered. One source of help for these problems is machine learning: rather than using a priori rules to interpret human behavior, we can potentially learn context-dependent rules by watching the user's behavior in the sensed context [34], [31]. Probabilistic graphical models may be a promising approach. Sets of such models can be learned in an iterative manner by building on previously acquired knowledge when learning new models. For instance, a probabilistic graphical model for affect recognition learned from data about a certain user can be used as a starting point for learning such a model for another user or for learning a new model for the same user and a different context. Though context sensing and the time needed to learn appropriate probabilistic graphical models are significant problems in their own right, many benefits could come from adaptive, contextual, multimodal, affect analyzers.

To develop and evaluate the envisioned contextual multimodal human-affect analyzers, large collections of training and test data are needed. Nonetheless, there is no comprehensive, readily accessible reference set of audiovisual data that could be used as a basis for benchmarks for efforts in the field. Benchmark databases should contain still and motion images of faces and upper bodies (to facilitate the research on human affect analysis from the face and the body), vocalizations and speech (to facilitate the research on vocal affect analysis), and metadata concerning both the context in which the recorded affective expressions were displayed and interpretations of these in terms of shown face and body actions, emotional and attitudinal states. Also, databases should contain existing research findings in order to facilitate the integration of efforts of researchers, highlighting contradictions and consistencies, and suggesting fruitful paths for new research. The lack of such easily accessible, suitable, common testing resources forms a major impediment to comparing and extending the issues concerned with automatic human affect analysis.

Two main issues that make this problem difficult to tackle are those of obtaining the ground truth for the observation data and getting data that genuinely correspond to a particular affective state. Even though there are cases when the data can be easily labeled (e.g., a singular strong emotion is captured, such as an episode of rage), in most cases the ground truth (which affective state was present) is difficult to establish. Furthermore, as any photographer can attest, getting a real smile can be challenging. Asking someone to smile often does not create the same picture as an authentic smile. The fundamental reason of course is that the subject often does not feel happy so his/her smile is artificial and in many subtle ways quite different than a genuine smile [12], [9].

Picard et al. [37] outlined five factors that influence the affective data collection:

- *Spontaneous* versus *posed*: Is the emotion elicited by a situation or stimulus that is outside the subject's control or the subject is asked to elicit the emotion?
- *Lab setting* versus *real-world*: Is the data recording taking place in a lab or in the usual environment of the subject?
- *Expression* versus *feeling*: Is the emphasis on external expression or on internal feeling?
- *Open recording* versus *hidden recording*: Is the subject aware that (s)he is being recorded?
- *Emotion-purpose* versus *other-purpose*: Does the subject know that (s)he is a part of an experiment and the experiment is about emotion?

Note that these factors are not necessarily independent. The most natural setup would imply that the subject feels the emotion internally (*feeling*), the emotion occurs *spontaneously*, while the subject is in his usual environment (*real-world*). Also, the subject should not know that (s)he is being recorded (*hidden recording*) and that (s)he is a part of an experiment (*other-purpose*). Such data are usual impossible to obtain because of privacy and ethics concerns. As a consequence, most researchers who tackled the problem of establishing a comprehensive human-affect expression database used a setup that is rather far from the natural setup. Cohn and Kanade [22], Pantic and Valstar [32], and Lyons [47] collected facial affect data, Banse and Scherer [3] and Nwe et al. [26] collected vocal affect data, while De Silva and Ng [8], and Chen [5] collected audiovisual human affect data using a *posed*, *lab-based*, *expression-oriented*, *open-recording*, and *emotion-purpose* methodology. So far only Sebe et al. [40] reported on efforts in collecting *spontaneous* audiovisual human affect data. They created a video kiosk (*lab setting*) with a hidden camera (*hidden-recording*) which displayed segments from recent movie

trailers. This setup had the main advantage that it naturally attracted people to watch and could potentially elicit emotions through different genres of video footage - i.e. horror films for shock, comedy for joy, etc.

Except of these problems concerned with acquiring valuable data and the related ground truth, another important issue is how does one construct and administer such a large audiovisual benchmark database. The related questions are the following. How does one facilitate efficient, fast, and secure retrieval and inclusion of objects constituting this database? How could fast and reliable object distribution over networks be achieved? How could the performance of a tested automated system be included in the database? How should the relationship between the performance and the database objects used in the evaluation be defined? Pantic et al. [30], [32] emphasized a number of specific research and development efforts needed to address the aforementioned problems. Nonetheless, note that their list of suggestions and recommendations is not exhaustive of worthwhile contributions.

7. CONCLUSSIONS

As remarked by Pentland [34] and Oviatt [26], multimodal context-sensitive (user-, task-, and application-profiled and affect-sensitive) HCI is likely to become the single most widespread research topic of 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 yet mature and need many improvements.

Main challenges in the field of affective multimodal HCI:

- How many and which behavioral channels like the face, the body, and the tone of the voice, should be combined for realization of robust and accurate human affect analysis? Too much information from different channels seems to be confusing for human judges. Does this pertain in HCI?
- At which abstraction level are these modalities to be fused? Humans simultaneously employ tightly coupled modalities of sight and sound. Does this tight coupling persists when the modalities are used for affective multimodal HCI, as suggested by, e.g., Chen and Rao [6], or not, as suggested by Cohn and Katz [8] and Scanlon and Reilly [39]?
- How can the grammar of human expressive behavior be learned? Should this be done in a human-centered manner or in an activity-centered manner as suggested by Norman [26]? How can this information be properly represented and then used to handle malfunctioning / inaccuracy of a particular observation channel and the resulting ambiguities in the observation data?
- How can the interpretation of the observed expressive behavior in terms of any emotion- / attitude- / mood be achieved? How can the system learn to distinguish between these interpretation classes when for the face channel only there are more than 7000 different facial expressions that humans can display [40]? We believe that researchers in the field should not focus on solving challenges in psychology of emotion and should not adhere to only one of emotion theories (e.g. discrete vs. dimensional emotion theory, emotion vs. behavioral ecology theory, etc.). Rather, they should focus on finding pragmatic context-sensitive solutions by learning appropriate models of expressive behavior and the related interpretations from available data and intended users.
- How to include information about the context (environment, user, user's task) in which the observed expressive behavior has been displayed so that a context-sensitive analysis of human behavior can be achieved?
- What properties should automated analyzers of human expressive behavior have in order to be able to analyze human spontaneous behavior? How can such analyzers be realized? How should one elicit spontaneous human expressive behavior including genuine emotional responses?
- United efforts of different research groups working in the field should be made to develop a comprehensive, readily accessible database of annotated, multi-sensory observations of human expressive behavior that could be used as a basis for benchmarks for efforts in the field. The related research questions include the following. How does one facilitate efficient, fast, and secure retrieval and inclusion of objects constituting this database? How could the performance of a tested automated system be included into the database? How should the relationship between the performance and the database objects used in the evaluation be defined?

8. ACKNOWLEDGMENTS

The work of Maja Pantic is supported by the Netherlands Organization for Scientific Research Grant EW-639.021.202. Jeffrey F Cohn is supported in part by NIH grant MHR01 MH051435 and Naval Research Laboratory grant N000140010915.

9. REFERENCES

[1] Ambady, N. and Rosenthal, R. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*, 2 (Feb. 1992), 256-274.

[2] Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K. and Kollias, S. Emotion Analysis in Man-Machine Interaction Systems. *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, vol. 3361, Bengio, S. and Bourlard, H., Eds. Springer-Verlag, Berlin, D, 2005, 318-328.

[3] Banse, R., & Scherer, K. R. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 1996, 614-636.

[4] Chen, L.S. and Huang, T.S. Emotional expressions in audiovisual human computer interaction. In *Proc. Int'l Conf. Multimedia and Expo*, 2000, 423-426.

[5] Chen, L.S., *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction.* PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[6] Chen, T. and Rao, R.R., Audio-visual integration in multimodal communication, *Proceedings of the IEEE, 86*, 5 (May 1998), 837-852.

[7] Cohen, I., Sebe. N., Garg, A., Chen, L., Huang, T.S., Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding, 91*, 1-2 (Jan/Feb 2003), 160-187.

[8] Cohn, J.F. and Katz, G.S., Bimodal expression of emotion by face and voice. In *Proc. ACM and ATR Workshop on Face and Gesture Recognition and Their Applications*, 1998, 41-44.

[9] Cohn, J. F. and Schmidt, K. L., The timing of facial motion in posed and spontaneous smiles. *Wavelets, Multiresolution and Information Processing, 2*, 2004, 1-12.

[10] De Silva, L.C. and Ng, P.C. Bimodal emotion recognition. In *Proc. Int'l Conf. Face and Gesture Recognition*, 2000, 332-335.

[11] Ekman, P. and Friesen, W.F. The repertoire of nonverbal behavioral categories – origins, usage, and coding. *Semiotica, 1*, 1969, 49-98.

[12] Frank, M.G., Ekman, P. and Friesen, W., Behavioral markers and recognizability of the smile of enjoyment. *Journal of Personality and Social Psychology, 64*, 1 (Jan. 1993), 83-93.

[13] Fridlund, A.J. The new ethology of human facial expression. *The psychology of facial expression*. Russell, J.A. and Fernandez-Dols, J.M., Eds. Cambridge University Press, Cambridge, MA, USA, 1997, 103-129.

[14] Garg, A., Naphade, M., Huang, T.S. Modeling video using input/output Markov models with application to multimodal event detection. *Handbook of Video Databases: Design and Applications,* B. Furth, O. Marques, and B. Furth, Eds., 2003.

[15] Garg, A., Pavlovic, V., Rehg, J. Audio-visual speaker detection using dynamic Bayesian networks. In *Proc. Int'l Conf. Face and Gesture Recognition*, 2000, 384-390.

[16] Garg, A., Pavlovic, V., Rehg, J. Boosted learning in dynamic Bayesian networks for multimodal speaker detection, *Proceedings of the IEEE, 91*, 9 (Sep. 2003), 1355-1369.

[17] Go, H.J., Kwak, K.C., Lee, D.J. and Chun, M.G. Emotion recognition from facial image and speech signal. In *Proc. Conf. of the Society of Instrument and Control Engineers*, 2003, 2890-2895.

[18] Goleman, D. *Emotional Intelligence*. Bantam Books, New York, NY, USA, 1995.

[19] Hu, W., Tan, T., Wang, L., Maybank, S. A survey on visual surveillance of object motion and behaviors, *IEEE Trans. On Systems, Man, and Cybernetics – Part C: Applications and Reviews, 34*, 3 (Aug. 2004), 334-352.

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

[21] Juslin, P.N. and Scherer, K.R. Vocal expression of affect. In *The New Handbook of Methods in Nonverbal Behavior Research*. Harrigan, J., Rosenthal, R. and Scherer, K., Eds. Oxford University Press, Oxford, UK, 2005.

[22] Kanade, T., Cohn, J.F. and Tian, Y. Comprehensive database for facial expression analysis. In *Proc. Int'l Conf. Face and Gesture Recognition*, 2000, 46-53.

[23] Keltner, D. and Ekman, P. Facial expression of emotion. *Handbook of Emotions*. Lewis, M. and Haviland-Jones, J.M., Eds. Guilford Press, New York, NY, USA, 2000, 236-249.

[24] Larsen, R.J. and Diener, E. Promises and problems with the circumplex model of emotion. *Emotion*, vol. 13, *Review of Personality and Social Psychology*, M. S. Clark, Ed., Sage Publications, Newbury Park, USA, 1992, 25-59.

[25] Matsumoto, D. Cultural similarities and differences in display rules. *Motivation and Emotion, 14*, 1990, 195-214.

[26] Norman, D.A., Human-centered design considered harmful, *ACM Interactions, 12*, 4 (July/Aug. 2005), 14-19.

[27] Nwe, T.L., Wei, F.S. and De Silva, L.C., Speaker Dependent Emotional Speech Recognition Using Hidden Markov Models. *Speech Communications, 41*, 4 (Nov. 2003), 603-623.

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

[29] Pan, H., Liang, Z.P., Anastasio, T.J., Huang, T.S., Exploiting the dependencies in information fusion, In Proc. *IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2*, 407-412, 1999.

[30] Pantic, M. and Rothkrantz, L.J.M. Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE, 91*, 9 (Sep. 2003), 1370-1390.

[31] Pantic, M. and Rothkrantz, L.J.M. Case-based reasoning for user-profiled recognition of emotions from face images. In *Proc. Int'l Conf. Multimedia and Expo*, 2004, 391-394.

[32] Pantic, M., Valstar, M.F., Rademaker, R. and Maat, L. Web-based database for facial expression analysis. In *Proc. Int'l Conf. Multimedia and Expo*, 2005. (www.mmifacedb.com)

[33] Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F. and Poggi, I. Embodied Contextual Agent in Information Delivering Application. In *Proc. Int'l Conf. Autonomous Agents & Multi-Agent Systems*, 2002.

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

[35] Pentland, A. Socially aware computation and communication, *IEEE Computer, 38,* 3 (Mar. 2005), 33-40.

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

[37] Picard, R.W., Vyzas, E., Healey, J., Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 10 (Oct. 2001), 1175-1191.

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

[39] Scanlon, P. and Reilly, R.B. Feature analysis for automatic speech reading. In *Proc. Int'l Workshop Multimedia Signal Processing*, 2001, 625-630.

[40] Scherer, K.R. and Ekman, P., Eds., *Handbook of methods in non-verbal behavior research*. Cambridge University Press, Cambridge, USA, 1982.

[41] Sebe. N, Lew, M.S., Cohen, I.,Sun, Y., Gevers, T., Huang, T.S., Authentic facial expression analysis. In *Proc. Int'l Conf. Face and Gesture Recognition*, 2004, 517-522.

[42] Song, M., Bu, J., Chen, C. and Li, N. Audio-visual based emotion recognition – A new approach. In *Proc. Int'l Conf. Computer Vision and Pattern Recognition*, 2004, 1020-1025.

[43] Watson, D., Clark, L.A., Weber, K., Smith-Assenheimer, J., Strauss, M.E. and McCormick, R.A. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology*, *104*, (Jan 1995), 15-25.

[44] Watson, D., Weber, K., Assenheimer, J.S., Clark, L.A., Strauss, M. E. and McCormick, R.A. Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology*, *104*, (Jan 1995), 3-14.

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

[46] Yoshitomi, Y., Kim, S., Kawano, T. and Kitazoe, T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Proc. Int'l Workshop on Robot-Human*, 2000, 178-183.

[47] JAFFE: www.mic.atr.co.jp/~mlyons/**jaffe**.html