Machine Understanding of Human Behavior*

Maja Pantic[1,3], Alex Pentland[2], Anton Nijholt[3] and Thomas Huang[4]

[1]Computing Department, Imperial College London, UK
[2]Media Lab, Massachusetts Institute of Technology, USA
[3]Faculty of EEMCS, University of Twente, The Netherlands
[4]Beckman Institute, University of Illinois at Urbana-Champaign, USA
m.pantic@imperial.ac.uk, pentland@media.mit.edu, a.nijholt@ewi.utwente.nl, huang@ifp.uiuc.edu

Abstract

A widely accepted prediction is that computing will move to the background, weaving itself into the fabric of our everyday living spaces and projecting the human user into the foreground. If this prediction is to come true, then next generation computing, which we will call *human computing*, should be about anticipatory user interfaces that should be human-centered, built for humans based on human models. They should transcend the traditional keyboard and mouse to include natural, human-like interactive functions including understanding and emulating certain human behaviors such as affective and social signaling. This article discusses a number of components of human behavior, how they might be integrated into computers, and how far we are from realizing the front end of human computing, that is, how far are we from enabling computers to understand human behavior.

1 Human Computing

Futuristic movies often contain visions of human environments of the future. Fitted out with arrays of intelligent, yet invisible devices, homes, transportation means and working spaces of the future can anticipate every need of their inhabitants (Fig. 1). This vision of the future is often referred to as "ubiquitous computing" [Weiser, 1991] or "ambient intelligence" [Aarts, 2005] . In this vision of the future, humans will be surrounded by intelligent interfaces that are supported by computing and networking technology embedded in all kinds of objects in the environment and that are sensitive and responsive to the presence of different individuals in seamless and unobtrusive way. This assumes a shift in computing – from desktop computers to a multiplicity of smart computing devices diffused into our environment. It assumes that computing will move to the background, weave itself into the fabric of everyday living spaces and disappear from the foreground, projecting the human user into it. However, as computing devices disappear from the scene, become invisible, weaved into our environment, a new set of issues is created concerning the interaction between this technology and humans [Nijholt et al., 2004, 2005, 2006; Streitz and Nixon, 2005; Zhai and Bellotti, 2005]. How can we design the interaction of humans with devices that are invisible? How can we design implicit interaction for sensor-based interfaces? What about users? What does a home dweller, for example, actually want? What are the relevant parameters that can be used by the systems to support us in our activities? If the context is key, how do we arrive at context-aware systems?

One way of tackling these problems is to move away from computer-centered designs toward human-centered designs for human computer interaction (HCI). The former involve usually the conventional interface devices like keyboard, mouse, and visual displays, and assume that the human will be explicit, unambiguous and fully attentive while controlling information and command flow. This kind of interfacing and categorical computing works well for context-independent tasks like making plane reservations and buying and selling stocks. However, it is utterly inappropriate for interacting with each of the (possibly hundreds) computer systems diffused throughout future smart environments and aimed at improving the quality of life by anticipating the users needs. The key to *human computing* and *anticipatory interfaces* is the ease of use, in this case the ability to unobtrusively sense certain behavioral cues of the users and to adapt automatically to his or hers typical behavioral patterns and the context in which he or she acts. Thus, instead of focusing on the computer portion of the HCI context, designs for human computing should focus on the human portion of the HCI context. They should go beyond the traditional keyboard and mouse to include natural, human-like interactive functions including understanding and emulating certain human behaviors like affective and social signaling. The design of these functions will require explorations of *what* is communicated (linguistic message, nonlin-

* This paper was originally published in the Proc. ACM Int'l Conf. Multimodal Interface 2006 (Copyright © ACM Press); see [Pantic et al., 2006]. The work of Maja Pantic and Anton Nijholt was partly supported by the EU 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

Fig. 1. Human environments of the future envisioned in movies: (left) hand-gesture-based interface and speech- & iris-id driven car (*Minority Report*, 2002), (right) multimedia diagnostic chart and a smart environment (*The Island*, 2005).

guistic conversational signal, emotion, attitude), *how* the information is passed on (the person's facial expression, head movement, nonlinguistic vocalization, hand and body gesture), *why*, that is, in which context the information is passed on (where the user is, what his or her current task is, are other people involved), and *which* (re)action should be taken to satisfy user needs and requirements.

This article discusses the front end of human computing, that is, what is communicated, how, and why [Pantic et al., 2006]. It focuses on certain human behaviors such as affective and social signaling, how they might be understood by computers, and how far we are from realizing the front end of human computing. For discussions about the back end of human computing, readers are referred to, e.g., [Nijholt et al., 2006; Ruttkay, 2006; Maat and Pantic, 2006].

2 Scientific and Engineering Issues

The scientific and engineering challenges related to the realization of machine sensing and understanding of human behaviors like affective and social signaling can be described as follows.

- **Which types of messages are communicated by behavioral signals?** This question is related to psychological issues pertaining to the nature of behavioral signals and the best way to interpret them.
- **Which human communicative cues convey information about a certain type of behavioral signals?** This issue shapes the choice of different modalities to be included into an automatic analyzer of human behavioral signals.
- **How are various kinds of evidence to be combined to optimize inferences about shown behavioral signals?** This question is related to issues such as how to distinguish between different types of messages, how best to integrate information across modalities, and what to take into account in order to realize context-aware interpretations.

Which types of messages are communicated by behavioral signals? The term behavioral signal is usually used to describe a set of temporal changes in neuromuscular and physiological activity that can last from a few milliseconds (a blink) to minutes (talking) or hours (sitting). Among the types of messages conveyed by behavioral signals are the following [Ekman and Friesen, 1969] (Fig. 2):

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

While there is agreement across different theories that at least some behavioral signals evolved to communicate information, there is lack of consensus regarding their specificity, extent of their innateness and universality, and whether they convey emotions, social motives, behavioral intentions, or all three [Izard, 1997]. Arguably the most often debated issue is whether affective states are a separate type of messages communicated by behavioral signals (i.e. whether behavioral signals communicate actually felt affect), or is the related behavioral signal (e.g. facial expression) just an illustrator / regulator aimed at controlling "the trajectory of a given social interaction", as suggested by Fridlund [1997]. Explanations of human behavioral signals in terms of internal states such as affective states are typical to psychological stream of thought, in particular to discrete emotion theorists who propose the existence of six or more basic emotions (happiness, anger, sadness, surprise, disgust, and fear) that are universally displayed and recognized from non-verbal behavioral signals (especially facial and vocal expression) [Keltner and Ekman, 2000; Juslin and Scherer, 2005]. Instead of explanations of human behavioral signals in terms of internal states, ethologists focus on consequences of behavioral displays for interpersonal interaction. As an extreme within the ethological line of thought, social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations. According to Fridlund, facial expressions should not be labeled in terms of emotions but in terms of Behavioral Ecology interpretations, which explain the influence a certain expression has in a particular context [Fridlund, 1997]. Thus, an "angry" face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. However, as proposed by Izard [1997], one may feel angry without the slightest intention of attacking anyone. In summary, is social communication the sole function of behavioral signals? Do they never represent visible manifestation of emotion / feeling / affective states? Since in some instances (e.g.

arachnophobia, acrophobia, object-elicited disgust, depression), affective states are not social, and their expressions necessarily have aspects other than "social motivation", we believe that affective states should be included into the list of types of messages communicated by behavioral signals. However, it is not only discrete emotions like surprise or anger that represent the affective states conveyed by human behavioral signals. Behavioral cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are essential components of human behavior. Hence, in contrast to traditional approach, which lists only (basic) emotions as the first type of message conveyed by behavioral signals [Ekman and Friesen, 1969], we treat affective states as being correlated not only to emotions but to other, aforementioned social signals and attitudinal states as well.

Which human communicative cues convey information about a certain type of behavioral signals? Manipulators are usually associated with self-manipulative gestures like scratching or lip biting and involve facial expressions and body gestures human communicative cues. Emblems, illustrators and regulators are typical social signals, spoken and wordless messages like head nods, bow ties, winks, 'huh' and 'yeah' utterances, which are sent by means of body gestures and postures, facial expressions and gaze, vocal expressions and speech. The most complex messages communicated by behavioral signals are affective and attitudinal states. Affective arousal modulates all human communicative signals. Hence, one could expect that automated analyzers of human behavior should include all human interactive modalities (audio, visual, and tactile) and should analyze all verbal and non-verbal interactive signals (speech, body gestures, facial and vocal expressions, and physiological reactions). However, we would like to make a few comments here. Although spoken language is between 200 thousand and 2 million years old [Gibson and Ingold, 1993], and speech has become the indispensable means for sharing ideas, observations, and feelings, findings in basic research indicate that in contrast to spoken messages [Furnas et al., 1987], nonlinguistic messages are the means to analyze and predict human behavior [Ambady and Rosenthal, 1992]. Anticipating a person's word choice and the associated intent is very difficult [Furnas et al., 1987]: even in highly constrained situations, different people choose different words to express exactly the same thing. As far as nonverbal cues are concerned, it seems that not all of them are equally important in the human judgment of behavioral signals. People commonly neglect physiological signals, 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 including pupillary diameter, heart rate, skin clamminess, temperature, respiration velocity [Cacioppo et al., 2000]. This and the recent advent of non-intrusive sensors and wearable computers, which promises less invasive physiological sensing [Starner, 2001], open up possibilities for including tactile modality into automatic analyzers of human behavior [Pentland, 2005]. However, the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioral cues [Ambady and Rosenthal, 1992]. Human judges seem to be most accurate in their judgment when they are able to observe the face and the body. Ratings that were based on the face and the body were 35% more accurate than the ratings that were based on the face alone. Yet, ratings that were based on the face alone were 30% more accurate than ratings that were based on the body alone and 35% more accurate than ratings that were based on the tone of voice alone [Ambady and Rosenthal, 1992]. These findings indicate that to interpret someone's behavioral cues, people rely on shown facial expressions and to a lesser degree on shown body gestures and vocal expressions. Note, however, that gestures like (Fig. 2) scratching (manipulator), thumbs up (emblem), finger pointing (illustrator), and head nods (regulator) are typical social signals. Basic research also provides evidence that observers tend to be accurate in decoding some negative basic emotions like anger and sadness from static body postures [Coulson, 2004] and that gestures like head inclination, face touching, and shifting posture often accompany social affective states like shame and embarrassment [Costa et al., 2001]. In addition, although cognitive scientists were unable to identify a set of vocal cues that reliably discriminate among affective and attitudinal states, listeners seem to be rather accurate in decoding some basic emotions from vocal cues like pitch and intensity [Juslin and Scherer, 2005] and some non-basic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns [Russell et al., 2003]. Thus, automated human behavior analyzers should at

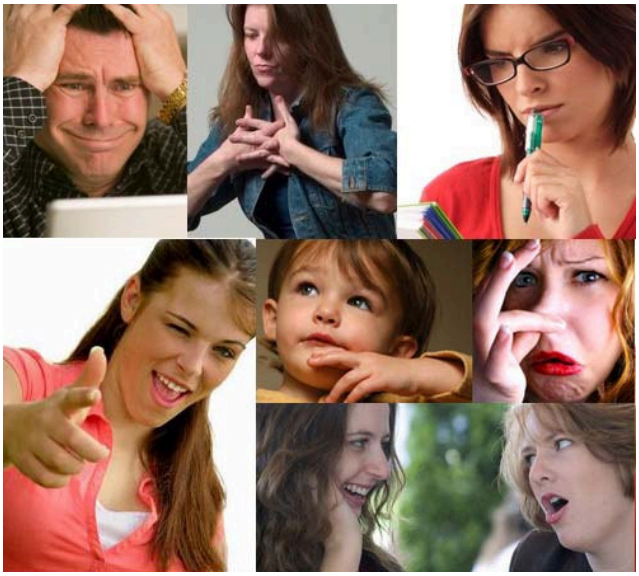

Fig. 2. Types of messages conveyed by behavioural signals: (1st row): affective/attitudinal states, (2nd row, clockwise from left) emblems, manipulators, illustrators, regulators.

least include facial expression and body gestures modalities and preferably they should also include modality for perceiving nonlinguistic vocalizations. Finally, while too much information from different channels seem to be confusing to human judges, resulting in less accurate judgments of shown behavior when three or more observation channels are available (face, body, and speech) [Ambady and Rosenthal, 1992], combining those multiple modalities (including physiology) may prove appropriate for realization of automatic human behavior analysis.

How are various kinds of evidence to be combined to optimize inferences about shown behavioral signals? Behavioral signals do not usually convey exclusively one type of messages but may convey any of the types (e.g. scratching is usually a manipulator but it may be displayed in an expression of confusion). It is crucial to determine to which class of behavioral signals a shown signal belongs since this influences the interpretation of it. For instance, squinted eyes may be interpreted as sensitivity of the eyes to bright light if this action is a reflex (a manipulator), as an expression of disliking if this action has been displayed when seeing someone passing by (affective cue), or as an illustrator of friendly anger on friendly teasing if this action has been posed (in contrast to being unintentionally displayed) during a chat with a friend, to mention just a few possibilities. To determine the class of an observed behavioral cue, one must know the context in which the observed signal has been displayed – where the expresser is (outside, inside, in the car, in the kitchen, etc.), what his or her current task is, are other people involved, and who the expresser is. The latter is of particular importance for recognition of affective and attitudinal states since it is not probable that each of us will express a particular affective state by modulating the same communicative signals in the same way, especially when it comes to affective states other than basic emotions. Since the problem of context-sensing is extremely difficult to solve (if possible at all) for a general case, we advocate that a pragmatic approach (e.g. activity/application- and user-centered approach) must be taken when learning the grammar of human expressive behavior. In addition, because of the impossibility of having users instructing the computers for each possible application, we propose that methods for unsupervised (or semi-supervised) learning must be applied. Moreover, much of human expressive behavior is unintended and unconscious; the expressive nonverbal cues can be so subtle that they are neither encoded nor decoded at an intentional, conscious level of awareness [Ambady and Rosenthal, 1992]. This suggests that the learning methods inspired by human unconscious problem solving processes may prove more suitable for automatic human behavior analysis than the learning methods inspired by human conscious problem solving processes [Valstar and Pantic, 2006a]. Another important issue is that of multimodal fusion. A number of concepts relevant to fusion of sensory neurons in humans may be of interest [Stein and Meredith, 1993]:

- *1+1 >2*: The response of multi-sensory neurons can be stronger for multiple weak input signals than for a single strong signal.
- *Context dependency*: The fusion of sensory signals is modulated depending on the sensed context – for different contexts, different combinations of sensory signals are made.
- *Handling of discordances*: Based on the sensed context, sensory discordances (malfunctioning) are either handled by fusing sensory signals 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).

Thus, humans simultaneously employ the tightly coupled audio, visual, and tactile modalities. As a result, analysis of the perceived information is highly robust and flexible. Hence, one could expect that in an automated analyzer of human behavior input signals should not be considered mutually independent and should not be combined only at the end of the intended analysis, as the majority of current studies do, but that they should be processed in a joint feature space and according to a context-dependent model [Pantic and Rothkrantz, 2003]. However, does this tight coupling persists when the modalities are used for multimodal interfaces as proposed by some researchers (e.g. [Gunes and Piccardi, 2005]), or not, as suggested by others (e.g. [Scanlon and Reilly, 2001])? This remains an open, highly relevant issue.

3 State of the Field

Human sensing: Sensing human behavioral signals including facial expressions, body gestures, nonlinguistic vocalizations, and vocal intonations, which seem to be most important in the human judgment of behavioral cues [Ambady and Rosenthal, 1992], involves a number of tasks.

- Face: face detection and location, head and face tracking, eye-gaze tracking, and facial expression analysis.
- Body: body detection and tracking, hand tracking, recognition of postures, gestures and activity.
- Vocal nonlinguistic signals: estimation of auditory features such as pitch, intensity, and speech rate, and recognition of nonlinguistic vocalizations like laughs, cries, sighs, and coughs.

Because of its practical importance and relevance to face recognition, face detection received the most attention of the tasks mentioned above. Numerous techniques have been developed for face detection, i.e., identification of all regions in the scene that contain a human face [Yang et al., 2002; Li and Jain, 2005]. However, virtually all of them can detect only (near-) upright faces in (near-) frontal view. Most of these methods emphasize statistical learning techniques and use appearance features, including the real-time face detection scheme proposed by Viola and Jones [2004], which is arguably the most commonly employed face de-

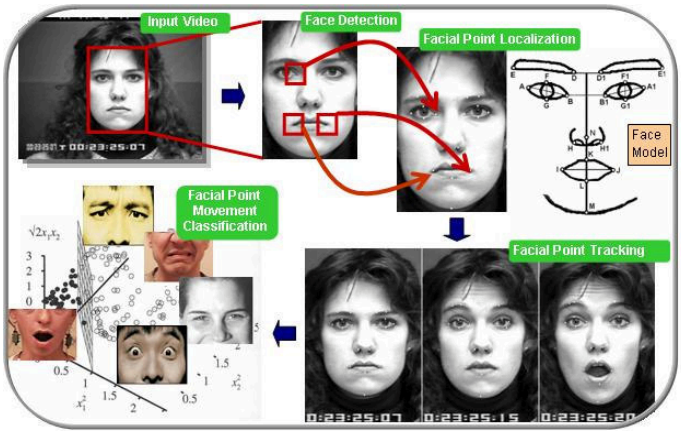

Fig. 3. An AU detection method [Valstar & Pantic, 2006b].

tector in automatic facial expression analysis. Note, however, that one of the few methods that can deal with tilted face images represents a feature-based rather than an appearance-based approach to face detection [Chiang and Huang, 2005].

Tracking is an essential step for human motion analysis since it provides the data for recognition of face/head/body postures and gestures. Optical flow has been widely used for head, face and facial feature tracking [Wang and Singh, 2003]. To omit the limitations inherent in optical flow techniques such as the accumulation of error and the sensitivity to occlusion, clutter, and changes in illumination, researchers in the field started to use sequential state estimation techniques like Kalman and particle filtering schemes [Haykin and Freitas, 2004]. Some of the most advanced approaches to head tracking and head-pose estimation are based on Kalman (e.g. [Huang and Trivedi, 2004]) and particle filtering frameworks (e.g. [Ba and Odobez, 2004]). Similarly, the most advanced approaches to facial feature tracking are based on Kalman (e.g. [Gu and Ji, 2005]) and particle filtering tracking schemes (e.g. [Valstar and Pantic, 2006b]). Although face pose and facial feature tracking technologies have improved significantly in the recent years with sequential state estimation approaches that run in real time, tracking multiple, possibly occluded, expressive faces, their poses, and facial feature positions simultaneously in unconstrained environments is still a difficult problem. The same is true for eye gaze tracking [Duchowski, 2002]. To determine the direction of the gaze, eye tracking systems employ either the so-called red-eye effect, i.e., the difference in reflection between the cornea and the pupil, or computer vision techniques to find the eyes in the input image and then determine the orientation of the irises. Although there are now several companies that sell commercial eye trackers like SMI GmbH, EyeLink, Tobii, Interactive Minds, etc., realizing non-intrusive (non-wearable), fast, robust, and accurate eye tracking remains a difficult problem even in computer-centred HCI scenarios in which the user is expected to remain in front of the computer but is allowed to shift his or her position in any direction for more than 30 cm.

Because of the practical importance of the topic for affective, perceptual, and ambient interfaces of the future and theoretical interest from cognitive scientists [Lisetti and Schiano, 2000; Pantic and Rothkrantz, 2003], automatic analysis of facial expressions attracted the interest of many researchers. Most of the facial expressions analyzers developed so far attempt to recognize a small set of prototypic emotional facial expressions such as happiness or sadness (see also the state of the art in facial affect recognition in the text below) [Pantic and Rothkrantz, 2003]. To facilitate detection of subtle facial signals like a frown or a smile and to make facial expression information available for usage in applications like anticipatory ambient interfaces, several research groups begun research on machine analysis of facial muscle actions (atomic facial cues, action units, AUs, [Ekman et al., 2002]). A number of promising prototype systems have been proposed recently that can recognize 15 to 27 AUs (from a total of 44 AUs) in either (near-) frontal view or profile view face image sequences [Li and Jain, 2005; Pantic and Patras, 2006]. Most of these employ statistical and ensemble learning techniques and are either feature-based (i.e., use geometric features like facial points or shapes of facial components, e.g., see Fig. 3) or appearance-based (i.e., use texture of the facial skin including wrinkles, bulges, and furrows). It has been reported that methods based on appearance features usually outperform those based on geometric features. Recent studies have shown that this claim does not always hold [Pantic and Patras, 2006]. Besides, it seems that using both geometric and appearance features might be the best choice for certain facial cues [Pantic and Patras, 2006]. However, the present systems for facial AU detection typically depend on accurate head, face and facial feature tracking as input and are still very limited in performance and robustness.

Vision-based analysis of hand and body gestures is nowadays one of the most active fields in computer vision. Tremendous amount of work has been done in the field in the recent years [Wang and Singh, 2003; Wang et al., 2003]. Most of the proposed techniques are either model-based (i.e., use geometric primitives like cones and spheres to model head, trunk, limbs and fingers) or appearance-based (i.e., use color or texture information to track the body and its parts). Most of these methods emphasize Gaussian models, probabilistic learning, and particle filtering framework (e.g. [Sand and Teller, 2006; Stenger et al., 2006]. However, body and hands detection and tracking in unconstrained environments where large changes in illumination and cluttered or dynamic background may occur still pose significant research challenges. Also, in casual human behavior, the hands do not have to be always visible (in pockets, under the arms in a crossed arms position, on the back of the neck and under the hair), they may be in a cross fingered position, and one hand may be (partially) occluded by the other. Although some progress has been made to tackle these problems using the knowledge on human kinematics, most of the present methods cannot handle such cases correctly.

In contrast to the linguistic part of a spoken message (*what* has been said) [Furnas et al., 1987], the nonlinguistic part of it (*how* it has been said) carries important information about the speaker's affective state [Juslin and Scherer, 2005] and attitude [Russell et al., 2003]. This finding instigated the research on automatic analysis of vocal nonlinguistic expressions. The vast majority of present work is aimed at discrete emotion recognition from auditory features like pitch, intensity, and speech rate (see the state of the art in vocal affect recognition in the text below) [Oudeyer, 2003; Pantic and Rothkrantz, 2003]. For the purposes of extracting auditory features from input audio signals, freely available signal processing toolkits like Praat[1] are usually used. More recently, few efforts towards automatic recognition of nonlinguistic vocalizations like laughs [Truong and van Leeuwen, 2005], cries [Pal et al., 2006], and coughs [Matos et al., 2006] have been also reported. Since the research in cognitive sciences provided some promising hints that vocal outbursts and nonlinguistic vocalizations like yelling, laughing, and sobbing, may be very important cues for decoding someone's affect/attitude [Russell et al., 2003], we suggest a much broader focus on machine recognition of these nonlinguistic vocal cues.

Context sensing: Context plays a crucial role in understanding of human behavioral signals, since they are easily misinterpreted if the information about the situation in which the shown behavioral cues have been displayed is not taken into account [Pantic and Rothkrantz, 2003]. For computing technology applications, context can be defined as any information that can be used to characterize the situation that is relevant to the interaction between users and the application [Dey et al., 2001]. Six questions summarize the key aspects of the computer's context with respect to nearby humans:

- *Who?* (Who the user is?)
- *Where?* (Where the user is?)
- *What?* (What is the current task of the user?)
- *How?* (How the information is passed on? Which behavioral signals have been displayed?)
- *When?* (What is the timing of displayed behavioral signals with respect to changes in the environment? Are there any co-occurrences of the signals?)
- *Why?* (What may be the user's reasons to display the observed cues? Except of the user's current task, the issues to be considered include the properties of the user's physical environment like lighting and noise level, and the properties of the current social situation like whether the user is alone and what is his or her psychological state.)

Here, we focus on answering context questions relating to the human-part of the computer's context. The questions related exclusively to the user's context and not to the computer's context like what kind of people are the user's communicators and what the overall social situation is, are considered irrelevant for adapting and tailoring the computing technology to its human users and are not discussed in this article.

Because of its relevance for the security, the *who* context question has received the most attention from both funding agencies and commercial enterprises and, in turn, it has seen the most progress. The biometrics market has increased dramatically in recent years, with multiple companies providing face recognition systems like Cognitec and Identix, whose face recognition engines achieved repeatedly top 2D face recognition scores in USA government testing (FRGC, FRVT 2002, FERET 1997). The problem of face recognition has been tackled in various ways in 2D and 3D, using feature-, shape-, and appearance-based approaches as well as the combinations thereof [Zhao et al., 2003; Li and Jain, 2005; Bowyer et al., 2006]. The majority of the present methods employ spectral methods for dimensionality reduction like PCA, LDA, and ICA. Except of the face, biometric systems can be based on other biometric traits like fingerprints, voice, iris, retina, gait, ear, hand geometry, and facial thermogram [Jain and Ross, 2004]. Biometric systems should be deployed in real-world applications and, in turn, should be able to handle a variety of problems including sensor malfunctioning, noise in sensed data, intra-class variations (e.g. facial expression which is treated as noise in face recognition), and spoof attacks (i.e. falsification attempts). Since most of these problems can be overcome by using multiple biometric traits [Jain and Ross, 2004], multimodal biometric systems have recently become a research trend. The most commonly researched multi-biometrics relate to audiovisual speaker recognition. For a survey of commercial systems for alternative biometrics, see [BTT Survey, 2006]. For current research efforts in multi-biometrics, see [MMUA, 2006].

Similarly to the *who* context question, security concerns also drive the research tackling the *where* context-sensing problem, which is typically addressed as a computer-vision problem of surveillance and monitoring. The work in this area is based on one or more unobtrusively mounted cameras used to detect and track people. The process usually involves [Wang et al., 2003]: scene (background) modeling, motion segmentation, object classification, and object tracking. The vast majority of scene modeling approaches can be classified as generative models [Buxton, 2003]. However, generative approaches, which require excessive amount of training data, are not appropriate for complex and incomplete problem domains like dynamic scene modeling. Unsupervised learning techniques are a better choice in that case. Motion segmentation aims at detecting regions in the scene which correspond to moving objects like cars and humans. It is one of the oldest computer vision problems and it has been tackled in various ways including [Wang et al., 2003]: background subtraction, temporal differencing, optical flow, watershed, region growing, scene mosaicing, statistical and Bayesian methods. Since natural scenes may contain multiple moving regions that may correspond to different entities, it is crucial to distinguish those that correspond to humans for the purposes of sensing the human part of the com-

[1]Praat: http://www.praat.org.

puter's context. Note that this step is superfluous where the moving objects are known to be humans. Present methods to moving object classification are usually either shape-based (e.g. human-silhouette-based) or motion-based (i.e. employ the premise that human articulated motion shows a periodic property) [Wang et al., 2003]. When it comes to human tracking for the purposes of answering the *where* context question, typically employed methods emphasize probabilistic methods like Dynamic Bayesian Networks and sequential state estimation techniques like Kalman and particle filtering schemes [Wang and Singh, 2003; Wang et al., 2003]. In summary, since most approaches base their analysis on segmentation and tracking, these present methods are adequate when a priori knowledge is available (e.g. the shape of the object to be tracked), but they are weak for unconstrained environments (e.g. gym, a house party), in which multiple occlusions and clutter may be present. For such cases, methods that perform analysis at the lowest semantic level (i.e. consider only temporal pixel-based behaviour) and use unsupervised learning represent a better solution (e.g. [Bicego et al., 2006]).

In desktop computer applications, the user's task identification (i.e., the *what* context question) is usually tackled by determining the user's current focus of attention by means of gaze tracking, finger pointing, or simply based on the knowledge of current events like keystrokes, mouse movements, and active software (e.g. web browser, e-mail manager). However, as traditional HCI and usability-engineering applications involve relatively well-defined user tasks, many of the methods developed for user task analysis in typical HCI domains are inappropriate for task analysis in the context of human computing and ubiquitous, anticipatory ambient interfaces, where the tasks are often ill-defined due to uncertainty in the sensed environmental and behavioral cues. Analysis of tasks that human may carry out in the context of anticipatory ambient interfaces require adaptation and fusion of existing methods for behavioral cues recognition (e.g. hand/body gesture recognition, focus of attention identification) and those machine learning techniques that can be applicable to solving ill-structured decision-making problems (e.g. Markov decision processes and hidden-state models). However, only a very limited research has been directed to multimodal user's task identification in the context of anticipatory ambient interfaces and the majority of this work is aimed at support of military activities (e.g. airplane cockpit control) and crisis management [Sharma et al., 2003]. Other methods for human activity recognition typically identify the task of the observed person in an implicit manner, by recognizing different tasks as different activities. The main shortcoming of these approaches is the increase of the problem dimensionality – for the same activity, different recognition classes are defined, one for each task (e.g. for the sitting activity, categories like watching TV, dining, and working with desktop computer, may be defined).

The *how* context question is usually addressed as a problem of human sensing (see the state of the art in human sensing in the text above; for a survey on speech recognition see [Deng and Huang, 2004]). When it comes to desktop computer application, additional modalities like writing, keystroke (choice and rate), and mouse gestures (clicks and movements) may be considered as well when determining the information that the user has passed on.

There is now a growing body of psychological research that argues that temporal dynamics of human behavior (i.e., the timing and the duration of behavioral cues) is a critical factor for interpretation of the observed behavior [Russell et al., 2003]. For instance, it has been shown that spontaneous smiles, in contrast to volitional smiles (like in irony), are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within 1s. In spite of these findings in basic research and except few studies on facial expression analysis [Valstar et al., 2006], present methods for human activity/behavior recognition do not address the *when* context question: the timing of displayed behavioral signals with respect to other behavioral signals is usually not taken into account. When it comes to the timing of shown behavioral signals with respect to changes in the environment, current methods typically approach the *when* question in an implicit way, by recognizing user's reactions to different changes in the environment as different activities.

The *why* context question is arguably the most complex and the most difficult to address context question. It requires not only detection of physical properties of the user's environment like the lighting and noise level (which can be easily determined based on the current illumination intensity and the level of auditory noise) and analysis of whether the user is alone or not (which can be carried out by means of the methods addressing the *where* context question), but understanding of the user's behavior and intentions as well (see the text below for the state of the art in human behavior understanding).

As can be seen from the overview of the current state of the art in so-called W5+ (who, where, what, when, why, how) technology, context questions are usually addressed separately and often in an implicit manner. Yet, the context questions may be more reliably answered if they are answered in groups of two or three using the information extracted from multimodal input streams. Some experimental evidence supports this hypothesis [Nock et al., 2004]. For example, solutions for simultaneous speaker identification (*who*) and location (*where*) combining the information obtained by multiple microphones and surveillance cameras had an improved accuracy in comparison to single-modal and single-aspect approaches to context sensing. A promising approach to realizing multimodal multi-aspect context-sensing has been proposed by Nock et al. [2004]. In this approach, the key is to automatically determine whether observed behavioral cues share a common cause (e.g. whether the mouth movements and audio signals complement to indicate an active known or unknown speaker (how, who, where) and whether his or her focus of attention is another person or a computer (what, why)). The main advantages of such an approach are effective handling of uncertainties due to noise in input data streams and the prob-

lem-dimensionality reduction. Therefore, we suggest a much broader focus on spatial and temporal, multimodal multi-aspect context-sensing.

Understanding human behavior: Eventually, automated human behavior analyzers should terminate their execution by translating the sensed human behavioral signals and context descriptors into a description of the shown behavior. The past work in this field can be roughly divided into the methods for understanding human affective / attitudinal states and those for understanding human social signaling (i.e., emblems, regulators, and illustrators).

Understanding Human Affect: As soon as research findings in HCI and usability engineering have suggested that HCI systems which will be capable of sensing and responding properly to human affective states are likely to be perceived as more natural, efficacious, and trustworthy, the interest in human affect machine analysis has surged. The existing body of literature in machine analysis of human affect is immense [Pantic and Rothkrantz, 2003; Oudeyer, 2003; Li and Jain, 2005]. Most of these works attempt to recognize a small set of prototypic expressions of basic emotions like happiness and anger from either face images/video or speech signal. They achieve an accuracy of 64% to 98% when detecting 3-7 emotions deliberately displayed by 5-40 subjects. However, the capabilities of these current approaches to human affect recognition are rather limited.

- Handle only a small set of volitionally displayed prototypic facial or vocal expressions of six basic emotions.
- Do not perform a context-sensitive analysis (either user-, or environment-, or task-dependent analysis) of the sensed signals.
- Do not analyze extracted facial or vocal expression information on different time scales (i.e., short videos or vocal utterances of a single sentence are handled only). Consequently, inferences about the expressed mood and attitude (larger time scales) cannot be made by current human affect analyzers.
- Adopt strong assumptions. For example, facial affect analyzers can typically handle only portraits or nearly-frontal views of faces with no facial hair or glasses, recorded under constant illumination and displaying exaggerated prototypic expressions of emotions. Similarly, vocal affect analyzers assume usually that the recordings are noise free, contain exaggerated vocal expressions of emotions, i.e., sentences that are short, delimited by pauses, and carefully pronounced by non-smoking actors.

Few exceptions from this overall state of the art in the field include a few tentative efforts to detect attitudinal and non-basic affective states such as boredom, fatigue, and pain from face video [e.g., El Kaliouby and Robinson, 2004; Bartlett et al., 2006], a few works on context-sensitive interpretation of behavioral cues like facial expressions [Pantic, 2006], and an attempt to discern spontaneous from volitionally displayed facial behavior [Valstar et al., 2006]. Few works have been also proposed that combine several modalities into a single system for human affect analysis. Although the studies in basic research suggest that the combined face and body are the most informative for the analysis of human expressive behavior [Ambady and Rosenthal, 1992], only 2-3 efforts are reported on automatic human affect analysis from combined face and body gestures [Gunes and Piccardi, 2005]. Existing works combining different modalities into a single system for human affective state analysis investigated mainly the effects of a combined detection of facial and vocal expressions of affective states [Pantic and Rothkrantz, 2003; Song et al., 2004; Zeng et al., 2006]. In general, these works achieve an accuracy of 72% to 85% when detecting one or more basic emotions from clean audiovisual input (e.g., noise-free recordings, closely-placed microphone, non-occluded portraits) from an actor speaking a single word and showing exaggerated facial displays of a basic emotion. Thus, present systems for multimodal human affect analysis have all (and some additional) drawbacks of single-modal analyzers. Hence, many improvements are needed if those systems are to be used for context-sensitive analysis of human behavioral signals where a clean input from a known actor/ announcer cannot be expected and a context-independent processing and interpretation of audiovisual data do not suffice.

An additional important issue is that we cannot conclude that a system attaining a 92% average recognition rate performs "better" than a system achieving a 74% average recognition rate when detecting six basic emotions from audio and/or visual input stream unless both systems are tested on the same dataset. The main problem is that no audiovisual database exists that is shared by all diverse research communities in the field [Pantic and Rothkrantz, 2003]. Although efforts have been recently reported towards development of benchmark databases that can be shared by the entire research community [Pantic et al., 2005; Gunes and Piccardi, 2005], this remains an open, highly relevant issue.

Understanding Human Social Signaling: As we already remarked above, research findings in cognitive sciences tend to agree that at least some (if not the majority) of behavioral cues evolved to facilitate communication between people [Izard, 1997]. Types of messages conveyed by these behavioral cues include emblems, illustrators, and regulators, which can be further interpreted in terms of social signaling like turn taking, mirroring, empathy, antipathy, interest, engagement, agreement, disagreement, etc. Although each one of us understands the importance of social signaling in everyday life situations, and although a firm body of literature in cognitive sciences exists on the topic [Ambady and Rosenthal, 1992; Russell and Fernandez-Dols, 1997; Russell et al., 2003] and in spite of recent advances in sensing and analyzing behavioral cues like blinks, smiles, winks, thumbs up, yawns, laughter, etc. (see the state of the art in human sensing in the text above), the research efforts in machine analysis of human social signaling are few and tentative. An important part of the existing research on understanding human social signaling has been conducted at

MIT Media Lab, under the supervision of Alex Pentland [2005]. Their approach aims to discern social signals like activity level, stress, engagement, and mirroring by analyzing the engaged persons' tone of voice. Other important works in the field include efforts towards analysis of interest, agreement and disagreement from facial and head movements [El Kaliouby and Robinson, 2004] and towards analysis of the level of interest from tone of voice, head and hand movements [Gatica-Perez et al., 2005]. Overall, present approaches to understanding social signaling are multimodal and based on probabilistic reasoning methods like Dynamic Bayesian Networks. However, most of these methods are context insensitive (key context issues are either implicitly addressed, i.e., integrated in the inference process directly, or they are ignored altogether) and incapable of handling unconstrained environments correctly. Thus, although these methods represent promising attempts toward encoding of social variables like status, interest, determination, and cooperation, which may be an invaluable asset in the development of social networks formed of humans and computers (like in the case of virtual worlds), in their current form, they are not appropriate for general anticipatory interfaces.

4 Research Challenges

According to the taxonomy of human movement, activity, and behavioral action proposed by Bobick [1997], movements are low-level semantic primitives, requiring no contextual or temporal knowledge for the detection. Activities are sequences of states and movements, where the only knowledge required to recognize them relates to statistics of the temporal sequence. As can be seen from the overview of the past work done in the field, most of the work on human gesture recognition and human behavior understanding falls in this category. Human behavioral actions, or simply human behavior, are high-level semantic events, which typically include interactions with the environment and causal relationships. An important distinction between these different semantic levels of human behavior representation is the degree to which the context, different modalities, and time must be explicitly represented and manipulated, ranging from simple spatial reasoning to context-constrained reasoning about multimodal events shown in temporal intervals. However, most of the present approaches to machine analysis of human behavior are neither multimodal, nor context-sensitive, nor suitable for handling longer time scales. In our survey of the state of the field, we have tried to explicitly mention most of the existing exceptions from this rule in an attempt to motivate researchers in the field to treat the problem of context-constrained analysis of multimodal behavioral signals shown in temporal intervals as one complex problem rather than a number of detached problems in human sensing, context sensing, and human behavior understanding. Besides this critical issue, there are a number of scientific and technical challenges that we consider essential for advancing the state of the art in the field.

Scientific challenges in human behavior understanding can be summarized as follows.

- *Modalities:* How many and which behavioral channels like the face, the body, and the tone of the voice, should be combined for realization of robust and accurate human behavior analysis? Too much information from different channels seems to be confusing for human judges. Does this pertain in HCI?
- *Fusion:* At which abstraction level are these modalities to be fused? Humans simultaneously employ modalities of sight and sound. Does this tight coupling persists when the modalities are used for human behavior analysis, as suggested by some researchers, or not, as suggested by others? Does this depend on the machine learning techniques employed or not?
- *Fusion & Context:* While it has been shown that the *1+1>2* concept relevant to fusion of sensory neurons in humans pertain in machine context sensing [Nock et al., 2004], does the same hold for the other two concepts relevant to multimodal fusion in humans (i.e. context-dependent fusion and discordance handling)? Note that context-dependent fusion and discordance handling were never attempted.
- *Dynamics & Context:* Since the dynamics of shown behavioral cues play a crucial role in human behavior understanding, how the grammar (i.e., temporal evolvement) of human behavioral displays can be learned? Since the grammar of human behavior is context-dependent, should this be done in a user-centered manner [Oviatt, 2003] or in an activity/application-centered manner [Norman, 2005]?
- *Learning vs. Education:* What are the relevant parameters in shown human behavior that an anticipatory interface can use to support humans in their activities? How this should be (re-) learned for novel users and new contexts? Instead of building machine learning systems that will not solve any problem correctly unless they have been trained on similar problems, we should build systems that can be educated, that can improve their knowledge, skills, and plans through experience. Lazy and unsupervised learning can be promising for realizing this goal.

Technical challenges in human behavior understanding can be summarized as follows.

- *Initialization:* A large number of methods for human sensing, context sensing, and human behavior understanding require an initialization step. Since this is typically a slow, tedious, manual process, fully automated systems are the only acceptable solution when it comes to anticipatory interfaces of the future.
- *Robustness:* Most methods for human sensing, context sensing, and human behavior under-

standing work only in (often highly) constrained environments. Noise, fast movements, changes in illumination, etc., cause them to fail.

- *Speed:* Many of the methods in the field do not perform fast enough to support interactivity. Researchers usually choose for more sophisticated (but not always smarter) processing rather than for real time processing. A typical excuse is that according to Moore's Law we'll have faster hardware soon enough.
- *Training & Validation Issues:* United efforts of different research communities working in the field should be made to develop a comprehensive, readily accessible database of annotated, multimodal displays of human expressive behavior recorded under various environmental conditions, which could be used as a basis for benchmarks for efforts in the field. The related research questions include the following. How one can elicit spontaneous expressive behavior including genuine emotional responses and attitudinal states? How does one facilitate efficient, fast, and secure retrieval and inclusion of objects constituting this database? How could the performance of a tested automated system be included into the database? How should the relationship between the performance and the database objects used in the evaluation be defined?

5 Conclusions

Human behavior understanding is a complex and very difficult problem, which is still far from being solved in a way suitable for anticipatory interfaces and human computing application domain. In the past two decades, there has been significant progress in some parts of the field like face recognition and video surveillance (mostly driven by security applications), while in the other parts of the field like in non-basic affective states recognition and multimodal multi-aspect context-sensing at least the first tentative attempts have been proposed. Although the research in these different parts of the field is still detached, and although there remain significant scientific and technical issues to be addressed, we are optimistic about the future progress in the field. The main reason is that anticipatory interfaces and their applications are likely to become the single most widespread research topic of AI and HCI research communities. Even nowadays, there are a large and steadily growing number of research projects concerned with the interpretation of human behavior at a deeper level.

References

[Aarts, 2005] E. Aarts. Ambient intelligence drives open innovation. *ACM Interactions,* 12(4): 66-68, July-Aug. 2005.

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

[Ba and Odobez, 2004] S.O. Ba and J.M. Odobez. A probabilistic framework for joint head tracking and pose estimation. In *Proc. Conf. Pattern Recognition, vol. 4*, pp. 264-267, 2004.

[Bartlett et al., 2006] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. Conf. Face & Gesture Recognition*, pp. 223-230, 2006.

[Bicego et al., 2006] M. Bicego, M. Cristani and V. Murino. Unsupervised scene analysis: A hidden Markov model approach. *Computer Vision & Image Understanding,* 102(1): 22-41, Apr. 2006.

[Bobick, 1997] A.F. Bobick. Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Trans. Roy. Soc. London B,* 352(1358): 1257-1265, Aug. 1997.

[Bowyer et al., 2006] K.W. Bowyer, K. Chang and P. Flynn. A survey of approaches and challenges in 3D and multimodal 3D+2D face recognition. *Computer Vision & Image Understanding,* 101(1): 1-15, Jan. 2006.

[Buxton, 2003] H. Buxton. Learning and understanding dynamic scene activity: a review. *Image & Vision Computing,* 21(1): 125-136, Jan. 2003.

[Cacioppo et al., 2000] 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. Haviland-Jones, Eds. Guilford Press, New York, 2000, pp. 173-191.

[Chiang and Huang, 2005] C.C. Chiang and C.J. Huang. A robust method for detecting arbitrarily tilted human faces in color images. *Pattern Recognition Letters*, 26(16): 2518-2536, Dec. 2005.

[Costa et al., 2001] M. Costa, W. Dinsbach, A.S.R. Manstead and P.E.R. Bitti. Social presence, embarrassment, and nonverbal behavior. *Journal of Nonverbal Behavior,* 25(4): 225-240, Dec. 2001.

[Coulson, 2004] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, & viewpoint dependence. *J. Nonverbal Behavior,* 28(2): 117-139, Jun. 2004.

[Deng and Huang, 2004] B.L. Deng and X. Huang. Challenges in adopting speech recognition. *Communications of the ACM,* 47(1): 69-75, Jan. 2004.

[Dey et al., 2001] A.K. Dey, G.D. Abowd and D.A. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *J. Human-Computer Interaction,* 16(2-4): 97-166, Dec. 2001.

[Duchowski, 2002] A.T. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments and Computing*, 34(4): 455-470, Nov. 2002.

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

[Ekman et al., 2002] P. Ekman, W.V. Friesen and J.C. Hager. *Facial Action Coding System*. A Human Face, Salt Lake City, 2002.

[El Kaliouby and Robinson, 2004] R. El Kaliouby and P. Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Proc. Int'l Conf. Computer Vision & Pattern Recognition, vol. 3*, p. 154, 2004.

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

[Furnas et al., 1987] G. Furnas, T. Landauer, L. Gomes and S. Dumais. The vocabulary problem in human-system communication, *Communications of the ACM,* 30(11): 964-972, Nov. 1987.

[Gatica-Perez et al., 2005] D. Gatica-Perez, I. McCowan, D. Zhang and S. Bengio. Detecting group interest level in meetings. In *Proc. Int'l Conf. Acoustics, Speech & Signal Processing, vol. 1*, pp. 489-492, 2005.

[Gibson and Ingold, 1993] K.R. Gibson and T. Ingold, Eds. *Tools, Language and Cognition in Human Evolution*. Cambridge University Press, Cambridge, UK, 1993.

[Gu and Ji, 2005] H. Gu and Q. Ji. Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications,* 16(2): 105-115, Feb. 2005.

[Gunes and Piccardi, 2005] H. Gunes and M. Piccardi. Affect Recognition from Face and Body: Early Fusion vs. Late Fusion, In *Proc. Int'l Conf. Systems, Man and Cybernetics*, pp. 3437- 3443, 2005.

[Haykin and de Freitas, 2004] S. Haykin and N. de Freitas, Eds. Special Issue on Sequential State Estimation. *Proceedings of the IEEE,* 92(3): 399-574, Mar. 2004.

[Huang and Trivedi, 2004] K.S. Huang and M.M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video. In *Proc. Conf. Pattern Recognition, vol. 3*, pp. 965-968, 2004.

[Izard, 1997] C.E. Izard. Emotions and facial expressions: A perspective from Differential Emotions Theory. In *The psychology of facial expression*. J.A. Russell and J.M. Fernandez-Dols, Eds. Cambridge University Press, Cambridge, UK, 1997, pp. 103-129.

[Jain and Ross, 2004] A.K. Jain and A. Ross. Multibiometric systems. *Communications of the ACM,* 47(1): 34-40, Jan. 2004.

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

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

[Li and Jain, 2005] S.Z. Li and A.K. Jain, Eds. *Handbook of Face Recognition*. Springer, New York, 2005.

[Lisetti and Schiano, 2000] C.L. Lisetti and D.J. Schiano. Automatic facial expression interpretation: Where human-computer interaction, AI and cognitive science intersect. *Pragmatics and Cognition*, 8(1): 185-235, Jan. 2000.

[Maat and Pantic, 2006] L. Maat and M. Pantic. Gaze-X: Adaptive affective multimodal interface for single-user office scenarios. In *Proc. Int'l Conf. Multimodal Interfaces*, pp. 171-178, 2006.

[Matos et al., 2006] S. Matos, S.S. Birring, I.D. Pavord and D.H. Evans. Detection of cough signals in continuous audio recordings using HMM. *IEEE Trans. Biomedical Engineering,* 53(6): 1078-1083, June 2006.

[Nijholt et al., 2004] A. Nijholt, T. Rist and K. Tuinenbreijer. Lost in ambient intelligence. In *Proc. Int'l Conf. Computer Human Interaction*, pp. 1725-1726, 2004.

[Nijholt et al., 2006] A. Nijholt, B. de Ruyter, B., D. Heylen, and S. Privender. Social Interfaces for Ambient Intelligence Environments. Chapter 14 in: *True Visions: The Emergence of Ambient Intelligence*. E. Aarts and J. Encarnaçao, Eds. Springer, New York, 2006, pp. 275-289.

[Nijholt and Traum, 2005] A. Nijholt and D. Traum. The Virtuality Continuum Revisited. In *Proc. Int'l Conf. Computer Human Interaction*, pp. 2132-2133, 2005.

[Nock et al., 2004] H.J. Nock, G. Iyengar and C. Neti. Multimodal processing by finding common cause. *Communications of the ACM,* 47(1): 51-56, Jan. 2004.

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

[Oudeyer, 2003] P.Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *Int'l J. Human-Computer Studies,* 59(1-2): 157-183, July 2003.

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

[Pal et al., 2006] P. Pal, A.N. Iyer and R.E. Yantorno. Emotion detection from infant facial expressions and cries. In *Proc. Int'l Conf. Acoustics, Speech & Signal Processing, 2*, pp. 721-724, 2006.

[Pantic, 2006] M. Pantic. Face for Ambient Interface. *Lecture Notes in Artificial Intelligence*, 3864: 35-66, 2006.

[Pantic and Patras, 2006] M. Pantic and I. Patras. Dynamics of Facial Expressions – Recognition of Facial Actions and their Temporal Segments from Face Profile Image

Sequences. *IEEE Trans. Systems, Man, and Cybernetics, Part B,* 36(2): 433-449, Apr. 2006.

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

[Pantic et al., 2005] M. Pantic, M.F. Valstar, R. Rademaker and L. Maat. Web-based database for facial expression analysis. In *Proc. Int'l Conf. Multimedia and Expo*, pp. 317-321, 2005. (www.mmifacedb.com)

[Pantic et al., 2006] M. Pantic, A. Pentland, A. Nijholt and T. Huang. Human Computing and Machine Understanding of Human Behavior: A Survey. In *Proc. Int'l Conf. Multimodal Interfaces*, pp. 239-248, 2006.

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

[Russell and Fernandez-Dols, 1997] J.A. Russell and J.M. Fernandez-Dols, Eds. *The psychology of facial expression*. Cambridge University Press, Cambridge, UK, 1997.

[Russell et al., 2003] J.A. Russell, J.A. Bachorowski and J.M. Fernandez-Dols. Facial and Vocal Expressions of Emotion. *Annual Review of Psychology,* 54: 329-349, 2003.

[Ruttkay et al., 2006] Z.M. Ruttkay, D. Reidsma and A. Nijholt. Human computing, virtual humans and artificial imperfection. In *Proc. Int'l Conf. Multimodal Interfaces*, pp. 179-184, 2006.

[Sand and Teller, 2006] P. Sand and S. Teller. Particle Video: Long-Range Motion Estimation using Point Trajectories. In *Proc. Int'l Conf. Computer Vision and Pattern Recognition*, pp. 2195-2202, 2006.

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

[Sharma et al., 2003] R. Sharma, M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, A.M. Maceachren and K. Sengupta. Speech-gesture driven multimodal interfaces for crisis management. *Proceedings of the IEEE,* 91(9): 1327-1354, Sep. 2003.

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

[Starner, 2001] T. Starner. The Challenges of Wearable Computing. *IEEE Micro,* 21(4): 44-67, July-Aug. 2001.

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

[Stenger et al., 2006] B. Stenger, P.H.S. Torr and R. Cipolla. Model-based hand tracking using a hierarchical Bayesian filter. *IEEE Trans. Pattern Analysis and Machine Intelligence,* 28(9): 1372-1384, Sep. 2006.

[Streitz and Nixon, 2005] N. Streitz and P. Nixon. The Disappearing Computer. *ACM Communications,* 48(3): 33-35, Mar. 2005.

[Truong and van Leeuwen, 2005] K.P. Truong and D.A. van Leeuwen. Automatic detection of laughter. In Proc. *Interspeech Euro. Conf.*, pp. 485-488, 2005.

[Valstar and Pantic, 2006a] M.F. Valstar and M. Pantic. Biologically vs. logic inspired encoding of facial actions and emotions in video. In *Proc. Int'l Conf. on Multimedia and Expo*, 2006.

[Valstar and Pantic, 2006b] M.F. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proc. Int'l Conf. Computer Vision and Pattern Recognition, vol. 3*, p. 149, 2006.

[Valstar et al., 2006] M.F. Valstar, M. Pantic, Z. Ambdar and J.F. Cohn. Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. In *Proc. Int'l Conf. Multimodal Interfaces*, pp. 162-170, 2006.

[Viola and Jones, 2004] P. Viola and M.J. Jones. Robust real-time face detection. *Int'l J. Computer Vision,* 57(2): 137-154, May 2004.

[Wang and Singh, 2003] J.J. Wang and S. Singh. Video analysis of human dynamics – a survey. *Real Time Imaging,* 9(5): 321-346, Oct. 2003.

[Wang et al., 2003] L. Wang, W. Hu and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3): 585-601, Mar. 2003.

[Weiser, 1991] M. Weiser. The Computer for the Twenty-First Century. *Scientific American,* 265(3): 94-104, Sep. 1991.

[Yang et al., 2002] M.H. Yang, D.J. Kriegman and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence,* 24(1): 34-58, Jan. 2002.

[Zhai and Bellotti, 2005] S. Zhai and V. Bellotti. Sensing-Based Interaction. *ACM Trans. Computer-Human Interaction,* 12(1): 1-2, Jan. 2005.

[Zhao et al., 2003] W. Zhao, R. Chellappa, P.J. Phillips and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys,* 35(4): 399-458, Dec. 2003.

[Zeng et al., 2006] Z. Zeng, Y. Hu, G.I. Roisman, Y. Fu and T.S. Huang. Audio-visual Emotion Recognition in Adult Attachment Interview. In *Proc. Int'l Conf. Multimodal Interfaces*, pp. 139-145, 2006.

[BTT Survey, 2006] BTT Survey on Alternative Biometrics. *Biometric Technology Today,* 14(3): 9-11, Mar. 2006.

[MMUA, 2006] MMUA: http://mmua.cs.ucsb.edu/