Toward an Affect-Sensitive Multimodal Human–Computer Interaction

MAJA PANTIC, MEMBER, IEEE, AND LEON J. M. ROTHKRANTZ

Invited Paper

The ability to recognize affective states of a person we are communicating with is the core of emotional intelligence. Emotional intelligence is a facet of human intelligence that has been argued to be indispensable and perhaps the most important for successful interpersonal social interaction. This paper argues that next-generation human–computer interaction (HCI) designs need to include the essence of emotional intelligence—the ability to recognize a user's affective states—in order to become more human-like, more effective, and more efficient. Affective arousal modulates all nonverbal communicative cues (facial expressions, body movements, and vocal and physiological reactions). In a face-to-face interaction, humans detect and interpret those interactive signals of their communicator with little or no effort. Yet design and development of an automated system that accomplishes these tasks is rather difficult. This paper surveys the past work in solving these problems by a computer and provides a set of recommendations for developing the first part of an intelligent multimodal HCI—an automatic personalized analyzer of a user's nonverbal affective feedback.

Keywords**—***Affective computing, affective states, automatic analysis of nonverbal communicative cues, human–computer interaction (HCI), multimodal human–computer interaction, personalized human–computer interaction.*

I. INTRODUCTION

The exploration of how we as human beings react to the world and interact with it and each other remains one of the greatest scientific challenges. Perceiving, learning, and adapting to the world around us are commonly labeled as "intelligent" behavior. But what does it mean being intelligent? Is IQ a good measure of human intelligence and the best predictor of somebody's success in life? There is now growing research in the fields of neuroscience, psychology, and cognitive science which argues that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely to how we do in life. This range of abilities is called *emotional intelligence* [44], [96] and includes the ability to have, express, and recognize affective states, coupled with the ability to regulate them, employ them for constructive purpose, and skillfully handle the affective arousal of others. The skills of emotional intelligence have been argued to be a better predictor than IQ for measuring aspects of success in life [44], especially in interpersonal communication, and learning and adapting to what is important [10], [96].

When it comes to the world of computers, not all of them will need emotional skills and probably none will need all of the skills that humans need. Yet there are situations where the man–machine interaction could be improved by having machines capable of adapting to their users and where the information about how, when, and how important it is to adapt involves information on the user's affective state. In addition, it seems that people regard computers as social agents with whom "face-to-(inter)face" interaction may be most easy and serviceable [11], [75], [90], [101], [110]. Human–computer interaction (HCI) systems capable of sensing and responding appropriately to the user's affective feedback are, therefore, likely to be perceived as more natural [73], more efficacious and persuasive [93], and more trustworthy [14], [78].

These findings, together with recent advances in sensing, tracking, analyzing, and animating human nonverbal communicative signals, have produced a surge of interest in *affective computing* by researchers of advanced HCI. This intriguing new field focuses on computational modeling of human perception of affective states, synthesis/animation of affective expressions, and design of affect-sensitive HCI.

Indeed, the first step toward an intelligent HCI having the abilities to sense and respond appropriately to the user's affective feedback is to detect and interpret affective states shown by the user in an automatic way. This paper focuses further on surveying the past work done on solving these problems and providing an advanced HCI with one of the key skills of emotional intelligence: the ability to recognize the user's nonverbal affective feedback.

Manuscript received October 25, 2002; revised March 5, 2003. The work of M. Pantic was supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202.

The authors are with the Delft University of Technology, Data and Knowledge Systems Group, Mediamatics Department, 2600 AJ Delft, The Netherlands (e-mail: M.Pantic@cs.tudelft.nl; L.J.M.Rothkrantz@cs.tudelft.nl).

Digital Object Identifier 10.1109/JPROC.2003.817122

A. Other Research Motivations

Besides the research on natural (human-like) HCI, various research areas and technologies brought to bear on the pertinent issues would reap substantial benefits from efforts to model human perception of affective feedback computationally [37], [86]. For instance, automatic recognition of human affective states is an important research topic for video surveillance as well. Automatic assessment of boredom, inattention, and stress will be highly valuable in situations where firm attention to a crucial but perhaps tedious task is essential, such as aircraft control, air traffic control, nuclear power plant surveillance, or simply driving a ground vehicle like a truck, train, or car. An advantage of affect-sensitive monitoring done by a computer is that human observers need not be present to perform privacy-invading monitoring; the automated tool could provide prompts for better performance based on the sensed user's affective states.

Another area where benefits could accrue from efforts toward computer analysis of human affective feedback is the automatic affect-based indexing of digital visual material [46]. A mechanism for detecting scenes/frames which contain expressions of pain, rage, and fear could provide a valuable tool for violent-content-based indexing of movies, video material, and digital libraries.

Other areas where machine tools for analysis of human affective feedback could expand and enhance research and applications include specialized areas in professional and scientific sectors. Monitoring and interpreting affective behavioral cues are important to lawyers, police, and security agents, who are often interested in issues concerning deception and attitude. Machine analysis of human affective states could be of considerable value in these situations where only informal interpretations are now used. It would also facilitate research in areas such as behavioral science (in studies on emotion and cognition), anthropology (in studies on cross-cultural perception and production of affective states), neurology (in studies on dependence between emotional abilities impairments and brain lesions), and psychiatry (in studies on schizophrenia) in which reliability, sensitivity, and precision are currently nagging problems.

B. Outline of the Paper

The paper will begin by examining the context in which affective computing research has arisen and by providing a taxonomy of the pertinent problem domain. The paper will then survey the past work done in tackling the problems of machine detection and interpretation of human affective states. According to the type of human nonverbal interactive signals conveying messages about an affective arousal, two areas will receive particular attention: facial expression analysis and vocal expression analysis. Finally, the paper will discuss the state of the art and consider some challenges and opportunities facing the researchers of affective computing.

II. The Problem Domain

While there is a general agreement that machine sensing and interpretation of human affective feedback would be quite beneficial for a manifold research and application areas, tackling these problems is not an easy task. The main problem areas concern the following.

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

This section discusses basic issues in these problem areas. It begins by examining the body of research literature on the human perception of affective states, which is large, but disunited when it comes to "basic" emotions that can be universally recognized. This lack of consensus implies that the selection of a list of affective states to be recognized by a computer requires pragmatic choices.

The capability of the human sensory system in the detection and understanding of the other party's affective state is explained next. It is meant to serve as an ultimate goal in efforts toward machine sensing and understanding of human affective feedback and as a basis for addressing two main issues relevant to affect-sensitive multimodal HCI: *which* modalities should be integrated and *how* should these be combined.

A. Psychological Issues

The question of how humans perceive affective states has become a concern of crucial importance for the researchers of affective computing. Ironically, the growing interest in affective computing comes at the time when the established wisdom on human affective states is being strongly challenged in the basic research literature.

On one hand, classic psychological research claims the existence of six basic expressions of emotions that are universally displayed and recognized: happiness, anger, sadness, surprise, disgust, and fear [8], [25], [58]. This implies that, apart from verbal interactive signals (spoken words), which are person dependent [43], nonverbal communicative signals (i.e., facial expression, vocal intonations, and physiological reactions) involved in these basic emotions are displayed and recognized cross culturally. For example, a well-known study by Ekman [34], which is commonly used as an important evidence for universals in facial expressions, found that spontaneously displayed, specific facial actions that signal the emotions of fear, sadness, disgust, surprise, and happiness occurred with virtually the same frequency by Japanese and

American subjects in response to watching emotion-inducing films.

On the other hand, there is now a growing body of psychological research that strongly challenges the classical theory on emotion. Russell argues that emotion in general can best be characterized in terms of a multidimensional affect space, rather than in terms of a small number of emotion categories [94], [95]. He also criticizes experimental design flaws applied in classic studies (i.e., using a single corpus of unnatural stimuli and forced-choice response format). Classic studies claim that the basic emotions are hardwired in the sense that some specific neural structures correspond to different emotions. Alternative studies like the study of Ortony and Turner [79] suggest that it is not emotions but some components of emotions that are universally linked with certain communicative displays like facial expressions. Social constructivists like Averill [4] argue, furthermore, that emotions are socially constructed ways of interpreting and responding to particular classes of situations and that they do not explain the genuine feeling (affect). Except for this lack of consensus on the nature of emotion, there is no consensus on how affective displays should be labeled/named. The key issue here, which is in contradiction to the classic studies' emphasis on emotions as a product of evolution, is that of culture dependency: the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent [12], [67], [106], [117]. For further details on issues debated in the basic research on emotion, readers are referred to [22] and [24].

In summary, the available body of basic research literature is excessively fragmented and does not provide firm conclusions that could be safely presumed and employed in studies on affective computing. Due to the unresolved debate concerning the classic emphasis on emotions as a product of evolution on one hand and evidence that they are culture dependent on the other hand, there are no grounds to assume the existence of a set of basic emotions that are displayed and recognized uniformly across different cultures. 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 the situation and the observer. The immediate implication is that pragmatic choices (e.g., application- and user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback.

B. Human Performance

Affective arousal modulates all verbal and nonverbal communicative signals. As shown by Furnas *et al.* [43], it is very difficult to anticipate a person's word choice and the associated intent: even in highly constrained situations, different people choose different words to express exactly the same thing. On the other hand, in usual interpersonal face-to-face interaction, people detect and interpret nonverbal communicative signals in terms of affective states expressed by their communicator with little or no effort [35]. Although the correct recognition of someone's affective state depends on many factors (the speaker's volition to reveal or to disguise his genuine feelings, the attention given to the speaker, the familiarity with the speaker's personality, face, vocal intonation, etc.), humans recognize nonverbal affective cues with apparent ease.

The human sensory system uses multimodal analysis of multiple communication channels to interpret face-to-face interaction and to recognize another party's affective states. A channel is a communication medium while a modality is a sense used to perceive signals from the outside world. Our communication channels are, for example, an auditory channel that carries speech and vocal intonation communicative signals and a visual channel that carries facial expression signals. The senses of sight, hearing, and touch are examples of modalities. Detection of another party's affective states in usual face-to-face interaction involves simultaneous usage of numerous channels and combined activation of multiple modalities. Hence, the analysis of the communicator's affective feedback becomes highly flexible and robust. Failure of one channel is recovered by another channel and a message in one channel can be explained by that in another channel (e.g., a mouth expression that might be interpreted as a smile will be seen as a display of sadness if at the same time we can see tears and hear sobbing).

The abilities of the human sensory system define, in some way, the expectations for a multimodal affect-sensitive analyzer. Although it may not be possible to incorporate all features of the human sensory system into an automated system (due to the complexity of the phenomenon, which involves an intricate interplay of knowledge, thoughts, language, and nonverbal behavioral cues), the affect-recognition capability of the human sensory system can serve as the ultimate goal and a guide for defining design recommendations for a multimodal analyzer of human affective states.

C. A Taxonomy of the Problem Domain

A taxonomy of the automatic human affect analysis problem domain can be devised by considering the following issues.

1) Which channels of information, corresponding to which human communicative channels, should be integrated into an automatic affect analyzer?
2) How should the data carried by multiple channels be fused to achieve a human-like performance in recognizing affective states?
3) How should temporal aspects of the information carried by multiple channels be handled?
4) How can automated human affect analyzers be made more sensitive to the context in which they operate (i.e., to the current situation and user)?

Modalities: Though one could expect that automated human affect analyzers should include all human interactive modalities (sight, sound, and touch) and should analyze all nonverbal interactive signals (facial expressions, vocal expressions, body gestures, and physiological reactions),

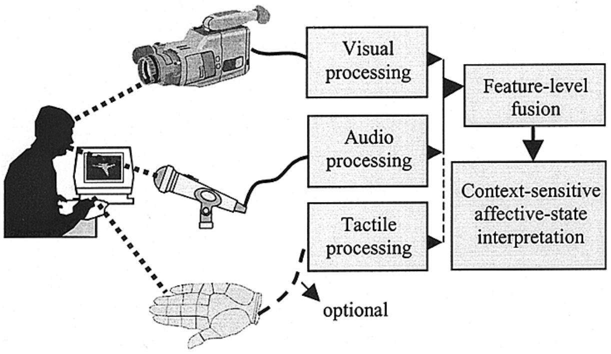


Fig. 1. Architecture of an "ideal" automatic analyzer of human affective feedback.

the reported research does not confirm this finding. The visual channel carrying facial expressions and the auditory channel carrying vocal intonations are widely thought of as most important in the human recognition of affective feedback [24]. According to Mehrabian [71], whether the listener feels liked or disliked depends only for 7% on the spoken word, for 38% on vocal utterances, and for even 55% on facial expressions. This indicates that, while judging someone's affective state, people rely less on body gestures and physiological reactions displayed by the observed person; they rely mainly on his facial expressions and vocal intonations. As far as body gestures are concerned, as much as 90% of body gestures are associated exclusively with speech [68]. Hence, it seems that they play a secondary role in the human recognition of affective states. As far as physiological signals are concerned, people commonly neglect these, since they cannot sense them at all times. Namely, in order to detect someone's clamminess or heart rate, the observer should be in a physical contact (touch) with the observed person. Yet the research in psychophysiology has produced firm evidence that affective arousal has a range of somatic and physiological correlates such as pupillary diameter, heart rate, skin clamminess, temperature, respiration velocity, etc. [12]. However, the integration of tactile channels carrying physiological reactions of the monitored subject into an automated human affect analyzer requires wiring the subject, which is usually perceived as being uncomfortable and unpleasant. Though the recent advent of nonintrusive sensors and wearable computers opened up possibilities for less invasive physiological sensing [64], yet another problem persists: currently available skin sensors are very fragile, and the accuracy of the measurements is commonly affected by hand washing and the amount of gel used [12]. In summary, automated affect analyzers should at least combine modalities for perceiving facial and vocal expressions of attitudinal states. Optionally, if provided with robust, nonintrusive sensory equipment, they could also include the modality for perceiving affective physiological reactions (see Fig. 1).

Multisensory Information Fusion: The performance of a multimodal analyzer of human affective feedback is not only influenced by the different types of modalities to be integrated; the abstraction level at which these modalities are to be integrated/fused and the technique which is to be applied to carry out multisensory data fusion are clearly of the utmost importance as well. If the goal is to ensure that HCI approaches the naturalness of human–human interaction [65], [105], then the ideal model for multisensory information fusion is the human handling of the multimodal information flow. Insight into how the modalities of sight, sound, and touch are combined in human–human interaction can be gained from neurological studies on fusion of sensory neurons [108]. Three concepts relevant to multimodal fusion can be distinguished.

1) $1+1>2$: The response of multisensory neurons can be stronger for multiple weak input sensory signals than for a single strong signal.
2) *Context dependency*: The fusion of sensory signals is modulated according to the signals received from the cerebral cortex: depending on the sensed context, different combinations of sensory signals are made.
3) *Handling of discordances*: Based upon the sensed context, sensory discordances (malfunctioning) are either handled by fusing sensory observations without any regard for individual discordances (e.g., when a fast response is necessary), or by attempting to recalibrate discordant sensors (e.g., by taking a second look), or by suppressing discordant and recombining functioning sensors (e.g., when one observation is contradictory to another).

Hence, humans simultaneously employ the tightly coupled modalities of sight, sound, and touch [68]. As a result, analysis of the perceived information is highly robust and flexible. Several studies confirmed that this tight coupling of different modalities persists when the modalities are used for multimodal HCI [17], [18], [81].

A question remains, nevertheless, as to whether such a tight coupling of multiple modalities can be achieved using the theoretical and computational apparatus developed in the field of sensory data fusion [26], [45]. As illustrated in Fig. 2, fusion of multisensory data can be accomplished at three levels: data, feature, and decision level. Data-level fusion involves integration of raw sensory observations and can be accomplished only when the observations are of the same type. Since the monitored human interactive signals are of different nature and are sensed using different types of sensors, data-level fusion is, in principle, not applicable to multimodal HCI. Feature-level fusion assumes that each stream of sensory information is first analyzed for features and then the detected features are fused. Feature-level fusion retains less detailed information than data-level fusion, but it is also less prone to noise and sensor failures, and, most importantly, it is the most appropriate type of fusion for tightly coupled and synchronized modalities. Though many feature-level techniques like Kalman fusion, artificial neural networks (ANN) based fusion, and hidden Markov models (HMM) based fusion have been proposed [26], [105], decision-level (i.e., interpretation-level) fusion is the approach applied most often for multimodal HCI [65], [88], [105]. This practice may follow from experimental studies that that have shown that a late integration approach (i.e., a decision-level fusion) might

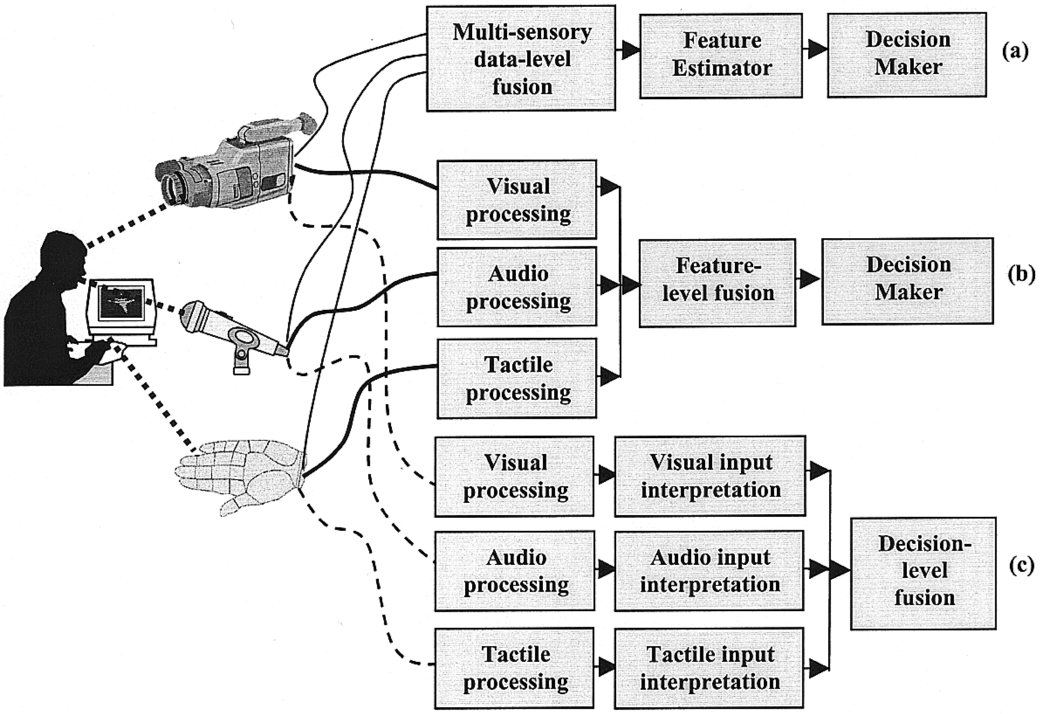


Fig. 2. Fusion of multiple sensing modalities. (a) Data-level fusion integrates raw sensory data. (b) Feature-level fusion combines features from individual modalities. (c) Decision-level fusion combines data from different modalities at the end of the analysis.

provide higher recognition scores than an early integration approach (e.g., [100]). The differences in the time scale of the features from different modalities and the lack of a common metric level over the modalities add and abet the underlying inference that the features from different modalities are not sufficiently correlated to be fused at the feature level. Yet it is almost certainly incorrect to use a decision-level fusion, since people display audio, visual, and tactile interactive signals in a complementary and redundant manner. In order to accomplish a multimodal analysis of human interactive signals acquired by multiple sensors, which resembles human processing of such information, input signals cannot be considered mutually independent and cannot be combined only at the end of the intended analysis. The input data should be processed in a joint feature space and according to a context-dependent model.

Temporal Information: Each observation channel, in general, carries information at a wide range of time scales. At the longest time scale are *static and semipermanent signals* like bony structure, fat deposits, metabolism, and phonetic peculiarities like accent. Those signals provide a number of social cues essential for interpersonal communication in our everyday life. They mediate person identification and gender, and provide clues on a person's origin and health. At shorter time scales are *rapid behavioral signals* that are temporal changes in neuromuscular and physiological activity that can last from a few milliseconds (e.g., a blink) to minutes (e.g., the respiration rate) or hours (e.g., sitting). Among the types of messages communicated by rapid behavioral signals are the following:

1) affective and attitudinal states (e.g., joy, inattention, frustration);
2) emblems (i.e., culture-specific communicators like a wink or thumbs up);
3) manipulators (i.e., actions used to act on objects in the environment or self-manipulative actions like scratching and lip biting);
4) illustrators (i.e., actions accompanying speech such as finger pointing and raised eyebrows);
5) regulators (i.e., conversational mediators such as the exchange of a look, palm pointing, head nods, and smiles).

In general, rapid behavioral signals can be recognized from 40-ms video frames and 10-ms audio frames. Yet the ability to discriminate subtle affective expressions requires a comparison over time. Namely, changes in human behavior observed in a time instance may be misinterpreted if the temporal pattern of those changes is not taken into account. For example, a rapid frown of the eyebrows, denoting a difficulty with understanding discussed matters, could be misinterpreted for an expression of anger if the temporal pattern of behavioral changes, indicating attentiveness to the discussed subject, is not taken into account. In addition, performing both time-instance and time-scale analysis of the information carried by multiple observation channels could be extremely useful for handling sensory discordances and ambiguities in general. This assumption bears on the existence of a certain grammar of neuromuscular actions and physiological reactions: only a certain subclass of these actions/reactions with respect to the currently encountered action/reaction (time instance) and the previously observed actions/reactions (time scale) is plausible. Thus, by considering the previously observed affective states (time scale) and the current data carried by functioning observation channels (time instance), a statistical prediction might be derived about both the current affective state and the information that have been lost due to malfunctioning or inaccuracy of a particular sensor.

Context Dependency: Rapid behavioral signals can be subdivided further into the following classes:

1) reflex actions under the control of afferent input (e.g., backward pull of the hand from a source of heat, scratching, squinting the eyes when facing the sun, etc.);
2) rudimentary reflex-like (impulsive) actions that appear to be controlled by innate motor programs and might accompany affective expressions (e.g., wide-open eyes by encountering an unexpected situation) and less differentiated information processing (e.g., orienting in space);
3) adaptable, versatile, and culturally and individually variable spontaneous actions that appear to be mediated by learned motor programs and form a firm part of affective expressions (e.g., smile of greeting or at a joke, raised eyebrows in wonder, etc.);
4) malleable and culturally and individually variable intentional actions that are controlled by learned motor programs and form a part of affective expressions (e.g., uttering a spoken message by raising the eyebrows, shaking hands to get acquainted with someone, tapping the shoulder of a friend in welcome, etc.).

Thus, some of the rapid behavioral signals demand relatively little of a person's information processing capacity and are free of deliberate control for their evocation, while others demand a lot of processing capacity, are consciously controlled, and are governed by complex and culturally specific rules for interpersonal communication. While some rapid behavioral signals belong exclusively to one class (e.g., scratching), others may belong to any of the classes (e.g., squinted eyes). It is crucial to determine to which class a shown behavioral signal belongs, since this influences the interpretation of the observed signal. For instance, squinted eyes may be interpreted as sensitivity of the eyes if this action is a reflex, as an expression of hate if this action is displayed unintentionally, or as an illustrator of friendly anger on friendly teasing if this action is displayed intentionally. To determine the class of an observed rapid behavioral signal and to interpret it in terms of affective/attitudinal states, one must know the context in which the observed signal has been displayed. In other words, it is necessary to know the interpretation of the observed behavioral signals that the expresser himself associates with those signals in the given situation (i.e., given the expresser's current environment and task).

Ideal Human Affect Analyzer: In summary, we conceive an "ideal" automatic analyzer of human nonverbal affective feedback to be able to emulate at least some of the capabilities of the human sensory system and, in turn, to be the following (see Fig. 1):

1) multimodal (modalities: facial expressions, vocal intonations, and physiological reactions);
2) robust and accurate (despite auditory noise, the frailness of skins sensors, occlusions, and changes in viewing and lighting conditions);
3) generic (independent of variability in subjects' physiognomy, sex, age, and ethnicity);
4) sensitive to the dynamics (time evolution) of displayed affective expressions (performing time-instance and time-scale analysis of the sensed data, previously combined by a multisensory feature-level data fusion);
5) context-sensitive (performing application- and task-dependent data interpretation in terms of user-profiled affect/attitude interpretation labels).

III. The State of the Art

This section will survey current state of the art in the machine analysis of the human affective feedback problem domain. Rather than presenting an exhaustive survey, this section focuses on the efforts recently proposed in the literature that had not been reviewed elsewhere or had the greatest impact on the community (as measured by, e.g., coverage of the problem domain, citations, and received testing).

Relatively few of the existing works combine different modalities into a single system for human affective state analysis. Examples are the works of Chen *et al.* [15], [16], De Silva and Ng [29], and Yoshitomi *et al.* [122], who investigated the effects of a combined detection of facial and vocal expressions of affective states. Virtually all other existing studies treat various human affective cues separately and present approaches to automatic single-modal analysis of human affective feedback.

Though a tactile computer-sensing modality for more natural HCI has been explored recently with increasing interest [65], [77] and although the research in psychophysiology has produced firm evidence that affective arousal has a range of somatic and physiological correlates [12], only a single work aimed at automatic analysis of affective physiological signals has been found in the existing body of literature: the work presented by Picard *et al.* [91]. The lack of interest in this research topic might be in part because of the lack of interest by research sponsors and in part because of the manifold of related theoretical and practical open problems. The pertinent problematic issues concern the following: the application of haptic technology might have a profound impact on the users' fatigue if done improperly [77]; currently available wearable sensors of physiological reactions imply wiring the subject, which is usually experienced as uncomfortable; skin sensors are very fragile and their measuring accuracy is easily affected [12].

The work of Picard *et al.* [91] concerns automatic recognition of eight user-defined affective states (neutral, anger, hate, grief, platonic love, romantic love, joy, and reverence) from a set of sensed physiological signals. Data have been collected over a period of 32 days from an actress intentionally expressing eight affective states during daily sessions (data obtained in 20 days have been used for further experiments). Five physiological signals have been recorded: electromyogram from jaw (coding the muscular tension of the jaw), blood volume pressure (BVP), skin conductivity, respiration, and heart rate calculated from the BVP. A total of 40 features has been used: 30 statistical features (for each raw signal, they calculated six statistical features such as mean and standard deviation) and ten features like mean slope of the skin conductivity, heart rate change, and power spectral density characteristics of the respiration signal. For emotional classification, an algorithm combining the sequential floating forward search and the Fisher projection has been used, which achieves an average correct recognition rate of 81.25%. As reported by Picard *et al.* , the features extracted from the raw physiological signals were highly dependent on the day the signals were recorded. Also the signals have been recorded in short (3 min) sessions.

The survey presented here is divided further into three parts. The first part is dedicated to the work done on the automation of human affect analysis from face images, while the second part explores and compares automatic systems for recognition of human affective states from audio signal. The third part of this survey focuses on the past efforts toward automatic bimodal analysis of human affective feedback from facial and vocal expressions.

A. Automatic Affect Recognition From Face Images

The major impulse to investigate automatic facial expression analysis comes from the significant role of the face in our emotional and social lives. The face provides conversational and interactive signals which clarify our current focus of attention and regulate our interactions with the surrounding environment and other persons in our vicinity [95]. As noted previously, facial displays are our direct, naturally preeminent means of communicating emotions [58], [95]. Automatic analyzers of subtle facial changes, therefore, seem to have a natural place in various vision systems including the automated tools for psychological research, lip reading, videoconferencing, animation/synthesis of life-like agents, and human-behavior-aware next-generation interfaces. It is this wide range of principle driving applications that has caused an upsurge of interest in the research problems of machine analysis of facial expressions. For exhaustive surveys of the pertinent problem domain, readers are referred to the following: Samal and Iyengar [97] for an overview of early works; Tian *et al.* [112] for a review of techniques for detecting micro facial actions [action units (AUs)]; and Pantic and Rothkrantz [85] for a survey of current efforts.

The problem of machine recognition of human affective states from images of faces includes three subproblem areas: finding faces, detecting facial features, and classifying these data into some affect categories.

The problem of *finding faces* can be viewed as a segmentation problem (in machine vision) or as a detection problem (in pattern recognition). Possible strategies for face detection vary a lot, depending on the type of input images. The existing systems for facial expression analysis process either static *face* images or *face* image sequences. In other words, current studies assume, in general, that the presence of a face in the scene is ensured. Posed portraits of faces (uniform background and good illumination) constitute input data processed by the majority of the current systems. Yet, in many instances, the systems do not utilize a camera mounted on the subject's head, as proposed by Otsuka and Ohya [80] and by Pantic [84], [87], which ascertains correctness of that assumption. Though much progress has been recently made in the development of vision systems for robust face detection in arbitrary scenes [120], except the works reported in [39] and [48], presently existing systems for facial affect recognition do not perform automatic face detection in an arbitrary scene.

The problem of *facial feature extraction* from input images may be divided into at least four dimensions.

1) Are the features extracted in an automatic way?
2) Is temporal information (image sequence) used?
3) Are the features holistic (spanning the whole face) or analytic (spanning subparts of the face)?
4) Are the features view- or volume based [two-dimensional (2-D) or three-dimensional (3-D)]?

Given this glossary, most of the proposed approaches to human affect analysis in facial images are directed toward automatic, static, analytic, 2-D facial feature extraction. Still, many of the proposed systems do not extract facial information in an automatic way (see Table 1). Although the techniques for facial affect classification employed by these systems are relevant to the present goals, the systems themselves are of limited use for machine human affect analysis, as analyses of human interactive signals should be fully automatic and preferably achieved in real time to obtain fluent, tight, and efficient HCI. The approaches to automatic facial-data extraction utilized by the existing systems include analyses of:

1) facial motion (e.g., [39], [80]; see Table 1);
2) holistic spatial pattern (e.g., [33], [48], [53]);
3) analytic spatial pattern (e.g., [21], [32], [87]).

In many instances, strong assumptions are made to make the problem of facial feature detection more tractable (e.g., images contain portraits of faces with no facial hair or glasses, the illumination is constant, the subjects are young and of the same ethnicity). Few of the current systems deal with rigid head motions (examples are the systems proposed by Hong *et al.* [48], Colmenarez *et al.* [21], and Ebine *et al.* [32]) and only the method of Essa and Pentland [39] can handle distractions like facial hair (beard, moustache) and glasses. None of the automated facial affect analyzers proposed in the literature up to date "fills in" missing parts of the observed face, that is, none "perceives" a whole face when a part of it is occluded (e.g., by a hand or some other object). Also, though the conclusions generated by an automated facial expression analyzer are affected by input data certainty, except for the system proposed by Pantic [87], none existing system for automatic facial expression analysis calculates the output data certainty based upon an input data certainty.

Eventually, automated facial affect analyzers should terminate their execution by translating the extracted facial features into a *description of the shown affective state*. This description should be identical, or at least very close, to a human's description of the examined facial affect. As already explained in Section II, human interpretation of (facial) affective feedback depends upon both the context in which the pertinent affective feedback has been observed and the dynamics (time evolution) of displayed affective expressions. Hence, in order

Table 1
Properties of the Proposed Approaches to Automatic Affect Recognition from Face Images

Reference	*Properties*																Test results
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Analysis from static facial images																	
Cohen '02 [19]	●	●	-	-	×	-	×	●	-	×	●	7	×	×	×	×	12600 frames, 5 subjects, Correct: 65.5%
Colmenarez '99 [21]	-	●	T	-	●	×	●	●	●	×	●	6	×	×	×	-	5970 frames, 18 subjects, Correct: 86%
Cottrell '91 [23]	×	●	●	-	-	×	×	×	-	-	●	8	×	×	×	×	40 images, 5 subjects Correct: 40%
Ebine '00 [32]	●	●	U	×	×	×	●	●	●	×	●	7	×	●	●	×	320 frames, 1 subject Correct: unknown
Edwards '98 [33]	●	●	T	●	●	×	×	●	-	×	●	7	×	×	×	×	200 images, 25 subjects Correct:74%
Hara '97 [47]	●	●	U	-	×	×	●	●	-	×	●	6	×	×	×	●	90 images, 15 subjects Correct: 85%
Hong '98 [48]	●	×	T	×	●	×	●	●	●	×	●	7	×	×	×	●	175 images, 25 subjects, Correct: 81%
Huang '97 [49]	●	●	U	×	×	×	●	●	-	×	●	6	×	×	×	×	90 images, 15 subjects Correct: 85%
Ji '99 [53]	●	●	U	-	×	×	●	●	●	×	●	6	×	●	●	×	37 images, # subjects unknown, Correct: 70%
Kearney '93 [57]	×	●	●	-	-	×	×	×	-	-	●	n	●	●	×	×	17 images, 1 subject, Correct: 91%
Kobayashi '92 [60]	×	●	●	-	-	×	×	×	-	-	●	6	×	●	●	×	54 images, 9 subjects, Correct: 80%
Kurozumi '99 [61]	×	●	●	-	-	×	×	×	-	-	●	7	×	×	×	×	7 images, 1 subject Correct:: 62%
Lyons '99 [63]	×	●	●	-	-	×	×	×	-	-	●	7	×	×	×	×	193 images, 9 women Correct: 92%
Padget '96 [82]	×	●	●	-	-	×	×	×	-	-	●	7	×	×	×	×	84 Ekman's photos Correct: 86%
Pantic '00 [84]	●	×	●	×	●	×	●	×	●	×	●	7	×	●	●	×	256 images, 8 subjects Correct: 91%
Pantic '01 [87]	●	×	●	×	●	×	●	●	●	●	●	n	●	●	●	×	196 images, 8 subjects Correct: 97%
Sebe '02 [104]	●	●	-	-	×	-	×	●	-	×	●	7	×	×	×	×	12600 frames, 5 subjects, Correct: 63.5%
Zhang '98 [124]	×	●	●	-	-	×	×	×	-	-	●	7	×	●	●	×	213 images, 9 women Correct: 90%
Analysis from facial image sequences																	
Black '97 [9]	●	●	-	●	●	×	×	●	×	×	●	6	×	×	●	×	70 sequences, 40 subjects Correct: 88%
Essa '97 [39]	●	×	●	●	-	●	●	●	●	×	●	4	×	×	×	●	30 sequences, 8 subjects Correct: 98%
Fellenz '00 [40]	×	●	●	-	-	×	×	×	-	-	●	6	×	×	×	×	20 sequences, # subjects unknown, Correct: 75%
Kimura '97 [59]	●	●	U	●	×	×	●	●	●	×	●	3	×	×	●	-	6 sequences, 1 subject unsuccessful testing
Mase '91 [66]	×	●	U	×	×	×	×	●	×	×	●	4	×	×	×	×	30 sequences, 1 subject Correct: 80%
Otsuka '98 [80]	●	×	U	-	●	×	-	●	×	×	●	6	×	×	×	×	120 sequences, 2 subjects Correct: unknown
Wang '98 [116]	●	●	U	×	×	×	×	●	-	×	●	3	×	×	●	×	29 sequences, 8 subjects Correct: 95%
Yacoob '94 [119]	●	●	-	-	×	×	×	●	●	×	●	7	×	×	●	×	46 sequences, 32 subjects Correct: 88%
Zhu '02 [126]	●	●	U	-	-	×	×	×	-	-	●	4	×	×	●	×	31 sequences, 10 subjects Correct: 94%

Legend: ● = "yes", × = "no", - = missing entry, U = unknown, T = handles images of subjects on which it has been trained

to achieve a human-like interpretation of shown facial affect, pragmatic choices (i.e., application-, user-, and task-profiled time-scale-dependent choices) must be made regarding the selection of moods and affective/attitudinal states to be recognized by a facial affect analyzer. For instance, if the intended application is the monitoring of a nuclear power plant operator, then the facial affect analyzer to be deployed will probably be aimed at discerning stress and inattention. In addition, facial affect display and interpretation rules differ from culture to culture and may even differ from person to person [95]. Hence, the interpretation of the user's facial affect is strongly dependent upon the affect labels that the pertinent user associates with the patterns of his facial behavior. So, for example, if the nuclear power plant operator distinguishes frustration, stress, and panic as variations of the generic category "stress," then the facial affect analyzer to be deployed for the surveillance of his affect/attitude/mood should adapt to these interpretation categories. Nonetheless, except of the automated system proposed by Pantic [87], which classifies facial expressions into multiple quantified user-defined interpretation classes, all the existing facial expression analyzers perform facial expression classification into a number of the six basic emotion categories. The classification techniques used by the existing systems include:

1) template-based classification in static images (e.g., [21], [33], [48], [61], [63]);
2) template-based classification in image sequences (e.g., [39], [80]);
3) (fuzzy) rule-based classification in static images (e.g., [53], [57], [84]);
4) (fuzzy) rule-based classification in image sequences (e.g., [9], [40], [119]);
5) ANN-based classification (e.g., [47], [60], [82]);
6) HMM-based classification (e.g., [80], [126]);
7) Bayesian classification (e.g., [19], [104]).

Further, as mentioned previously, a shown facial expression may be misinterpreted if the current task of the user is not taken into account. For example, a frown may be displayed by the speaker to emphasize the difficulty of the currently discussed problem and it may be shown by the listener to denote that he did not understand the problem at issue. Yet existing facial affect analyzers do not perform a task-dependent interpretation of shown facial behavior. Finally, the timing (dynamics) of facial expressions is a critical factor in interpretation of facial affect [95]. However, current systems for facial affect recognition do not analyze extracted facial information on different time scales. Proposed interframe analyses are either used to handle the problem of partial data or to achieve detection of facial expression time markers (onset, apex, and offset; short time scale). Consequently, automatic recognition of the expressed mood and attitude (longer time scales) is still not within the range of current facial affect analyzers.

Table 1 summarizes the features of existing systems for facial affect recognition with respect to the following issues.

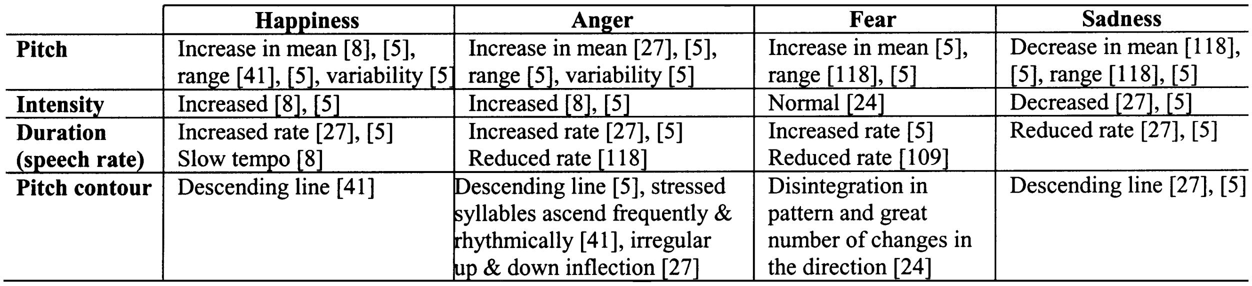
1) Is the input image provided automatically?
2) Is the presence of the face assumed?
3) Is the performance independent of variability in subjects' sex, physiognomy, age, and ethnicity?
4) Can variations in lighting be handled?
5) Can rigid head movements be handled?
6) Can distractions like glasses and facial hair be handled?
7) Is the face detected automatically?
8) Are the facial features extracted automatically?
9) Can inaccurate input data be handled?
10) Is the data uncertainty propagated throughout the facial information analysis process?
11) Is the facial expression interpreted automatically?
12) How many interpretation categories (labels) have been defined?
13) Are the interpretation labels user profiled?
14) Can multiple interpretation labels be scored at the same time?
15) Are the interpretation labels quantified?
16) Is the input processed in real time?

● stands for "yes," × stands for "no," and – represents a missing entry. A missing entry either means that the matter at issue has not been reported or that the pertinent matter is not applicable to the system in question. The inapplicable issues, for instance, are the issues of dealing with variations in lighting, rigid head movements, and inaccurate facial information in the cases where the input data were hand measured (e.g., by [23]). Further, the value "U" of column 3 indicates that it is unknown whether the system in question can handle images of an arbitrary subject (usually, this is a consequence of the fact that the pertinent system has not been tested on images of unknown subjects and/or on images of subjects of different ethnicity). The value "T" of column 3 indicates that the surveyed system cannot handle images of subjects for which it has not been previously trained. Finally, the value "n" of column 12 indicates that the system in question is not limited to a predefined, rigid number of interpretation categories; it is "dynamic" in the sense that new user-defined interpretation labels can be learned with experience.

In summary, current facial affect machine analysis research is largely focused at attempts to recognize a small set of posed prototypic facial expressions of basic emotions from portraits of faces or nearly frontal view face image sequences under good illumination. Yet, given that humans detect six basic emotional facial expressions with an accuracy ranging from 70% to 98% [7], it is rather significant that the automated systems achieve an accuracy of 64% to 98% when detecting three to seven emotions deliberately displayed by 5 to 40 subjects. 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.

In spite of repeated references to the need for a readily accessible, reference set of images (image sequences) that could provide a basis for benchmarks for efforts in automatic facial affect analysis, no database of images exists that is shared by all diverse facial expression research communities [37], [88], [85]. In general, only isolated pieces of such a facial database exist, each of which has been made and exploited by a particular facial research community. To our best knowledge, the only example of a facial database used by more than one facial research community is the unpublished database of Ekman–Hager facial action exemplars [38]. It has been used by Bartlett *et al.* [6], Donato *et al.* [31], and Tian *et al.* [112] to train and test their methods for detecting facial micro actions (i.e., facial muscle actions) from face image sequences. The facial database made publicly available, but used only by Tian *et al.* up to now, is the Cohn–Kanade AU-coded face expression image database [55]. Like the database of Ekman–Hager facial action exemplars, it can be used as a basis for benchmarks for efforts in the research area of facial micro action detection from face image sequences.

Table 2
The Speech Correlates of the Prototypic ("Basic") Emotions of Happiness, Anger, Fear, and Sadness (Only the Auditory Variables Commonly Reported in Psycholinguistic Research Studies Have Been Included)



	Happiness	Anger	Fear	Sadness
Pitch	Increase in mean [8], [5], range [41], [5], variability [5]	Increase in mean [27], [5], range [5], variability [5]	Increase in mean [5], range [118], [5]	Decrease in mean [118], [5], range [118], [5]
Intensity	Increased [8], [5]	Increased [8], [5]	Normal [24]	Decreased [27], [5]
Duration (speech rate)	Increased rate [27], [5] Slow tempo [8]	Increased rate [27], [5] Reduced rate [118]	Increased rate [5] Reduced rate [109]	Reduced rate [27], [5]
Pitch contour	Descending line [41]	Descending line [5], stressed syllables ascend frequently & rhythmically [41], irregular up & down inflection [27]	Disintegration in pattern and great number of changes in the direction [24]	Descending line [27], [5]

This glaring lack of a common testing resource, which forms a major impediment to comparing, resolving, and extending the issues concerned with automatic facial expression analysis and understanding, also represents our main incentive to avoid labeling some of the surveyed systems as being better than others. We believe that a well-defined, validated, and commonly used database of images of faces (both still and motion) is a necessary prerequisite for ranking the performances of the proposed facial affect analyzers in an objective manner. Since such a benchmark database has not been established yet, the reader is left to rank the surveyed systems according to his own priorities and based on the overall properties of these systems that have been summarized in Table 1.

B. *Automatic Affect Recognition From Audio Signal*

The auditory aspect of a communicative message carries various kinds of information. If we consider the verbal part (strings of words) only, without regarding the manner in which it was spoken, we might miss important aspects of the pertinent utterance and even misunderstand the spoken message by not attending to the nonverbal aspect of the speech. Nevertheless, in contrast to spoken language processing, which has witnessed significant advances in the last decade [54], the processing of "emotional" speech has not been widely explored by the auditory research community. Yet recent data show that the accuracy of automated speech recognition, which is about 80% to 90% for neutrally spoken words, tends to drop to 50% to 60% if it concerns emotional speech [107]. The same has been shown in the case of automatic speaker verification systems [102]. Although such findings triggered some efforts at automating human affect recognition from speech signal, most researchers in this field have focused on synthesis of emotional speech [72].

The problem of vocal affect analysis includes two subproblem areas: specifying auditory features to be estimated from the input audio signal, and classifying the extracted data into some affect categories.

The research in psychology and psycholinguistics provides an immense body of results on acoustic and prosodic *features* which can be used to encode affective states of a speaker (e.g., [5], [24], [42]). Table 2 lists the speech correlates of the archetypal emotions of happiness, anger, fear, and sadness that have been commonly reported in these studies. The speech measures which seem to be reliable indicators of these "basic" emotions are the continuous acoustic measures, particularly pitch-related measures (range, mean, median, variability), intensity, and duration. The works on automatic affect-sensitive analysis of vocal expressions presented in the literature up to date commonly use this finding. The auditory features usually estimated from the input audio signal are (see Table 3):

1) *pitch* (the fundamental frequency of the acoustic signal delimited by the rate at which vocal cords vibrate);
2) *intensity* (the vocal energy);
3) *speech rate* (the number of words spoken in a time interval; words can be identified from time-varying spectra of harmonics, which are generated by vocal cord vibrations and filtered as they pass through the mouth and nose);
4) *pitch contour* (pitch variations described in terms of geometric patterns);
5) *phonetic features* (features that deal with the types of sounds involved in speech, such as vowels and consonants and their pronunciation).

As mentioned previously, in order to accomplish a human-like *interpretation of perceived vocal affective feedback*, pragmatic choices (i.e., application-, user-, and task-profiled time-scale-dependent choices) must be made regarding the selection of affective/attitudinal states and moods to be recognized by a vocal affect analyzer. Nevertheless, existing automated systems for auditory analysis of human affective feedback do not perform a context-sensitive analysis (i.e., application-, user-, and task-dependent analysis) of the input audio signal. Also, they do not analyze extracted vocal expression information on different time scales. Proposed interframe analyses are used either for the detection of suprasegmental features (such as the pitch and intensity over the duration of a syllable, word, or sentence, [92]) or for the detection of phonetic features [50]. Computer-based recognition of moods and attitudes (longer time scales) from input audio signal remains, therefore, a significant research challenge. Virtually all the existing work on automatic vocal affect analysis performs singular classification of input audio signals into a few emotion categories such as anger, irony, happiness, sadness/grief, fear,

Table 3
Properties of the Proposed Approaches to Automatic Affect Recognition From Audio Signal

	Properties																
Reference	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Test results
Amir '98 [2]	●	✕	●	●	●	●	●	✕	●	✕	●	5	✕	●	●	●	spontaneous speech, 2 subjects,Correct: unknowı
Banse '96 [5]	✕	●	●	●	●	●	✕	✕	●	✕	●	14	✕	✕	✕	U	2 sentences (5 words), 12 subjects, Correct: 53%
Dellaert '96 [30]	U	✕	●	●	✕	●	●	✕	✕	✕	●	4	✕	✕	✕	U	50 short sentences, 5 subjects, Correct: 80%
Fellenz '00 [24], [40]	U	U	●	●	●	✕	●	✕	●	●	●	4	✕	✕	✕	✕	test data set unknown Correct: unknown
Huber '00 [50]	●	●	●	✕	✕	✕	✕	●	●	✕	●	2	✕	✕	✕	✕	spontaneous speech, 39 subjects, Correct:66%
Kang '00 [56]	✕	✕	●	●	●	✕	✕	✕	✕	✕	●	6	✕	✕	✕	✕	5 words, 8 subjects Correct: 85%
Li '98 [62]	●	✕	●	●	●	✕	●	✕	✕	✕	●	6	✕	✕	✕	✕	20 sentences (7 words), 5 subjects, Correct: 62%
Nakatsu '00 [74]	●	●	●	●	●	✕	✕	●	●	✕	●	8	✕	●	●	✕	100 words, 100 subjects, Correct: 50%
Nwe '01 [76]	U	T	●	✕	●	✕	✕	✕	✕	✕	●	6	✕	✕	✕	✕	90 short sentences, 2 subjects, Correct: 66%
Petrushin '00 [89]	●	T	●	●	●	●	✕	✕	●	✕	●	5	✕	✕	✕	✕	4 sentences (4-6 words), 30 subjects, Correct: 66%
Polzin '00 [92]	✕	T	●	●	●	✕	✕	✕	●	✕	●	5	✕	✕	✕	✕	50 sentences (2-12 words) subjects, Correct: 73%
Sato '01 [98]	U	T	●	●	✕	✕	●	✕	✕	✕	●	4	✕	✕	✕	✕	1 word, 13 subjects Correct: 74%
Tosa '96 [113]	●	T	●	●	●	✕	●	●	●	✕	●	8	✕	✕	✕	✕	100 words, 10 subjects Correct: 60%
Zhao '00 [125]	✕	✕	●	●	●	✕	✕	✕	●	✕	●	4	✕	✕	✕	✕	4 sentences, 5 males Correct: 90%

Legend: ● = "yes", ✕ = "no", U = unknown, T = handles speech samples of (known) subjects on which it has been trained

disgust, surprise, and affection [24]. Utilized classification techniques include:

- ANNs (e.g., [40], [50], [74], [89], [113]);
- HMMs (e.g., [56], [76], [92]);
- Gaussian mixture density models (e.g., [62]);
- Fuzzy membership indexing (e.g., [2]);
- maximum-likelihood Bayes classifiers (e.g., [30], [56]).

Table 3 summarizes the properties of current systems for vocal affect analysis with respect to the following.

1) Can nonprofessionally spoken input samples be handled?
2) Is the performance independent of variability in subjects, their sex and age?
3) Are the auditory features extracted automatically?
4) Are the pitch-related variables utilized?
5) Is the vocal energy (intensity) utilized?
6) Is the speech rate utilized?
7) Are pitch contours utilized?
8) Are phonetic features utilized?
9) Are some other auditory features utilized?
10) Can inaccurate input data be handled?
11) Is the extracted vocal expression information interpreted automatically?
12) How many interpretation categories (labels) have been defined?
13) Are the interpretation labels scored in a context-sensitive manner (application-, user-, task-profiled manner)?
14) Can multiple interpretation labels be scored at the same time?
15) Are the interpretation labels quantified?
16) Is the input processed in real time?

In general, people can recognize emotion in a neutral-content speech with an accuracy of 55%–70% when choosing from among six basic affective states [8]. Automated vocal affect analyzers match this accuracy when recognizing two to eight emotions deliberately expressed by subjects recorded while pronouncing sentences having a length of 1 to 12 words. Nonetheless, in many instances strong assumptions are made to make the problem of automating vocal expression analysis more tractable. For example, the recordings are noise free, the recorded sentences are short, delimited by pauses, and carefully pronounced by nonsmoking actors to express the required affective state. Overall, the test data sets are small (one or more words or one or more short sentences spoken by few subjects) containing exaggerated vocal expressions of affective states. Similarly to the case of automatic facial affect analysis, no readily accessible reference set of speech material exists that could provide a basis for benchmarks for efforts in automatic vocal affect analysis. In summary, the state of the art in automatic affective state recognition from speech is similar to that of speech recognition several decades ago when computers could classify the carefully pronounced digits spoken with pauses in between, but could not accurately detect these digits if they were spoken in a way not previously encountered and forming a part of a longer continuous conversation.

C. Automatic Bimodal Affect Recognition

Today there are in total four reported studies on bimodal, audiovisual interpretation of human affective feedback. These are the works of Chen *et al.* [16], De Silva and Ng [29], Yoshitomi *et al.* [122], and Chen and Huang [15].

Chen *et al.* proposed a rule-based method for singular classification of input audiovisual data into one of the following emotion categories: happiness, sadness, fear, anger, surprise, and dislike. The input data utilized by Chen *et al.* were 36 video clips of a Spanish speaker and 36 video clips of a Sinhala speaker. The speakers were asked to portray each of the six emotions 6six times using both vocal and facial expressions. Based on this data set, Chen *et al.* defined the rules for classification of acoustic and facial features into the pertinent emotion categories. From the speech signals, pitch, intensity, and pitch contours were estimated as acoustic features. Facial features such as lowering and rising of the eyebrows, opening of the eyes, stretching of the mouth, and presence of a frown, furrow, and wrinkles were manually measured from the input images. The rules for emotion classification have been evaluated only on the mentioned data set. Hence, it is not known whether or not are these rules suitable for emotion recognition from audiovisual data of an unknown subject. Besides, a quantification of the recognition results obtained by this method has not been reported. A clear picture on the actual performance of this method cannot be obtained, therefore, from the research efforts presented in [16].

De Silva and Ng also proposed a rule-based method for singular classification of input audiovisual data into one of six emotion categories used by Chen *et al.* The input data utilized by De Silva and Ng were 144 2-s-long video clips of two English speakers. Each speaker has been asked to portray 12 emotion outbursts per category by displaying the related prototypic facial expression while speaking a single English word of his choice. The pertinent audio and visual material has been processed separately. The optical flow method proposed in [3] was used to detect the displacement and its velocity of the following facial features: the mouth corners, the top and the bottom of the mouth, and the inner corners of the eyebrows. From the speech signals, pitch and pitch contours were estimated by using the method proposed in [70]. A nearest-neighbor method has been used to classify the extracted facial features, and an HMM-based method has been used to classify the estimated acoustic features into one of the emotion categories. Per subject, the results of the classifications have been plotted in a graph. Based upon the two resulting graphs, De Silva and Ng defined the rules for emotion classification of the input audiovisual material. They reported a correct recognition rate of 72% for a reduced input data set (i.e., 10% of the input samples for which the utilized rules could not yield a classification into one of the emotion categories were excluded from the data set). It is not known, therefore, whether and with which precision the method of De Silva and Ng could be used for emotion classification of similar audiovisual data obtained by recording an unknown subject.

The method proposed by Yoshitomi *et al.* represents a hybrid approach to singular classification of input audiovisual data into one of the following "basic" emotion categories: happiness, sadness, anger, surprise, and neutral. Yoshitomi *et al.* utilized 100 video clips of one female Japanese professional announcer. She was asked to pronounce a Japanese name "Taro" while portraying each of the five emotions 20 times. This input audiovisual material has been processed further as follows. From the speech signals, pitch, intensity, and pitch contours have been estimated as acoustic features. These features were classified further into one of the emotion categories by applying an HMM-based method. Yoshitomi *et al.* utilized both a visible rays (VR) camera and an infrared (IR) camera to obtain ordinary and thermal face images, respectively. From the VR and IR part of each input sample, only two VR and two corresponding IR images were utilized for further processing. The images correspond to the points where the intensity of the speech signal was maximal for the syllables "Ta" and "Ro," respectively. The typical regions of interest such the mouth region, the eyebrow and eye region, etc., were extracted from each of the selected images separately [123]. Then, each image segment has been compared to the relevant "neutral" image segment in order to generate a "differential" image. Based on the VR and the IR differential images, a discrete cosine transformation has been applied to yield a VR and an IR feature vector, respectively. An ANN-based approach has been used further to classify each of these feature vectors into one of the emotion categories. These and the classification obtained for the speech signal only were further summed to decide the final output category. Yoshitomi reported a correct recognition rate of 85% for a reduced input data set (i.e., 34% of the input samples for which the proposed method could not yield a classification into one of the emotion categories were excluded from the data set). Similarly to the works reported by Chen *et al.* and De Silva and Ng, it is not known whether and with which precision the method of Yoshitomi *et al.* could be used for emotion classification of audiovisual data from an unknown subject.

Chen and Huang proposed a set of methods for singular classification of input audiovisual data into one of the "basic" emotion categories: happiness, sadness, disgust, fear, anger, and surprise. They utilized 180 video sequences being about 70 samples long [19]. These data were from five subjects, each of which displayed six basic emotions six times by producing the appropriate facial expression right before or after speaking a sentence with the appropriate vocal emotion. Each of these single-emotion sequences started and ended with a neutral expression. For facial motion tracking, Chen and Huang utilized the Tao–Huang algorithm based upon a piecewise Bezier volume deformation model (PBVD) [111]. First a 3-D facial mesh model embedded in multiple Bezier volumes was constructed by manual selection of landmark facial feature points in the first video frame (frontal view of a neutral facial expression). Then, for each adjacent pair of frames, the 2-D motion vectors of multiple mesh nodal points were estimated using a multiresolution template matching method. To alleviate the shifting problem, the templates from both the previous and the first frame have been used. From these motion vectors, 3-D rigid head motions and 3-D nonrigid facial motions were computed using a least squares estimator. The algorithm also utilized a set of 12 predefined basic facial movements (i.e., facial muscle actions) to describe the motions around the mouth, eyes, and eyebrows. The final output of the algorithm was a vector containing the strengths of these facial muscle actions. A classifier based upon a sparse network of winnows with naive Bayes output nodes [121] has been used further to classify this vector into one of the emotion categories. From the speech signals, pitch and intensity have been computed using the ESPS *get_f0* command, and the speech rate has been found using a recursive convex-hull algorithm. These features were classified further into one of the emotion categories, each of which has been modeled with a Gaussian distribution. Given that in each of the utilized video clips a pure facial expression occurs right before or after a sentence spoken with the appropriate vocal emotion, Chen and Huang applied the single-modal methods described above in a sequential manner. They performed two types of experiments: person-dependent and person-independent experiments. In person-dependent experiments, half of the available data have been used as the training data and the other half as the test data. A 79% average recognition rate has been achieved in this experiment. In person independent experiments, data from four subjects have been used as the training data, and the data from the remaining subject have been used as the test data. A 53% average recognition rate has been reported for this experiment.

In brief, the existing works on human affect analysis from bimodal data assume, in general, clean audiovisual input (e.g., noise-free recordings, closely placed microphone, nonoccluded portraits) from an actor speaking a single word and displaying exaggerated facial expressions of "basic" emotions. Though audio and image processing techniques in these systems are relevant to the present goals, the systems themselves need many improvements if they are to be used for a multimodal context-sensitive HCI where a clean input from a known actor/announcer cannot be expected and a context-independent data interpretation does not suffice.

IV. Challenges and Opportunities

The limitations of existing systems for human affect recognition are probably the best place to start a discussion about the challenges and opportunities which researchers of affective computing face. The issue that strikes and surprises us most is that, although the recent advances in video and audio processing make *bimodal*, audiovisual analysis of human affective feedback tractable and although all agreed that solving this problem would be extremely useful, merely four efforts aimed at an actual implementation of such a bimodal human affect analyzer have been reported up to date (Section III). Further, there is no record of a research endeavor toward inclusion of all nonverbal modalities into a single system for affect-sensitive analysis of human behavior. Besides the problem of achieving a deeper integration of detached visual, auditory, and tactile research communities, there are a number of related additional issues.

A. Visual Input

The acquisition of video input for machine analysis of human affective states concerns, at least, the detection of the monitored subject's face in the observed scene. The problematic issue here, typical for all visual sensing including gaze, lip movement, and facial-gesture tracking, is that of scale, resolution, pose, and occlusion. Namely, in most real-life situations, it cannot be assumed that the subject will remain immovable; rigid head and body movements can be expected, causing changes in the viewing angle and in the visibility and illumination of the tracked facial features.

Although highly time consuming, the scale problem can be solved as proposed by Viennet and Soulie [115], i.e., by forming a multiresolution representation of the input image/frame and performing the same detection procedure for different resolutions. To do so, however, a high spatial resolution of the original input image/frame is necessary if the discrimination of subtle facial changes is to be achieved. A standard NTSC or PAL video camera provides an image that, when digitized, measures approximately 720×480 or 720×576 pixels, respectively. Since images of 100–200 pixels form a lower limit for the detection of a face and its expression by a human observer [13], it may be sufficient for a typical detection/tracking of facial expressions to arrange the camera setting so that there are at least 200 pixels across the width of the subject's face. The field of view can then be about two and one-half times the width of the face. This camera setting should be sufficient for machine recognition of human facial affect, in which the subject is seated but otherwise free to move his head. On the other hand, such a camera setting may not be sufficient for the pertinent task if the subject is free to walk in front of the camera and to approach and move away from the camera. Hence, it would be highly beneficial to develop strategies for extending the field of regard while maintaining a high resolution. The investigation and development of such strategies is the subject of research in the field of active vision [1]. In general terms, the objective of active camera control is to focus sensing resources on relatively small regions of the scene that contain critical information. In other words, the aim of active vision systems is to observe the scene with a wide field of view at a low spatial resolution and then to determine where to direct high spatial resolution observations. This is analogous to human vision in the fovea—a small depression in the retina where vision is most acute. The fovea provides the resolution needed for discriminating patterns of interest, while the periphery provides broad-area monitoring for altering and gaze control. Although the work in the field of active vision has not been extended yet for the purposes of face detection and tracking, it is certain that the field of automatic facial affect analysis and affective computing in general could highly benefit from the progress in this area of research.

Pose and occlusion are even more difficult problems, initially thought to be intractable or at least the hardest to solve [88]. Yet significant progress is being made using methods for the monitored object's representation at several orientations, employing data acquired by multiple cameras. Those methods are currently thought to provide the most promising solution to the problems of pose and occlusion [88]. For an extensive review of such methods utilized for video surveillance, the reader is referred to [20]. In addition, interesting progress is being made using statistical methods, which essentially try to predict the appearance of monitored objects from whatever image information is available. As explained in Section II, a statistical facial expression predictor based upon a task- and user-profiled temporal grammar of human facial behavior could prove most serviceable for the purposes of handling partial data within machine human affect analysis from images of faces.

Besides these standard visual-processing problems, there is another cumbersome issue typical for face image processing: the universality of the employed technique for detection of the face and its features. Namely, the employed detection method must not be prone to the physiognomic variability and the current looks of monitored subjects. As explained in Section II, an "ideal" automated system for facial affect recognition should perform a generic analysis of the sensed facial information, independently of possibly present static facial signals like birthmarks and facial hair, slow facial signals such as wrinkles, and artificial facial signals like glasses and makeup. Essa and Pentland [39] proposed such a method.

B. Audio Input

As mentioned previously, virtually all of the work done on automating vocal affect analysis assumes a fixed listening position, a closely placed microphone, nonsmoking actors or radio announcers, and noise-free recordings of short, neutral-content sentences that are delimited by pauses and carefully pronounced to express the required affective state. Such a clean audio input is not realistic to expect, however, especially not in unconstrained environments in which human affect analyzers are most likely to be deployed (multimodal context-sensitive HCI and ubiquitous computing). An obvious, necessary extension of the current research on automatic vocal affect analysis is to extend the range of test material to include speech samples that:

1) are naturally spoken rather than read by actors;
2) have meaningful rather than semantically neutral content (e.g., "You came late again," rather than "What time is it?");
3) include rather than exclude continuous speech;
4) are drawn from a genuine range of speakers, in terms of sex, age, smoking pattern, and social background;
5) use a range of languages.

However, the extension of the range of test material will not provide a solution to the original problem which incited, in the first place, the acquisition of the currently used, heavily constrained speech material. Processing and interpreting unconstrained audio input with minimal degradation in performance remains a significant research challenge facing both the speech understanding research field [54], [88] and the affective computing research field [24], [90]. As far as the area of affective computing is concerned, a basic problem that needs to be solved is that of defining a better computational mapping between affective states and speech patterns. Specifically, it is necessary to find features which can be extracted by a computer and which, at the same time, may be used for discerning affective states from less constrained speech material.

The psycholinguistic research results are consistent in general on some neutral-speech continuous acoustic correlates of a few "basic" emotions like anger, fear, happiness, and sadness (see Table 2). Yet, even within the research on this small set of prototypic emotions, there are many contradictory reports. For example, there is a disagreement on duration facets of anger, fear, and happiness—some researchers report a faster speech rate and some report a slower speech rate (see Table 2). Furthermore, while some researchers adhere to the standpoint that the features should be solely acoustic and different from the phonetic features used for speech understanding, others adhere to the standpoint that acoustic and phonetic features are tightly combined when uttering speech. The later standpoint accedes to the assumption that it is impossible for us to express and recognize vocal expressions of affective states by considering only acoustic features. There are a number of fundamental reasons to adhere to this standpoint. The best known example that the features which signal affective arousal are easy to confound with the features which are determined by linguistic rules involves questions. Uttering a question gives rise to distinctive pitch contours that could easily be taken as evidence of affective inflection if linguistic context is ignored. Similar examples are turn taking and topic introduction [69]. Another reason for considering linguistic content in connection with the detection of affective arousal concerns the dependence of our performance from whether we understood the linguistic content of not. As shown by De Silva *et al.* [28], observers who did not speak the Sinhala language recognized six different emotions correctly in Sinhala spoken speech with an average of only 32.3%. Research efforts toward the inclusion of phonetic features in vocal affect analysis have been reported in [50], [74], and [113]. Another interesting observation is that the information encoded in the speech signal becomes far more meaningful if the pitch and intensity can be observed over the duration of a syllable, word, or phrase [51], [92]. Although a robust detection of boundaries at different levels still poses a significant research challenge in speech processing (that is why recognition of connected speech lingers far behind recognition of discrete words), there is a large body of literature that could be used as a source of help while tackling the problems of locating pauses between phrases and detecting boundaries between words and phonemes [54].

In summary, establishing reliable knowledge about a set of suitable features sufficient for discriminating any affective state (archetypal or nonarchetypal) from unconstrained audio input still lags in the distant future. In order to approach this goal, research toward temporal analysis of both acoustic and phonetic characteristics of spontaneous affective speech is needed. Finally, the large gaps in "related research" headings of the works surveyed in Table 3 indicate a detachment of the existing auditory affect research communities. To raise the overall quality of research in this field, a deeper integration of currently detached research communities is necessary.

C. Multimodal Input

An ideal analyzer of human nonverbal affective feedback (see Fig. 1) should generate a reliable result based on multiple input signals acquired by different sensors. There are a number of related issues, which pose interesting but significant research challenges.

If we consider the state of the art in audio, visual, and tactile processing, noisy and partial input data should be expected. An (ideal) analyzer of human nonverbal affective feedback should be able to deal with these imperfect data and generate its conclusion so that the certainty associated with it varies in accordance with the input data. A way of achieving this is to consider the time-instance versus time-scale dimension of human nonverbal communicative signals. As already explained in Section II, there is a certain grammar of neuromuscular actions and physiological reactions. By considering previously observed actions/reactions (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. Generative probability models such as HMM provide a principled way of handling temporal and treating missing information. Another alternative is to exploit discriminative methods such as support vector machines or kernel methods [114], whose classification performance is often superior to that of generative classifiers, in a combination with generative probability models for extracting features for use in discriminative classification [52]. Yet such a temporal analysis also involves untangling the grammar of human behavior, which still represents a rather unexplored topic even in the psychological and sociological research areas. The issue that makes this problem even more difficult to solve in a general case is the dependency of a person's behavior on his/her personality, cultural and social vicinity, current mood, and the context in which the observed behavioral cues were 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 application-, user-, and context-dependent rules by watching the user's behavior in the sensed context [88]. Though context sensing and the time needed to learn appropriate rules are significant problems in their own right, many benefits could accrue from such an adaptive affect-sensitive HCI tool (Section I).

Another typical issue of multimodal data processing is that the multisensory data are processed separately and only combined at the end (Section III). Yet this is almost certainly incorrect; people display audio, visual, and tactile communicative signals in a complementary and redundant manner. Chen *et al.* [16] have proven this experimentally for the case of audio and visual input. 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 (Section II). The input data should be processed in a joint feature space and according to a context-dependent model. In practice, however, except the acute problems of context sensing and developing context-dependent models for combining multisensory information, there are two additional major difficulties: the size of the required joint feature space, which is usually huge and results in a heavy computational burden, and different feature formats and timing. A potential way to achieve the target tightly coupled multisensory data fusion is to develop context-dependent versions of a suitable method such as the Bayesian inference method proposed in [83].

D. Affect-Sensitive Interpretation of Multimodal Input

As explained in Section II, accomplishment of a human-like interpretation of sensed human affective feedback requires pragmatic choices (i.e., application-, user- and task-profiled choices). Nonetheless, as already noted, currently existing methods aimed at the automation of human affect analysis are not context sensitive (see also Tables 1 and 3). Initially thought to be the research topic that would be the hardest to solve, context sensing—that is, answering questions such as who the user is, where he is, and what he is doing—has witnessed recently a number of significant advances [20], [88]. However, the complexity of this wide-ranging problem makes the problem of context-sensitive human affect analysis perhaps the most significant of the research challenges facing researchers of affective computing.

Another issue concerns the actual interpretation of human nonverbal interactive signals in terms of affective/attitudinal states. The existing work employs usually singular classification of input data into one of the "basic" emotion categories (Section III). This approach has many limitations. As mentioned previously, the theory on the existence of universal emotion categories is nowadays strongly challenged in the psychological research area (Section II). Further, pure expressions of "basic" emotions are seldom elicited; most of the time, people show blends of emotional displays. Hence, the classification of human nonverbal affective feedback into a single basic-emotion category is not realistic. Automatic analyzers of sensed nonverbal affective cues must at least realize quantified classification into multiple-emotion categories, as proposed, for example, in [32], [53], [84], and [124] for automatic facial affect analysis and in [2] and [74] for automatic vocal affect analysis. Yet not all nonverbal affective cues can be classified as a combination of the "basic" emotion categories. Think, for instance, about the frustration, stress, skepticism, or boredom attitudinal states. Also, 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 (Section II). Hence, the definition of interpretation categories in which any set of displayed human nonverbal affective cues can be classified is a key challenge in the design of realistic affect-sensitive monitoring tools. The lack of psychological scrutiny on the topic makes this problem even harder. One source of help is (again) machine learning: instead of integrating rigid generic rules for the interpretation of human nonverbal affective feedback into the intended tool, the system can potentially learn its own expertise by allowing the user to define his own (context-dependent) interpretation categories (e.g., as proposed in [57] and [87]).

E. Validation Issues

In general, validation studies on an automated system address the question of whether the developed system does what it should do while complying with the predefined set of requirements. Automated analyzers of human affective feedback are usually envisioned as machine tools for sensing, analyzing, and translating human communicative cues into a description of the expressed affective state. This description should be identical, or at least very close, to a human's description of the pertinent affective state. Hence, validation studies on automated analyzers of human affective feedback address commonly the question of whether the interpretations reached automatically are equal to those given by human observers judging the same stimulus material. In turn, evaluating the performance of an automated human affect analyzer involves obtaining a set of test material coded by human observers for ground truth (i.e., in terms of affective states shown in the pertinent material). In order to enable comparing, resolving, and extending the issues concerned with automatic human affective feedback analysis, this set of test material should be a standard, commonly used database of test material.

Nevertheless, no readily accessible database of test material that could be used as a basis for benchmarks for efforts in the research area of automated human affective feedback analysis has been established yet. In fact, even in the research on facial affect analysis, which attracted the interest of many researchers and became one of the hot topics in machine vision and artificial intelligence (AI) research, there is a glaring lack of an existing benchmark facial database. This lack of common testing resources forms the major impediment to comparing, resolving, and extending the issues concerned with automatic human affective feedback analysis and understanding. It slowed down not only the progress in applying computers to analyze human facial and/or vocal affective feedback but also overall cooperation and collaboration among investigators of affective computing. The benefits that could accrue from a commonly used audiovisual database of human affect expressions are numerous.

1) Avoiding redundant collection of facial and/or vocal expression exemplars can reduce research costs: investigators can use one another's exemplars.
2) Having a centralized repository for retrieval and exchange of audio and/or visual training and test material can improve research efficiency.
3) Maintaining various test results obtained for a reference audio and/or visual data set and, hence, providing a basis for benchmarks for research efforts can increase research quality. This would also reduce the currently existing abundance of reports presenting rather insignificant achievements.

Thus, the establishment of a readily accessible, benchmark audiovisual database of human affect expressions has become an acute problem that needs to be resolved if fruitful avenues for new research in affective computing are to be opened. A number of issues make this problem complex and, in turn, rather difficult to tackle.

Which objects should be included into such an audiovisual database so that it meets multiple needs of scientists working in the field? Facial expression is an important variable for a large number of studies in computer science (lipreading, audiovisual speech and face synthesis, etc.; see Section I) as well as for research in behavioral science, psychology, psychophysiology, anthropology, neurology, and psychiatry [37]. While motion records are necessary for studying temporal dynamics of facial behavior, static images are important for obtaining configurational information about facial expressions [37]. Hence, the benchmark database should include both still and motion images of faces. For their relevance in evaluating the achievements in tackling past and present challenges in automatic human affect analysis, images of nonoccluded and partially occluded faces in various poses acquired under various lighting conditions should be included. In order to support both the research efforts directed toward specific monodisciplinary research goals and the integration of research efforts done in various fields including lipreading, facial, vocal, and audiovisual human affect recognition, the database should include all video recordings of silent subjects, audio recordings of speech and other vocalizations, and audiovisual recordings of speaking subjects. Another important variable is the distinction between deliberate actions performed on request versus spontaneous actions not under volitional control. Examples of both categories should be included in the database in order to study the essential question of the difference between these expressions. Examples of prototypic emotional expressions defined in classic studies on emotion (e.g., [5], [8], [58]) should be made available. To facilitate experiments directed toward alternative approaches to human affect expression interpretation (e.g., application- and/or user-profiled interpretation), expressions in which only some components of prototypic expressions are present should be also included in the database.

A crucial aspect of the benchmark database of human affect expressions is the metadata that is to be associated with each database object and to be used as the ground truth in validating automated human affect analyzers. For general relevance, the images should be scored in terms of facial muscle actions such as the facial AUs defined in Ekman and Friesen's FACS system [36] and used as a standard measure of activity in the face [37]. The interpretations of displayed (facial and/or vocal) expressions in terms of affective state(s) should be associated with each database object. For spontaneous expressions, the associated metadata should also identify the conditions under which the expression has been elicited. This provision is important, since the eliciting circumstances can produce different types of expressions (e.g., a conversation versus listening and/or watching stimulus material alone).

How should objects and the related metadata be collected for inclusion into the benchmark audiovisual database? First, several technical considerations for the database should be resolved including criteria for sensor data rates (i.e., field of sensing, spatial resolution, and frame rate), data formats, and compression methods. The choice should enable sharing the database between different research communities all over the world. Further, the database objects should represent a number of demographic variables including ethnic background, gender, and age, and should provide a basis for generality of research findings. For each category of database objects, audio and/or visual recordings of several individuals should be included in order to avoid effects of the unique properties of particular people. For the acquisition of deliberate actions performed on request, the recorded subjects should be either experts in production of facial and/or vocal expressions (e.g., individuals having a formal training in using FACS, behavioral scientists) or individuals being instructed by such experts on how to perform the required expressions. For spontaneous expressions, the decision should be made about how to acquire this kind of data. The problematic issue here is that hidden recordings, which promise the acquisition of truly spontaneous behavior of people in their usual environment, are privacy intruding and unethical. Given the large number of expressions that should be included into the database, provision should be made for individual researchers to add their own research material to the database. However, a secure handling of such additions has to be facilitated. At least, an automatic control of whether an addition matches the specified technical and other formats defined for the database objects should be realized.

A number of issues should be taken into account when generating the metadata to be associated with the database objects. Humans detect six basic emotional expressions in face images and/or neutral-content speech with an accuracy ranging from 55% to 98% [7], [8]. Thus, human observers may sometimes disagree in their judgments, and they may make mistakes occasionally. Since validation studies on automated analyzers of human affective feedback address commonly the question of whether the interpretations reached automatically are equal to those given by human observers judging the same stimulus material, the validation of an automated system is only as sound as the validity of the metadata associated with test samples. It is necessary, therefore, that the consensus of the experts involved in generating metadata to be associated with the database objects is excessive. Yet this is by no means an easily achievable goal. First, it is extremely difficult to ascertain whether the human experts involved in generating metadata are sufficiently concentrated on their task. Second, the facial and vocal affect analyses are perhaps tasks that always yield inconsistencies, even between the judgments made by a single human observer in two successive analyses of the same stimulus material. A source of help for these problems would be the usage of more accurate means for recognizing facial and vocal expressions such as measures of muscular electrical activity to double-check human visual and auditory judgments of stimulus material. However, this involves wiring the subjects, and that, in turn, results in visual occlusions of recorded facial expressions. Perhaps the only available means for addressing these problems is, therefore, the involvement of many human experts in the task of generating metadata to be associated with the database objects. This, however, implies occupying valuable time of many trained human observers for a longer period, and that, in turn, might trigger unwillingness of behavioral science research communities to participate in the process of establishing the subject audiovisual benchmark database.

How does one construct and administer such a large audiovisual database? How does one facilitate efficient, fast, and secure retrieval and inclusion of objects constituting this database? The benchmark audiovisual database envisioned in this section could be valuable to hundreds or even thousands of researchers in various scientific areas if it would be easy to access and use. A relaxed level of security, which allows any user a quick access to the database and frees database administrators of time-consuming identity checks, can attain such an easy access. Yet, in this case, nonscientists such as journalists and hackers would be able to access the database. If the database is likely to contain images that can be made available only to certain authorized users (e.g., images of psychiatric patients with emotional disorders), then a more comprehensive security strategy should be used. For example, a mandatory multilevel access control model could be used in which the users can get rights to use database objects at various security levels (e.g., confidential, for internal use only, no security, etc.) [37]. However, the usage of any such access control model implies that the database would neither be easily accessible nor be easily usable. Hence, perhaps the best strategy would be to encourage primary researchers to include into the database just the recordings (imagery and/or speech samples) without restriction on use and then to allow a relaxed level of access control as just described. Another important issue is the problem of secure inclusion of objects into the database. For this purpose, procedures for determining whether a recording to be added matches the specified technical and other formats defined for the database objects need to be developed. Other important issues that should be resolved involve the following questions: 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 pertinent evaluation be defined? How could fast and reliable object distribution over networks be achieved?

In summary, the development of a readily accessible, benchmark audiovisual database, which meets multiple needs of scientists working in the field, involves many questions that need to be answered. We believe that these questions could appropriately be answered only if an interdisciplinary team of computer vision experts, spoken language processing experts, database designers, and behavioral scientists is set to investigate the pertinent aspects.

V. Conclusion

As remarked by scientists like Nass [75], [93] and Pentland [88], multimodal context-sensitive (user-, task-, and application-profiled and affect-sensitive) HCI is likely to become the single most widespread research topic of the AI research community. Breakthroughs in such HCI designs could bring about the most radical change in 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, are not mature yet and need many improvements. Current approaches to computer-based analysis of human affective feedback are as follows (see Tables 1 and 3):

1) single modal (except for the audiovisual analyzers discussed in Section III-C)—information processed by the computer system is limited either to acquired face images or recorded speech signal; and
2) context insensitive—no attention is paid to who the current user is, where he works, and what his current task is.

In summary, although the fields of machine vision, audio processing, and affective computing witnessed rather significant advances in the past few years, the realization of a robust, multimodal, adaptive, context-sensitive analyzer of human nonverbal affective feedback still lies in a rather distant future. Except the problems involved in the integration of multiple sensors and pertinent modalities according to the model of the human sensory system and the lack of a better understanding of individual- and context-dependent human behavior in general, there are two additional related issues.

The first issue that could jeopardize a future wide deployment of this new HCI technology concerns the efficiency of the pertinent HCI tools. Namely, since it is generally thought that pervasive computing devices will be all around us in the future [99], it will be inefficient if the user should train each of those devices separately. The computers of our future must know enough about the people and the environment in which they act to be capable of adapting to the current user with a minimum of explicit instruction [88]. A long-term way of achieving this is the following.

1) Develop multimodal affect-sensitive HCI tools conforming to the recommendations provided in this paper (Section IV), which will be able to monitor human nonverbal behavior and to adapt to the current user (i.e., to who he is and to what the grammar of his behavioral actions/reactions is), to his context (i.e., to where he is and to what he is doing at this point), and to the current scenario (e.g., stress sensed by a nuclear power plant operator while he reads his e-mail is not a cause for an alarm).
2) Make those adaptive tools commercially available to the users who will profile them in the context in which the tools are to be used.
3) Withdraw the trained systems after some time and combine the stored knowledge in order to derive generic statistical rules/models for interpretation of human nonverbal behavior in the given context/environment.

Although the unwillingness of people to participate in such a privacy-intruding large-scale project is a significant problem in its own right, this approach could resolve many intriguing questions. The most important is that this could resolve the social impact of interaction in electronic media, that is, the effects of computing and information technology on our interaction patterns and related behavior and on our social and cultural profile.

Another issue that might jeopardize a future wide deployment of the pertinent HCI technology concerns the design of the HCI tools in question. Computer technology and especially affect-sensitive monitoring tools might be perceived as "big brother is watching you" tools. As remarked by Schneiderman [103], a large proportion of the population would in fact be terrified by the vision of the universal use of computers in the coming era of ubiquitous computing. Therefore, the actual deployment of context-sensitive multimodal HCI tools proposed in this paper will only be attainable if the design of those tools *will not*:

1) invade the user's privacy (the pertinent HCI tools' capacity to monitor and concentrate information about somebody's behavior must not be misused);
2) cause the user to worry about being unemployed (air-traffic or production controllers do not want computer programs that could cause their layoff but help them in performing their job faster and/or more accurately);
3) reduce the user's professional responsibility (insisting on the "intelligent" capabilities of computing devices could have negative effects like blaming machines for our own poor performance or seeing machines as infallible devices instead of tools that can merely empower us by retrieving and processing information faster and according to our own preferences and the context in which we act).

Multimodal HCI tools envisioned in this paper could enable the development of smart, perceptually aware environments that can adapt to their users, recognize the context in which they act, understand how they feel, and respond appropriately. They might represent the coming of human-like (natural) HCI and the means for determining the impact the information technology has on our social behavior. Yet we should recognize that the realization of this human-centered HCI technology still lies in a relatively distant future and that its commercialization and actual deployment depends upon its user-friendliness and trustworthiness.

Acknowledgment

The authors would like to thank Prof. T. Huang, Prof. H. Koppelaar, and anonymous reviewers for their helpful comments and suggestions.

References

[1] J. Y. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 333–356, 1988.

[2] N. Amir and S. Ron, “Toward automatic classification of emotions in speech,” in *Proc. ICSLP*, 1998, pp. 555–558.

[3] P. Anandan, “A computational framework and an algorithm for the measurement of visual motion,” *Int. J. Comput. Vis.*, vol. 2, no. 3, pp. 283–310, 1989.

[4] J. R. Averill, “Acquisition of emotions in adulthood,” in *The Social Construction of Emotions*, R. Harre, Ed. Oxford, U.K.: Blackwell, 1986, p. 100.

[5] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression,” *J. Personality Social Psychol.*, vol. 70, no. 3, pp. 614–636, 1996.

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

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

[8] R. V. Bezooijen, *Characteristics and Recognizability of Vocal Expression of Emotions*. Dordrecht, The Netherlands: Floris, 1984.

[9] M. J. Black and Y. Yacoob, “Recognizing facial expressions in image sequences using local parameterized models of image motion,” *Int. J. Comput. Vis.*, vol. 25, no. 1, pp. 23–48, 1997.

[10] E. Boyle, A. H. Anderson, and A. Newlands, “The effects of visibility on dialogue in a cooperative problem solving task,” *Lang. Speech*, vol. 37, no. 1, pp. 1–20, 1994.

[11] V. Bruce, “What the human face tells the human mind: Some challenges for the robot-human interface,” in *Proc. ROMAN*, 1992, pp. 44–51.

[12] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, and T. A. Ito, “The psychophysiology of emotion,” in *Handbook of Emotions*, M. Lewis and J. M. Havil-Jones, Eds. New York: Guilford, 2000, pp. 173–191.

[13] F. W. Campbell and D. G. Green, “Optical and retinal factors affecting visual resolution,” *J. Physiol.*, vol. 181, pp. 576–593, 1965.

[14] J. Cassell and T. Bickmore, “External manifestations of trust-worthiness in the interface,” *Commun. ACM*, vol. 43, no. 12, pp. 50–56, 2000.

[15] L. S. Chen and T. S. Huang, “Emotional expressions in audiovisual human computer interaction,” in *Proc. ICME*, 2000, pp. 423–426.

[16] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, “Multimodal human emotion/expression recognition,” in *Proc. FG*, 1998, pp. 396–401.

[17] T. Chen and R. R. Rao, “Audio-visual integration in multimodal communication,” *Proc. IEEE*, vol. 86, pp. 837–852, May 1998.

[18] T. Chen, “Audiovisual speech processing,” *IEEE Signal Processing Mag.*, vol. 18, pp. 9–21, Jan. 2001.

[19] I. Cohen, N. Sebe, A. Garg, M. S. Lew, and T. S. Huang, “Facial expression recognition from video sequences,” in *Proc. ICME*, 2002, pp. 121–124.

[20] “Special section on video surveillance,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 745–887, Aug. 2000.

[21] A. Colmenarez, B. Frey, and T. S. Huang, “Embedded face and facial expression recognition,” in *Proc. ICIP*, vol. 1, 1999, pp. 633–637.

[22] R. Cornelius, *The Science of Emotion*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[23] G. W. Cottrell and J. Metcalfe, “EMPATH: Face, emotion, and gender recognition using holons,” *Adv. Neural Inf. Process. Syst.*, vol. 3, pp. 564–571, 1991.

[24] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Mag.*, vol. 18, pp. 32–80, Jan. 2001.

[25] C. Darwin, *The Expression of the Emotions in Man and Animals*. Chicago, IL: Univ. of Chicago Press, 1965.

[26] B. V. Dasarathy, “Sensor fusion potential exploitation—innovative architectures and illustrative approaches,” *Proc. IEEE*, vol. 85, pp. 24–38, Jan. 1997.

[27] J. R. Davitz, Ed., *The Communication of Emotional Meaning*. New York: McGraw-Hill, 1964.

[28] L. C. De Silva, T. Miyasato, and R. Nakatsu, “Facial emotion recognition using multimodal information,” in *Proc. ICSP*, 1997, pp. 397–401.

[29] L. C. De Silva and P. C. Ng, “Bimodal emotion recognition,” in *Proc. FG*, 2000, pp. 332–335.

[30] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” in *Proc. ICSLP*, 1996, pp. 1970–1973.

[31] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Classifying facial actions,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 974–989, Oct. 1999.

[32] H. Ebine, Y. Shiga, M. Ikeda, and O. Nakamura, “The recognition of facial expressions with automatic detection of reference face,” in *Proc. Canadian Conf. ECE*, vol. 2, 2000, pp. 1091–1099.

[33] G. J. Edwards, T. F. Cootes, and C. J. Taylor, “Face recognition using active appearance models,” in *Proc. ECCV*, vol. 2, 1998, pp. 581–695.

[34] P. Ekman, “Universals and cultural differences in facial expressions of emotion,” in *Proc. Nebraska Symp. Motivation*, J. Cole, Ed., 1972, pp. 207–283.

[35] P. Ekman and W. F. Friesen, “The repertoire of nonverbal behavioral categories—origins, usage, and coding,” *Semiotica*, vol. 1, pp. 49–98, 1969.

[36] P. Ekman and W. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologist, 1978.

[37] P. Ekman, T. S. Huang, T. J. Sejnowski, and J. C. Hager, Eds., “Final report to NSF of the planning workshop on facial expression understanding,” Human Interaction Lab., Univ. California, San Francisco, 1993.

[38] P. Ekman, J. Hager, C. H. Methvin, and W. Irwin, “Ekman-Hager facial action exemplars,” Human Interaction Lab., Univ. California, San Francisco.

[39] I. Essa and A. Pentland, “Coding analysis interpretation recognition of facial expressions,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 757–763, July 1997.

[40] W. A. Fellenz, J. G. Taylor, R. Cowie, E. Douglas-Cowie, F. Piat, S. Kollias, C. Orovas, and B. Apolloni, “On emotion recognition of faces and of speech using neural networks, fuzzy logic and the assess system,” in *Proc. IJCNN*, vol. 2 , 2000, pp. 93–98.

[41] I. Fonagy and K. Magdics, “Emotional patterns in intonation and music,” *Phonet. Sprachwiss. Kommunikationsforsch*, vol. 16, pp. 293–326, 1963.

[42] R. Frick, “Communicating emotion. The role of prosodic features,” *Psychol. Bull.*, vol. 97, no. 3, pp. 412–429, 1985.

[43] G. Furnas, T. Landauer, L. Gomes, and S. Dumais, “The vocabulary problem in human-system communication,” *Commun. ACM*, vol. 30, no. 11, pp. 964–972, 1987.

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

[45] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” *Proc. IEEE*, vol. 85, pp. 6–23, Jan. 1997.

[46] A. Hanjalic and L.-Q. Xu, “User-oriented affective video content analysis,” in *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries* , 2001, pp. 50–57.

[47] F. Hara and H. Kobayashi, “Facial interaction between animated 3D face robot and human beings,” in *Proc. SMC*, 1997, pp. 3732–3737.

[48] H. Hong, H. Neven, and C. von der Malsburg, “Online facial expression recognition based on personalised galleries,” in *Proc. FG*, 1998, pp. 354–359.

[49] C. L. Huang and Y. M. Huang, “Facial expression recognition using model-based feature extraction and action parameters classification,” *Vis. Commun. Image Representat.*, vol. 8, no. 3, pp. 278–290, 1997.

[50] R. Huber, A. Batliner, J. Buckow, E. Noth, V. Warnke, and H. Niemann, “Recognition of emotion in a realistic dialogue scenario,” in *Proc. ICSLP*, 2000, pp. 665–668.

[51] G. Izzo. (1998) Review of Existing techniques for human emotion understanding and applications in HCI. [Online]. Available: http://www.image.ece.ntua.gr/physta/reports/emotionreview.

[52] T. S. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” in *Proc. Advances in Neural Information Processing*, vol. 11, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds., 1998, pp. 487–493.

[53] S. Ji, C. Yoon, J. Park, and M. Park, “Intelligent system for automatic adjustment of 3D facial shape model and face expression recognition,” in *Proc. IFSC*, vol. 3, 1999, pp. 1579–1584.

[54] “Special issue on spoken language processing,” *Proc. IEEE*, vol. 88, pp. 1139–1366, Aug. 2000.

[55] T. Kanade, J. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Proc. FG*, 2000, pp. 46–53.

[56] B. S. Kang, C. H. Han, S. T. Lee, D. H. Youn, and C. Lee, “Speaker dependent emotion recognition using speech signals,” in *Proc. ICSLP*, 2000, pp. 383–386.

[57] G. D. Kearney and S. McKenzie, "Machine interpretation of emotion: Memory-based expert system for interpreting facial expressions in terms of signaled emotions (JANUS)," *Cogn. Sci.*, vol. 17, no. 4, pp. 589–622, 1993.

[58] D. Keltner and P. Ekman, "Facial expression of emotion," in *Handbook of Emotions*, M. Lewis and J. M. Havil-Jones, Eds. New York: Guilford, 2000, pp. 236–249.

[59] S. Kimura and M. Yachida, "Facial expression recognition and its degree estimation," in *Proc. CVPR*, 1997, pp. 295–300.

[60] H. Kobayashi and F. Hara, "Recognition of mixed facial expressions by a neural network," in *Proc. ROMAN*, 1992, pp. 387–391 (see also pp. 381-386).

[61] T. Kurozumi, Y. Shinza, Y. Kenmochi, and K. Kotani, "Facial individuality and expression analysis by eigenface method based on class features or multiple discriminant analysis," in *Proc. ICIP*, vol. 1, 1999, pp. 648–652.

[62] Y. Li and Y. Zhao, "Recognizing emotions in speech using short-term and long-term features," in *Proc. ICSLP*, 1998, pp. 2255–2258.

[63] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1357–1362, Dec. 1999.

[64] S. Mann, "Wearable computing: A first step toward personal imaging," *Computer*, vol. 30, no. 2, pp. 25–32, 1997.

[65] I. Marsic, A. Medl, and J. Flanagan, "Natural communication with information systems," *Proc. IEEE*, vol. 88, pp. 1354–1366, Aug. 2000.

[66] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E74, no. 10, pp. 3474–3483, 1991.

[67] D. Matsumoto, "Cultural similarities and differences in display rules," *Motivat. Emotion*, vol. 14, pp. 195–214, 1990.

[68] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: Univ. of Chicago Press, 1992.

[69] G. McRoberts, M. Studdert-Kennedy, and D. P. Shankweiler, "Role of fundamental frequency in signaling linguistic stress and affect: Evidence for a dissociation," *Percept. Psychophys.*, vol. 57, pp. 159–174, 1995.

[70] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Signal Processing*, vol. 39, pp. 40–48, Jan. 1991.

[71] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 4, pp. 53–56, 1968.

[72] I. R. Murray and J. L. Arnott, "Synthesizing emotions in speech: Is it time to get excited?," in *Proc. ICSLP*, 1996, pp. 1816–1819.

[73] R. Nakatsu, "Toward the creation of a new medium for the multimedia era," *Proc. IEEE*, vol. 86, pp. 825–836, May 1998.

[74] R. Nakatsu, J. Nicholson, and N. Tosa, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities," *Knowl.-Based Syst.*, vol. 13, no. 7–8, pp. 497–504, 2000.

[75] C. I. Nass, J. S. Steuer, and E. Tauber, "Computers are social actors," in *Proc. CHI*, 1994, pp. 72–78.

[76] T. L. Nwe, F. S. Wei, and L. C. De Silva, "Speech-based emotion classification," in *Proc. TENCON*, vol. 1, 2001, pp. 297–301.

[77] I. Oakley, M. R. McGee, S. Brewster, and P. Grey, "Putting the feel in 'look and feel'," in *Proc. CHI* , 2000, pp. 415–422.

[78] J. S. Olson and G. M. Olson, "Trust in e-commerce," *Commun. ACM*, vol. 43, no. 12, pp. 41–44, 2000.

[79] A. Ortony and T. J. Turner, "What is basic about emotions?," *Psychol. Rev.*, vol. 74, pp. 315–341, 1990.

[80] T. Otsuka and J. Ohya, "Spotting segments displaying facial expression from image sequences using HMM," in *Proc. FG*, 1998, pp. 442–447.

[81] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration-synchronization of input modes during multimodal human-computer interaction," in *Proc. CHI*, 1997, pp. 415–422.

[82] C. Padgett and G. W. Cottrell, "Representing face images for emotion classification," in *Proc. Advances in Neural Information Processing Systems*, 1996, pp. 894–900.

[83] H. Pan, Z. P. Liang, T. J. Anastasio, and T. S. Huang, "Exploiting the dependencies in information fusion," *Proc. CVPR*, vol. 2, pp. 407–412, 1999.

[84] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expression," *Image Vis. Comput. J.*, vol. 18, no. 11, pp. 881–905, 2000.

[85] ——, "Automatic analysis of facial expression: The state of the art," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1424–1445, Dec. 2000.

[86] ——, "Affect-sensitive multi-modal monitoring in ubiquitous computing: Advances and challenges," in *Proc. ICEIS*, 2001, pp. 466–474.

[87] M. Pantic, "Facial expression analysis by computational intelligence techniques," Ph.D. dissertation, Delft Univ. Technol., Delft, The Netherlands, 2001.

[88] A. Pentland, "Looking at people: Sensing for ubiquitous and wearable computing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 107–119, Jan. 2000.

[89] V. A. Petrushin, "Emotion recognition in speech signal," in *Proc. ICSLP*, 2000, pp. 222–225.

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

[91] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 1175–1191, Oct. 2001.

[92] T. S. Polzin, "Detecting verbal and non-verbal cues in the communications of emotions," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, 2000.

[93] B. Reeves and C. I. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge Univ. Press, 1996.

[94] J. A. Russell, "Is there universal recognition of emotion from facial expression?," *Psychol. Bull.*, vol. 115, no. 1, pp. 102–141, 1994.

[95] J. A. Russell and J. M. Fernez-Dols, Eds., *The Psychology of Facial Expression*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

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

[97] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognit.*, vol. 25, no. 1, pp. 65–77, 1992.

[98] H. Sato, Y. Mitsukura, M. Fukumi, and N. Akamatsu, "Emotional speech classification with prosodic parameters by using neural networks," in *Proc. Australian and New Zealand Intelligent Information Systems Conf.*, 2001, pp. 395–398.

[99] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Pers. Commun.*, vol. 8, pp. 10–17, Aug. 2001.

[100] P. Scanlon and R. B. Reilly, "Feature analysis for automatic speech reading," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, 2001, pp. 625–630.

[101] D. J. Schiano, S. M. Ehrlich, K. Rahardja, and K. Sheridan, "Face to interface: Facial affect in human and machine," in *Proc. CHI*, 2000, pp. 193–200.

[102] K. Scherer, T. Johnstone, and T. Banziger, "Automatic verification of emotionally stressed speakers: The problem of individual differences," in *Proc. SPECOM*, 1998, pp. 233–238.

[103] B. Schneiderman, "Human values and the future of technology: A declaration of responsibility," in *Sparks of Innovation in Human-Computer Interaction*, B. Schneiderman, Ed. Norwood, NJ: Ablex, 1993.

[104] N. Sebe, M. S. Lew, I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition using a cauchy naive bayes classifier," in *Proc. ICPR*, vol. 1, 2002, pp. 17–20.

[105] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, pp. 853–869, May 1998.

[106] S. Shigeno, "Cultural similarities and differences in recognition of audio-visual speech stimuli," in *Proc. ICSLP*, 1998, pp. 281–284.

[107] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. ICASSP*, vol. 4, 1999, pp. 2079–2082.

[108] B. Stein and M. A. Meredith, *The Merging of Senses*. Cambridge, MA: MIT Press, 1993.

[109] J. Sulc, "Emotional changes in human voice," *Activitas Nervosa Superior*, vol. 19, pp. 215–216, 1977.

[110] A. Takeuchi and K. Nagao, "Communicative facial displays as a new conversational modality," in *Proc. INTERCHI*, 1993, pp. 187–193.

[111] H. Tao and T. S. Huang, "Explanation-based facial motion tracking using a piecewise bezier volume deformation model," in *Proc. CVPR*, vol. 1, 1999, pp. 611–617.

[112] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 97–115, Feb. 2001.

[113] N. Tosa and R. Nakatsu, "Life-like communication agent—emotion sensing character MIC and feeling session character MUSE," in *Proc. Int. Conf. Multimedia Computing and Systems*, 1996, pp. 12–19.

[114] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1999.

[115] E. Viennet and F. F. Soulie, “Multi-resolution scene segmentation by MLP’s,” in *Proc. IJCNN*, vol. 3, 1992, pp. 55–59.

[116] M. Wang, Y. Iwai, and M. Yachida, “Expression recognition from time-sequential facial images by use of expression change model,” in *Proc. FG*, 1998, pp. 324–329.

[117] A. Wierzbicka, “Reading human faces,” *Pragmat. Cogn.*, vol. 1, no. 1, pp. 1–23, 1993.

[118] C. E. Williams and K. N. Stevens, “Emotions and speech: Some acoustic correlates,” *J. Acoust. Soc. Amer.*, vol. 52, pp. 1238–1250, 1972.

[119] Y. Yacoob and L. Davis, “Recognizing facial expressions by spatio-temporal analysis,” in *Proc. ICPR*, vol. 1, 1994, pp. 747–749.

[120] M. H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 34–58, Jan. 2002.

[121] M. H. Yang, D. Roth, and N. Ahuja, “A SNoW-based face detector,” in *Proc. Neural Information Processing Systems*, vol. 12, 2000, pp. 855–861.

[122] Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, “Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face,” in *Proc. ROMAN*, 2000, pp. 178–183.

[123] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, “Facial expression recognition using thermal image processing and neural network,” in *Proc. ROMAN*, 1997, pp. 380–385.

[124] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and gabor wavelets-based facial expression recognition using multi-layer perceptron,” in *Proc. FG*, 1998, pp. 454–459.

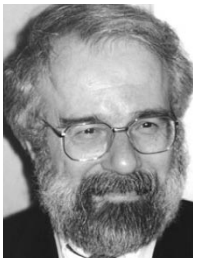
[125] L. Zhao, W. Lu, Y. Jiang, and Z. Wu, “A study on emotional feature recognition in speech,” in *Proc. ICSLP*, 2000, pp. 961–964.

[126] Y. Zhu, L. C. De Silva, and C. C. Ko, “Using moment invariants and HMM in facial expression recognition,” *Pattern Recognit. Lett.*, vol. 23, no. 1–3, pp. 83–91, 2002.

Maja Pantic (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Delft University of Technology, Delft, The Netherlands, in 1997 and 2001, respectively.

She is an Assistant Professor with the Data and Knowledge Systems Group, Mediamatics Department, Delft University of Technology. Her research interests pertain to the application of artificial intelligence and computational intelligence techniques in the analysis of different aspects of human behavior for the realization of perceptual, context-aware, multimodal human-computer interfaces.

Dr. Pantic is a Member of the Association for Computing Machinery and the American Association of Artificial Intelligence.



Leon J. M. Rothkrantz received the M.Sc. degree in mathematics from the University of Utrecht, Utrecht, The Netherlands, in 1971, the Ph.D. degree in mathematics from the University of Amsterdam, Amsterdam, The Netherlands, in 1980, and the M.Sc. degree in psychology from the University of Leiden, Leiden, The Netherlands, in 1990.

He is currently an Associate Professor with the Data and Knowledge Systems Group, Mediamatics Department, Delft University of Technology, Delft, The Netherlands, in 1992. His current research focuses on a wide range of the related issues, including lip reading, speech recognition and synthesis, facial expression analysis and synthesis, multimodal information fusion, natural dialogue management, and human affective feedback recognition. The long-range goal of his research is the design and development of natural, context-aware, multimodal man–machine interfaces.