Human-Centred Intelligent Human–Computer Interaction (HCI²): how far are we from attaining it?

Maja Pantic*

Computing Department,
Imperial College London,
London SW7 2AZ, UK
and
EEMCS,
University of Twente,
Enschede 7500 AE, The Netherlands
E-mail: m.pantic@imperial.ac.uk
*Corresponding author

Anton Nijholt

EEMCS,
University of Twente,
Enschede 7500 AE, The Netherlands
E-mail: a.nijholt@ewi.utwente.nl

Alex Pentland

Media Lab,
Massachusetts Institute of Technology,
Cambridge, MA 02139 4307, USA
E-mail: pentland@media.mit.edu

Thomas S. Huanag

University of Illinois at Urbana-Champaign,
Beckman Institute,
Urbana, IL 61801, USA
E-mail: 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. To realise this prediction, next-generation computing should develop anticipatory user interfaces that are human-centred, built for humans and based on naturally occurring multimodal human communication. These interfaces should transcend the traditional keyboard and mouse and have the capacity to understand and emulate human communicative intentions as expressed through behavioural cues, such as affective and social signals. This article discusses how far we are to the goal of human-centred computing and Human-Centred Intelligent Human-Computer Interaction (HCI^2) that can understand and respond to multimodal human communication.

Keywords: context sensing; human affect recognition; introductory and survey; sensing humans; social signals detection.

Reference to this paper should be made as follows: Pantic, M., Nijholt, A., Pentland, A. and Huanag, T.S. (2008) 'Human-Centred Intelligent Human–Computer Interaction (HCI²): how far are we from attaining it?', *Int. J. Autonomous and Adaptive Communications Systems*, Vol. 1, No. 2, pp.168–187.

Biographical notes: Maja Pantic received the MSc degree and PhD in Computer Science from the Delft University of Technology, The Netherlands, in 1997 and 2001. She is a Reader in Multimodal HCI at Imperial College London, Computing Department and a Professor in Affective and Behavioural Computing at the University of Twente, Computer Science Department. Her research interests include computer vision and machine learning applied to face and body gesture recognition, human communicative behaviour analysis, multimodal HCI, affective computing and e-learning tools. She published more than 70 research papers on these topics. She is an IEEE Senior member. She is an Associate Editor of IEEE Trans. on Systems, Man and Cybernetics – Part B, and of Image and Vision Computing Journal. She is a guest Editor, Organiser and Committee member of over 10 major journals and conferences in the field.

Anton Nijholt received his MS degree in Mathematics and Computer Science from the Delft University of Technology and his PhD from the Vrije Universiteit of Amsterdam, the Netherlands. He held positions at various universities in the Netherlands, Belgium and Canada. Currently, he is a Full Professor and the Chair of the Human Media Interaction group of the University of Twente's Department of Computer Science. His main research interests are multiparty and multimodal interaction, virtual environments and social and intelligent (embodied) agents. He published more than 200 research papers on these topics. Except of several large Dutch national projects, he is currently involved in the research of the FP6 EC projects AMI and AMIDA (on Augmented Multi-party Interaction) and the FP6 EC NoE HUMAINE (the role of affect in the interface).

Alex Pentland is the Toshiba Professor of Media Arts and Sciences at the Massachusetts Institute of Technology (MIT) and Director of Human Dynamics Research. He was the Founding Director of the Media Lab Asia and the academic head of the MIT Media Laboratory. His research interests include wearable computers, health systems, smart environments and technologies for developing countries. He is a Pioneer in organisational engineering, mobile information systems and computational social science. His focus is the development of human-centred technology and the creation of ventures that take this technology into the real world. For his work, he has won numerous international awards in the arts, sciences and engineering.

Thomas S. Huang received his ScD from the Massachusetts Institute of Technology in Electrical Engineering, and was on the faculty of MIT and Purdue University. He joined University of Illinois at Urbana-Champaign in 1980 and is currently William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor of Coordinated Science Laboratory and Co-Chair of the Human Computer Intelligent Interactive major research theme of the Beckman Institute. He is a Member of the National Academy of Engineering and has received numerous honors and awards, including the IEEE Jack S. Kilby Signal Processing Medal and King-Sun Fu Prize of the International Association of Pattern Recognition. He has published 21 books and more than 600 papers in network theory, digital holograph, image and video compression, multimodal human computer Interfaces and multimedia databases.

1 Introduction

We entered an era of enhanced digital connectivity. Computers and internet have become so embedded in the daily fabric of people's lives that they simply cannot live without them. We use this technology to work, to communicate, to shop, to seek out new information and to entertain ourselves. These processes shift human activity away from real physical objects, emphasising virtual over physical environments.

It is widely believed that the next shift in computing technology will be embedding computers into our homes, transportation means and working spaces, emphasising once again physical environments. In this vision of the future, often referred to as ambient intelligence (Aarts, 2005) humans will be surrounded by arrays of intelligent, yet invisible computing devices that can anticipate every their need. Chairs and tables will be equipped with sensors and devices that can inform us if our sitting position can cause lower back pain, cars will pull over or sound an alarm if the driver becomes drowsy, and lights will be dimmed and our favourite background music will play when we come home showing signs of weariness.

Although profoundly appealing, this vision of the digital future creates a set of novel, greatly challenging issues concerning the interaction between the technology and humans as discussed by e.g. Nijholt and Traum (2005), Streitz and Nixon (2005) and Pantic et al. (2007). 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?

Human Computer Interaction (HCI) designs were first dominated by direct manipulation and then delegation. As observed by several researchers (e.g. Oviatt, 2003; Maat and Pantic, 2007) both styles of interaction involve usually conventional interface devices such as 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 such as 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. Clearly, 'business as usual' will not work in this case. We must approach HCI in a different way, moving away from computer-centred designs toward human-centred designs for HCI, made for humans, and based on naturally occurring human interactive behaviour. More specifically, Human-Centred Intelligent HCI (HCI^2) must have the ability to detect subtleties of and changes in the user's communicative behaviour (as expressed through, e.g. affective and social signals), and to initiate interactions based on this information, rather than simply responding to the user's commands. Sensing and understanding human communicative intentions including affective and social signals is a challenging task, however. How far are we from attaining it?

Figure 1 Behavioural cues such as facial expressions, gestures, vocalisations, etc. convey communicative intentions such as cognitive and affective states, emblems, manipulators, illustrators and regulators (see online version for colours)

2 Human behaviour perception: challenges

Machine analysis of human communicative behaviour is inherently a multi-disciplinary enterprise involving different research fields including psychology, linguistics, computer vision, signal processing and machine learning. There is no doubt that the progress in machine understanding of human interactive behaviour is contingent on the progress in the research in each of those fields (Ekman et al., 1992). The main scientific and engineering challenges related to realisation of machine sensing and understanding of human communicative intentions such as affective and social signals can be summarised as follows:

- Which type of communicative intention is expressed through displayed behavioural cues such as body postures, vocal and facial expressions (e.g. linguistic message, non-linguistic interactive cue, affect, attitude and mood)?
- Which human behavioural cues convey information about human communicative intentions such as social and emotional signals and, in turn, which modalities should be considered when building an automatic analyser of human communicative behaviours?
- What to take into account in order to understand shown behavioural cues (e.g. is the context important, such as the person's identity, current task, etc.) and in turn, how to discern between different types of communicative intentions (e.g. emotions vs. social signals).

2.1 Types of communicative intentions

The term behavioural signal (cue) 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). As summarised by Ekman and Friesen (1969) among the types of messages (communicative intentions) conveyed by behavioural cues are the following (Figure 1):

- affective/attitudinal/cognitive states (e.g. fear, joy, stress, disagreement, ambivalence and inattention)
- emblems (culture-specific interactive signals like wink or thumbs up)

- manipulators (actions used to act on objects in the environment or self-manipulative actions such as lip biting and scratching)
- 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 an agreement across different theories that at least some behavioural signals evolved to communicate information, there is a lack of consensus regarding their specificity, extent of their innateness and universality, and whether they convey emotions, social motives, behavioural intentions or all three. For detailed discussions of these topics, readers are referred to Russell and Fernandez-Dols (1997) and Lewis and Haviland-Jones (2000).

Arguably the most often debated issue is whether affective states are a separate type of messages communicated by behavioural signals (i.e. whether behavioural signals communicate actually felt emotions), or is the related behavioural signal (e.g. facial expression) just an illustrator/regulator aimed at controlling 'the trajectory of a given social interaction', as suggested by Fridlund (Russell and Fernandez-Dols, 1997). Explanations of human behavioural 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; Figure 2) that are universally displayed and recognised from non-verbal behavioural signals, especially facial and vocal expression, as suggested by Ekman, Scherer and others (Lewis and Haviland-Jones, 2000). Instead of explanations of human behavioural signals in terms of internal states, ethologists focus on consequences of behavioural 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 labelled in terms of emotions but in terms of behavioural ecology interpretations, which explain the influence a certain expression has in a particular context. Thus, an 'angry' face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. However, as proposed by Izard, one may feel angry without the slightest intention of attacking anyone (Russell and Fernandez-Dols, 1997).

Figure 2 Prototypic facial expressions of six basic emotions (disgust, happiness, sadness, anger, fear and surprise)

In summary, is social communication the sole function of behavioural signals? Do they never represent visible manifestation of emotion/feeling/affective states? Since in some instances (e.g. arachnophobia, acrophobia, object-elicited disgust and 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 behavioural signals. However, it is not only discrete emotions like surprise or anger that represent the affective states conveyed by human behavioural signals, as suggested by Ekman and Friesen (1969). Behavioural cues identifying attitudinal states like interest and boredom, via those underlying moods, to those representing cognitive states like agreement and disagreement, and to those disclosing social signals like empathy and antipathy are essential components of human behaviour as well. Hence, in contrast to traditional approach of Ekman and Friesen (1969) who list only basic emotions as the first type of messages conveyed by behavioural cues, we suggest that for the aims of HCI² technology, affective states should be treated as being correlated not only to basic emotions, but also to more complex mental states like depression or pain as well as to the aforementioned attitudinal states and social signals.

2.2 Relevant modalities

We speak, move, gesture, shift our gaze in an effective flow of communication. But which of these interactive cues convey information about human behaviours like affective and social signals? From the types of messages conveyed by behavioural signals, manipulators are usually associated with self-manipulative gestures like scratching or lip biting and involve facial expressions and body gestures. Emblems, illustrators and regulators are typical social signals, spoken and wordless messages such as 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 behavioural signals are affective and attitudinal states. Affective arousal modulates all human communicative signals (Lewis and Haviland-Jones, 2000). Hence, one could expect that automated analysers of human behaviour should include all human interactive modalities (audio, visual and tactile) and should analyse 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.

It seems that not all behavioural cues are equally important in the human interpretation of the communicative intention. For instance, although the research in psycho-physiology has produced firm evidence that affective arousal has a range of somatic and physiological correlates including heart rate, skin clamminess, temperature and respiration velocity (Lewis and Haviland-Jones, 2000) 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 recent advent of wearable computers, which promises robust physiological sensing, opens up possibilities for including tactile modality into automatic analysers of human behaviour (Pentland, 2005).

Similarly, although speech has become the indispensable means for sharing ideas, spoken messages do not represent a reliable means to analyse and predict human behaviour. Let us explain this issue in more detail. Speech conveys affective information

through explicit (linguistic) and implicit (paralinguistic) messages that reflect the way the words are spoken. As the linguistic content is concerned, some information about the speaker's affective state can be inferred directly from the surface features of words, which were summarised in some affective word dictionaries and lexical affinity (e.g. Whissell, 1989) and the rest of affective information lies below the text surface and can only be detected when the semantic context (e.g. discourse information) is taken into account. However, findings in basic research like those reported by Furnas et al. (1987) and Ambady and Rosenthal (1992) indicate that linguistic messages are rather unreliable means to analyse human (affective) behaviour, and it is very difficult to anticipate a person's word choice and the associated intent in affective expressions. In addition, the association between linguistic content and emotion is language-dependent and generalising from one language to another is very difficult to achieve.

When it comes to implicit, paralinguistic messages that convey affective information, basic researchers have not identified an optimal set of voice cues that reliably discriminate among emotions. Nonetheless, listeners seem to be accurate in decoding some basic emotions from prosody (Juslin and Scherer, 2005) as well as some non-basic affective states such as distress, anxiety, boredom and sexual interest from non-linguistic vocalisations, such as laughs, cries, sighs and yawns (Russell and Fernandez-Dols, 1997; Russell, Bachorowski and Fernandez-Dols, 2003). For a comprehensive summary of acoustic cues related to vocal expressions of basic emotions, readers are referred to Cowie et al. (2001).

It seems that the visual channel carrying facial expressions and body gestures is the most important in the human judgment of behavioural cues. As indicated by numerous researchers, the human face is our pre-eminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression (Lewis and Haviland-Jones, 2000; Ekman and Rosenberg, 2005). Basic research also provides evidence that observers tend to be accurate in decoding some negative basic emotions such as anger and sadness from static body postures (e.g. Coulson, 2004) and that gestures such as head inclination, face touching and shifting posture often accompany social affective states such as shame and embarrassment (e.g. Costa et al., 2001). However, as shown by 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. According to this study, to interpret someone's behavioural cues, people usually rely on shown facial expressions, to a lesser degree on shown body gestures, and to an even lesser degree on displayed vocal expressions. However, note that the relative contributions of body gestures, facial and vocal expressions to affect judgment depend on the affective state and the environment where the affective behaviour occurs (Ekman, 1982).

In a summary, a large number of studies in psychology and linguistics confirm the correlation between some affective displays (especially prototypical emotions) and specific audio and visual signals (e.g. Lewis and Haviland-Jones, 2000). The human judgment agreement is typically higher for facial expression modality than it is for vocal expression modality. However, the amount of the agreement drops considerably when the stimuli are spontaneously displayed expressions of affective behaviour rather than posed exaggerated displays. In addition, facial expression, body postures/gestures and vocal expression of emotion are often studied separately. This precludes finding evidence of the temporal correlation between them. On the other hand, numerous studies have theoretically and empirically demonstrated the advantage of integration of multiple

modalities (at least audio and visual) in human affect perception over single modalities (e.g. Russell, Bachorowski and Fernandez-Dols, 2003).

Finally, a growing body of research in cognitive sciences argues that the dynamics of human behaviour are crucial for its interpretation (e.g. Russell, Bachorowski and Fernandez-Dols, 2003; Ekman and Rosenberg, 2005). For instance, it has been shown that temporal dynamics of facial behaviour represents a critical factor for distinction between spontaneous and posed facial behaviour as well as for categorisation of complex behaviours such as pain, shame and amusement (e.g. Cohn and Schmidt, 2004; Ekman and Rosenberg, 2005). Based on these findings, we may expect that temporal dynamics of each modality separately and temporal correlations between the modalities play an important role in interpretation of human naturalistic, audiovisual affective behaviour. However, these are virtually unexplored areas of research.

2.3 Context-sensitive interpretation

A smile can be a display of politeness (social signal), joy (affective state), irony/irritation (affective state), empathy (emotional response/social signal), greeting (social signal), etc. In other words, behavioural cues do not usually convey exclusively one type of communicative intention, but may convey any of the types listed above. For instance, a frown may be a sign of short-sightedness if this action is a reflex (a manipulator), a sign of anger/dislike if this action is displayed unintentionally when seeing someone passing by (affective cue), a sign of posed anger if this action is displayed deliberately as a response on friendly teasing (illustrator), or a sign of rapt attention and understanding if this action occurs during a conversation (regulator), to mention just a few possibilities. From this example, it is obvious that in order to determine the communicative intention conveyed by an observed behavioural 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, when the signal has been displayed (i.e. what is the timing of displayed behavioural signals with respect to changes in the environment), and who the expresser is (i.e. it is not probable that each of us will express a particular affective state by modulating the same communicative signals in the same way).

However, note that while W4 (where, what, when and who) methodology is dealing only with the apparent perceptual aspect of the context in which HCI takes place, human-centred computing is about W5 + (where, what, when, who, why and how) methodology, where the why and how are directly related to recognising communicative intention including social signals, affective and cognitive states of the user. Hence, W5+ designs for HCI will yield the transition from HCI to HCI², where people and computers with embodied/embedded cognition can augment each other's capabilities and display collaborative team behaviour. However, since the problem of context-sensing is extremely difficult to solve, especially for a general case (i.e. general-purpose W4 technology does not exist yet; see also Section 3.2), answering the why and how questions in a W4-context-sensitive manner is virtually unexplored area of research. In turn, we also need to recognise the likelihood that W5 + designs for HCI and human-centred computing, in general, still linger in the relatively distant future.

3 Machine analysis of human behaviour: the state of the art

Modelling human behaviour and understanding displayed patterns of behavioural signals, involve a number of tasks.

- Sensing and analysing displayed behavioural cues including facial expressions, body gestures, non-linguistic vocalisations and vocal intonations.
- Sensing the perceptual aspects of the context in which the observed behavioural cues were displayed (W4-context sensing).
- Understanding the observed behaviour by translating the sensed behavioural and context cues into a description of communicative intentions (W5+-context sensing).

3.1 Sensing human behavioural cues methodology

Computer vision technology applied to detect, track and recognise human behavioural signals such as face, hand and body gestures has some notable success to date.

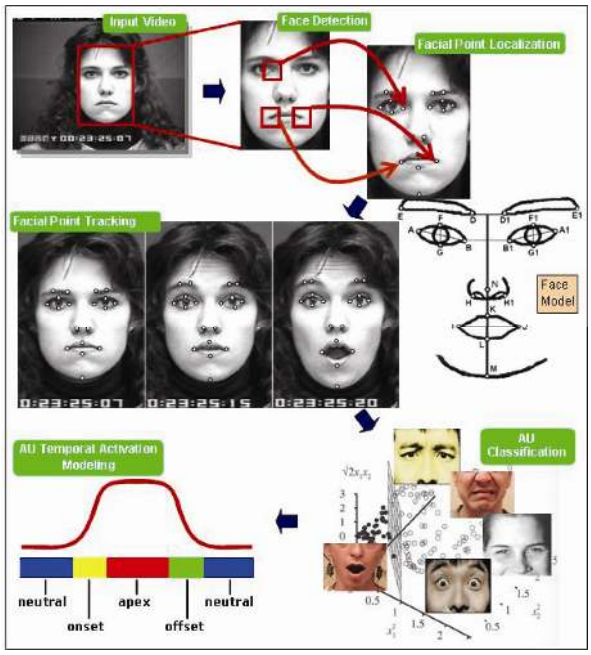
Because of their relevance to face recognition and, in turn, security, face detection, tracking and facial expression analysis attracted the interest of many researchers. Numerous techniques have been developed for face detection, i.e. identification of all regions in the scene that contain a human face. For surveys of the past efforts in the field, the readers are referred to Yang, Kriegman and Ahuja (2002), Li and Jain (2005) and Sato et al. (2005). Most of these methods emphasise statistical learning techniques and use appearance (skin texture based) features. However, virtually all of them can detect only (near-) upright faces in (near-) frontal view.

Head, face and facial feature tracking are essential steps for human motion analysis, since they provide data for recognition of face/head pose and facial expression. Optical flow has been widely used for head, face and facial feature tracking (Pantic and Bartlett, 2007). 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 such as Kalman and particle filtering schemes (Haykin and de Freitas, 2004). Some of the most advanced approaches to head tracking and head-pose estimation are based on Kalman and particle filtering frameworks (e.g. Smith et al., 2008). Similarly, the most advanced approaches to facial feature tracking are based on Kalman and particle filtering tracking schemes (Pantic and Bartlett, 2007). 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.

Most of the facial expressions analysers developed so far attempt to recognise a small set of prototypic emotional facial expressions such as happiness or sadness displayed on command. For comprehensive surveys of the past efforts in the field, readers are referred to Pantic and Rothkrantz (2003), Tian et al. (2005) and Zeng et al. (2008). To facilitate detection of subtle facial signals like a frown or a smile and to make facial expression information available for usage in HCI² applications like face-based assessment of consumer's satisfaction, several research groups begun research on machine analysis of facial muscle actions (i.e., atomic facial cues, Action Units (AUs)) defined by Ekman and

colleagues (2002). The focus of the research efforts in the field was first on automatic recognition of AUs in either static face images or face image sequences picturing facial expressions produced on command. Several promising prototype systems were reported that can recognise few deliberately produced AUs in either (near-) frontal view face images (see Tian et al., 2005) or profile view face images (Pantic and Patras, 2006). These systems employ different approaches including expert rules and machine learning methods such as neural networks, and use either feature-based image representations (i.e. use geometric features like facial points; e.g. see Figure 3) or appearance-based image representations (i.e. use texture of the facial skin including wrinkles and furrows). One of the main criticisms that these works received, is that the methods are not applicable in real-life situations, where subtle changes in facial expression typify the displayed facial behaviour rather than the exaggerated changes that typify posed expressions. Hence, the focus of the research in the field started to shift to automatic (non-basic-) emotion and AU recognition in spontaneous facial expressions (produced in a reflex-like manner). Several works have recently emerged on machine analysis of non-basic emotions or AUs in spontaneous facial expression data. For comprehensive overviews of the efforts in the field, see Pantic and Bartlett (2007) and Zeng et al. (2008). These methods employ probabilistic, statistical and ensemble learning techniques, which seem to be particularly suitable for automatic facial expression analysis in video recordings. However, the present systems for facial expression analysis typically depend upon accurate head, face and facial feature tracking as input and are still very limited in performance and robustness.

Figure 3 Outline of a geometric-feature-based system for detection of facial AUs and their temporal phases (onset, apex, offset and neutral) (see online version for colours)



Source: Valstar and Pantic (2006)

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. For exhaustive surveys of the past efforts in the field, readers are referred to Wang and Singh (2003), Ong and Ranganath (2005), Mitra and Acharya (2007) and Poppe (2007). Most of the proposed techniques are either model-based (i.e. use geometric primitives such as cones and spheres to model head, trunk, limbs and fingers) or appearance-based (i.e. use colour or texture information to track the body and its parts). Most of these methods emphasise Gaussian models, probabilistic learning and particle filtering framework (e.g. Bray, Koller-Meier and van Gool, 2007). 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 behaviour, the hands do not have to be always visible (they may be 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), the non-linguistic part of it (how it has been said) carries important information about the speaker's affective state and attitude (Russell, Bachorowski and Fernandez-Dols, 2003; Juslin and Scherer, 2005). This finding instigated the research on automatic analysis of vocal non-linguistic expressions. The vast majority of present work is aimed at basic emotion recognition from prosodic features (e.g. pitch, intensity and speech rate) and spectral features (e.g. Mel Frequency Cepstral Coefficients (MFCC) and cepstral features). For comprehensive surveys in the field, the readers are referred to Section 3.3 of this article, and to Cowie et al. (2001), Pantic and Rothkrantz (2003) and Zeng et al. (2008). More recently, few efforts towards automatic recognition of non-linguistic vocal outbursts such as laughs, cries and coughs have been also reported (Pantic et al., 2007). Most of these efforts are based only on audio signals (e.g. Truong and van Leeuwen, 2007). However, since it has been shown by several experimental studies in either psychology or signal processing that integrating the information from audio and video leads to an improved performance of human behaviour recognition (e.g. Russell, Bachorowski and Fernandez-Dols, 2003) few pioneering efforts towards audiovisual recognition of non-linguistic vocal outbursts have been recently reported including audiovisual analysis of infants' cries proposed by Pal, Iyer and Yantorno (2006) and audiovisual laughter recognition proposed by Petridis and Pantic (2008). Since, the research in cognitive sciences provided some promising hints that vocal outbursts and non-linguistic vocalisations such as yelling, laughing and sobbing, may be very important cues for decoding someone's affect/attitude (Russell, Bachorowski and Fernandez-Dols, 2003), we suggest a much broader focus on machine recognition of these non-linguistic vocal cues.

3.2 W4 methodology

Context plays a crucial role in understanding of human behavioural signals, since they are easily misinterpreted if the information about the situation in which the shown behavioural cues have been displayed is not taken into account. Hence, the so-called W4 (who, where, what and when) technology is essential for interpreting human behaviour.

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 selling biometric systems (Pantic et al., 2007). The problem of face recognition has been tackled most often. For comprehensive surveys of past works in the field, see Zhao et al. (2003) and Bowyer, Chang and Flynn (2006). However, due to their reliability and robustness, multimodal biometric systems based on multiple biometric traits including face, voice, iris, retina, fingerprints, gait, ear, hand, brainwaves and facial thermogram, have recently become a research trend (Bowyer et al., 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. Most of the current approaches base their analysis on scene (background) modelling, motion segmentation, object classification and object tracking (Wang and Singh, 2003; Poppe, 2007). In turn, these present methods are adequate when a priori knowledge is available (e.g. scene model, human-silhouette-based shape to be tracked), but they are weak for unconstrained environments (e.g. gym and a house party), in which dynamic scene changes, 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; e.g. see Elgammal and Lee, 2007) and use unsupervised learning (e.g. Bicego, Cristani and Murino, 2006) will represent a better solution.

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 such as keystrokes, mouse movements and active software (e.g. Maat and Pantic, 2007). 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 ubiquitous, anticipatory, HCI² interfaces, where the tasks are often ill-defined due to uncertainty in the sensed environmental and behavioural cues. Analysis of tasks that human may carry out in the context of HCI² require adaptation and fusion of existing methods for behavioural 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 (Pantic et al., 2007). Current methods for human activity recognition typically identify the task of the observed person in an implicit manner, by recognising 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 such as watching TV, dining and working with desktop computer, may be defined).

As we have already mentioned above, temporal dynamics of behavioural cues (i.e. their timing, co-occurrence, speed, etc.) are crucial for the interpretation of the observed behaviour (Ekman and Rosenberg, 2005). However, present methods for human activity/behaviour recognition do not address the when context question – dynamics of displayed behavioural signals is usually not taken into account when analysing the

observed behaviour, let alone analysing the timing of displayed behavioural signals with respect to changes in the environment. Exceptions of this rule include few recent studies such as that by Tong, Liao and Ji (2007) on modelling semantic and temporal relationships between AUs forming a facial expression, that by Valstar, Gunes and Pantic (2007) on discrimination between spontaneous and posed smiles based on temporal dynamics of AUs, head and shoulder gestures (2007), and few studies on multimodal analysis of audio and visual dynamic behaviours for emotion recognition (Zeng et al., 2008). In general, present methods cannot handle longer time scales, model grammars of users' behaviours, and take temporal and context-dependent evolvement of observations into account for more robust performance. When it comes to the timing of shown behavioural signals with respect to changes in the environment, current methods typically approach the when question in an implicit way, by recognising user's reactions to changes in the environment as different activities.

Overall, context questions are usually addressed separately and often in an implicit manner. Yet, they may be more reliably answered if they are answered in groups of two or three using the information extracted from multimodal input streams. For example, as shown by Nock, Iyengar and Neti (2004) 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. They suggested further that a key to successful realisation of multimodal multi-aspect context-sensing is to automatically determine whether observed behavioural cues share a common cause – e.g. whether the mouth movements and audio signals complement to indicate an active known or unknown speaker (who and where) and whether his or her focus of attention is another person or a computer (what). The main advantages of such an approach are effective handling of uncertainties due to noise in input data streams and the problem-dimensionality reduction. Therefore, we suggest a much broader focus on spatial and temporal, multimodal multi-aspect context-sensing.

3.3 W5+ methodology

The past work in translating the sensed human behavioural signals and context descriptors into a description of the shown behaviour, that is, answering the question why and how in the context of human-centred computing and W5 + (where, what, when, who, why and how) methodology, can be roughly divided into the methods for understanding human affective/attitudinal states and those for understanding human social signals (i.e. emblems, regulators and illustrators).

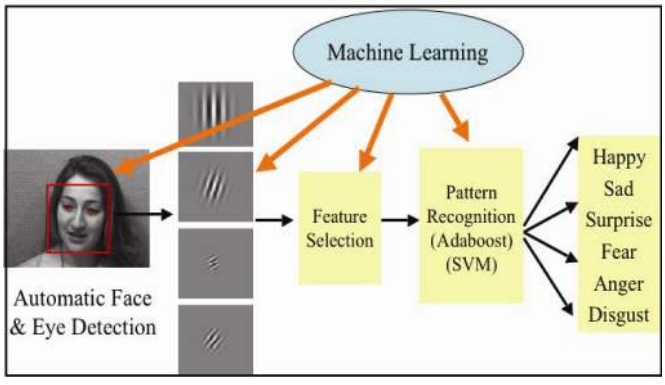
The existing body of literature in machine analysis of human affect is immense. For exhaustive reviews of the past work in the field, the readers are referred to Cowie et al. (2001), Pantic and Rothkrantz (2003) and Zeng et al. (2008). Most of these works attempt to recognise a small set of prototypic expressions of basic emotions such as happiness and anger from either face images/video or speech signal (e.g. Figure 4). They achieve an accuracy of 64–98% 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:

- Only a small set of deliberately displayed prototypic facial or vocal expressions of six basic emotions can be handled.
- Context-sensitive analysis (user-, or environment-, or task-dependent analysis) of the sensed signals cannot be performed.

- Extracted facial or vocal expression information cannot be analysed 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 analysers.
- Strong assumptions are usually adopted. For example, facial affect analysers 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 analysers 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.

The main criticism that these works receive is that the methods are not applicable in real-life situations, where the displayed behaviour is typified by subtle rather than exaggerated changes in facial and vocal expressions and other behavioural cues. Hence, the focus of the research in the field started to shift to automatic (non-basic-) emotion recognition in recordings of spontaneous human behaviour (produced in a reflex-like manner). Several efforts have been recently reported on a automatic analysis of volatile facial affect data (e.g. Littlewort, Bartlett and Lee, 2007), few studies investigated automatic, vision-based discrimination between spontaneous and deliberate affective behaviour (e.g. Valstar, Gunes and Pantic, 2007), several efforts have been reported on automatic emotion analysis from spontaneous vocal affect data (e.g. Neiberg, Elenius and Laskowski, 2006), and few studies have been reported on audiovisual analysis of spontaneously produced affect data (e.g. Fragopanagos and Taylor, 2005). For a comprehensive overview of the current efforts in the field, see Zeng et al. (2008). However, many improvements are needed if these systems are to be used for context-sensitive analysis of subtle (unexaggerated) human behavioural 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. Of these, we would like to stress the importance of two issues: using information of language and achieving temporal multimodal data fusion.

Figure 4 Outline of the facial affect recognition system (see online version for colours)



Source: Littlewort et al. (2006)

The importance of the former became obvious with the research shift towards analysis of spontaneous human behaviour – analysis of acoustic information only, does not suffice for identifying subtle changes in vocal expression. In turn, several recent studies investigated the combination of acoustic features and linguistic features (language and discourse) to improve recognition of emotions from speech signal (e.g. Fragopanagos and Taylor, 2005). However, these systems typically depend upon both accurate recognition of verbal content of emotional speech, which still cannot be reliably achieved by existing automatic speech recognition systems, and on accurate extraction of semantic discourse information, which is attained manually in the present systems. Hence, we suggest a much broader focus on solving these issues which will enable the use of linguistic features for attaining more accurate and robust vocal affect analysis.

Most of the present audiovisual and multimodal systems in the field perform decision-level data fusion (i.e. classifier fusion) in which the input coming from each modality is modelled independently and these single-modal recognition results are combined at the end. Since humans display audio and visual expressions in a complementary and redundant manner, the assumption of conditional independence between audio and visual data streams in decision-level fusion is incorrect and results in the loss of information of mutual correlation between the two modalities. To address this problem, a number of model-level fusion methods have been proposed that aim at making use of the correlation between audio and visual data streams, and relax the requirement of synchronisation of these streams (e.g. Fragopanagos and Taylor, 2005; Zeng et al., 2006). However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is largely unexplored. Hence, we suggest a much broader focus on the issues relevant to multimodal temporal fusion including the optimal level of integrating these different streams, the optimal function for the integration, how estimations of reliability of each stream can be included in the inference process. In addition, how to build context-dependent multimodal fusion is another open and highly relevant issue.

Messages conveyed by behavioural cues such as emblems, illustrators and regulators, can be interpreted in terms of social signals such as turn taking, mirroring, empathy, interest, engagement, agreement, disagreement, etc. Although each one of us understands the importance of social signals in everyday life situations, and although a firm body of literature in cognitive sciences exists on the topic (e.g. Chartrand and Bargh, 1999) and in spite of recent advances in sensing and analysing behavioural cues such as blinks, smiles, winks, thumbs up, yawns, laughter, etc. the research efforts in machine analysis of human social signals are few and tentative. Important works in the field include efforts to discern social signals such as activity level, stress, engagement and mirroring by analysing the engaged persons' tone of voice, efforts towards analysis of interest, agreement and disagreement from facial and head movements, and efforts towards analysis of the level of interest from tone of voice, head and hand movements. For an overview of efforts in the field, see Pantic et al. (2007). Overall, present approaches to understand social signals are multimodal and based on either statistical or probabilistic reasoning methods (e.g. Schuller et al., 2007). However, most of these methods are context insensitive (W4 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 such as status, interest, determination and

cooperation, which may be an invaluable asset in the development of HCI² and social networks formed of humans and computers, in their current form, they are not appropriate for general-purpose anticipatory interfaces.

4 How to attain HCI²: guidelines

Human behavioural actions or simply human behaviour, are high-level semantic events, which typically include interactions with the environment and causal relationships. An important distinction between the analysis of these high-level semantic events and the analysis of low-level semantic events like the occurrence of an individual behavioural cue like the blink, 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 behaviour are neither multimodal, nor context-sensitive, nor suitable for handling longer time scales. Hence, the focus of future research efforts in the field should be primarily on tackling the problem of context-constrained analysis of multimodal behavioural signals shown in temporal intervals. As suggested throughout the text, this problem should be treated as one complex problem rather than a number of detached problems in human sensing, context sensing and human behaviour 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.

- Modalities: which behavioural channels such as the face, the body and the tone of the voice, are minimally needed for realisation of robust and accurate human behaviour analysis? Does this hold independently of the target communicative intention to be recognised? No comprehensive study on the topic is available yet.
- Fusion: how to model temporal multimodal fusion which will take into account temporal correlations within and between different modalities? What is the optimal level of integrating these different streams? Does this depend upon the time scale at which the fusion is achieved? What is the optimal function for the integration?
- Fusion and context: do context-dependent fusion of modalities and discordance handling, which are typical for fusion of sensory neurons in humans, pertain in machine context sensing? Note that context-dependent fusion and discordance handling were never attempted within an automated system.
- Learning vs. education: what are the relevant parameters in shown human behaviour 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 realising this goal.

- Technical aspects: most methods for human sensing, context sensing and human behaviour understanding work only in (often highly) constrained environments. Noise, fast movements, changes in illumination, etc. cause them to fail. Also, many of the methods in the field do not perform fast enough to support interactivity. Researchers usually choose for more sophisticated processing rather than for real-time processing. The aim of future efforts in the field should be the realisation of more robust, real-time systems, if they are to be deployed in anticipatory interfaces defused throughout smart environments of the future.

In summary, although the research in sensing and understanding human communicative intentions including affective and social signals has witnessed a good deal of progress in recent years, there remain significant scientific and technical challenges to be addressed. However, we are optimistic about the future progress in the field. The main reason is that W5+ methodology and HCI² technology are likely to become the single most widespread research topic of AI (if not of the whole computing) community. This is aided and abetted by large and steadily growing number of research projects concerned with the interpretation of human behaviour at a deeper level for the purposes of ambient intelligence applications, independent living and personal wellness technologies, educational tools, etc. (e.g. EC FP6 AMI and AMIDA, EC FP7 CoFRIEND, LIREC, PROMETHEUS, SEMAINE and CHRIS).

Acknowledgements

The work of Maja Pantic and Anton Nijholt has been funded in part by the EU IST Programme Project FP6-0027787 (AMIDA) and the EC's 7th Framework Programme [FP7/2007–2013] under grant agreement no 211486 (SEMAINE).

References

Aarts, E. (2005) 'Ambient intelligence drives open innovation', *ACM Interactions*, Vol. 12, pp.66–68.

Ambady, N. and Rosenthal, R. (1992) 'Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis', *Psychological Bulletin*, Vol. 111, pp.256–274.

Bicego, M., Cristani, M. and Murino, V. (2006) 'Unsupervised scene analysis: a hidden Markov model approach', *Computer Vision and Image Understanding*, Vol. 102, pp.22–41.

Bray, M., Koller-Meier, E. and van Gool, L. (2007) 'Smart particle filtering for high-dimensional tracking', *Computer Vision and Image Understanding*, Vol. 106, pp.116–129.

Bowyer, K.W., Chang, K. and Flynn, P.J. (2006) 'A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition', *Computer Vision and Image Understanding*, Vol. 101, pp.1–15.

Bowyer, K.W., Chang, K., Yan, P., Flynn, P.J., Hansley, E. and Sarkar, S. (2006) 'Multimodal biometrics: an overview', Paper presented in the Proceedings of the *Workshop on Multimodal User Authentication (MMUA'06)*. Available at: http://www.nd.edu/~kwb/publications.html.

Chartrand, T.L. and Bargh, J.A. (1999) 'The chameleon effect: the perception-behavior link and social interaction', *Journal of Personality and Social Psychology*, Vol. 76, pp.893–910.

Cohn, J.F. and Schmidt, K.L. (2004) 'The timing of facial motion in posed and spontaneous smiles', *Journal of Wavelets, Multi-resolution and Information Processing*, Vol. 2, pp.121–132.

Costa, M., Dinsbach, W., Manstead, A.S.R. and Bitti, P.E.R. (2001) 'Social presence, embarrassment, and nonverbal behavior', *Journal of Nonverbal Behavior*, Vol. 25, pp.225–240.

Coulson, M. (2004) 'Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence', *Journal of Nonverbal Behavior*, Vol. 28, pp.117–139.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J.G. (2001) 'Emotion recognition in human–computer interaction', *IEEE Signal Processing Magazine*, Vol. 18, pp.32–80.

Ekman, P. (Ed.) (1982) *Emotion in the Human Face*. New York, NY: Cambridge University Press.

Ekman, P. and Friesen, W.F. (1969) 'The repertoire of nonverbal behavioral categories – origins, usage, and coding', *Semiotica*, Vol. 1, pp.49–98.

Ekman, P., Friesen, W.V. and Hager, J.C. (2002) *Facial Action Coding System*. Salt Lake City, USA: A Human Face.

Ekman, P., Huang, T.S., Sejnowski, T.J. and Hager, J.C. (Eds) (1992) *NSF Understanding the Face*. Salt Lake City, USA: A Human Face eStore, (see Library).

Ekman, P. and Rosenberg, E. (Eds) (2005) *What the Face Reveals*. Oxford, UK: Oxford University Press.

Elgammal, A. and Lee, C-S. (2007) 'Nonlinear manifold learning for dynamic shape and dynamic appearance', *Computer Vision and Image Understanding*, Vol. 106, pp.31–46.

Fragopanagos, F. and Taylor, J.G. (2005) 'Emotion recognition in human–computer interaction', *Neural Networks*, Vol. 18, pp.389–405.

Furnas, G., Landauer, T., Gomes, L. and Dumais, S. (1987) 'The vocabulary problem in human–system communication', *Communications of the ACM*, Vol. 30, pp.964–972.

Haykin, S. and de Freitas, N. (Eds) (2004) 'Special issue on sequential state estimation', Paper presented in the Proceedings of the *IEEE*, Vol. 92, pp.399–574.

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

Lewis, M. and Haviland-Jones, J.M. (Eds) (2000) *Handbook of Emotions*. New York, NY: Guilford Press.

Li, S.Z. and Jain, A.K. (Eds) (2005) *Handbook of Face Recognition*. New York, NY: Springer.

Littlewort, G.C., Bartlett, M.S. and Lee, K. (2007) 'Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain', Paper presented in the Proceedings of the *ACM International Conference of Multimodal Interfaces*, pp.15–21.

Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J. and Movellan, J. (2006) 'Dynamics of facial expression extracted automatically from video', *Journal of Image and Vision Computing*, Vol. 24, pp.615–625.

Maat, L. and Pantic, M. (2007) 'Gaze-X: adaptive affective multimodal interface for single-user office scenarios', *Lecture Notes in Artificial Intelligence*, Vol. 4451, pp.251–271.

Mitra, S. and Acharya, T. (2007) 'Gesture recognition: a survey', *IEEE Transactions on Systems, Man and Cybernetics, Part C*, Vol. 37, pp.311–324.

Neiberg, D., Elenius, K. and Laskowski, K. (2006) 'Emotion recognition in spontaneous speech using GMM', Paper presented in the Proceedings of the *International Conference on Spoken Language Processing*, pp.809–812.

Nijholt, A. and Traum, D. (2005) 'The virtuality continuum revisited', Paper presented in the Proceedings of the *International Conference on Computer Human Interaction*, pp.2132–2133.

Nock, H.J., Iyengar, G. and Neti, C. (2004) 'Multimodal processing by finding common cause', *Communications of the ACM*, Vol. 47, pp.51–56.

Ong, S.C.W. and Ranganath, S. (2005) 'Automatic sign language analysis: a survey and the future beyond lexical meaning', *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 27, pp.873–891.

Oviatt, S. (2003) 'User-centered modeling and evaluation of multimodal interfaces', Paper presented in the Proceedings of the *IEEE*, Vol. 91, pp.1457–1468.

Pal, P., Iyer, A.N. and Yantorno, R.E. (2006) 'Emotion detection from infant facial expressions and cries', Paper presented in the Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.721–724.

Pantic, M. and Bartlett, M.S. (2007) 'Machine analysis of facial expressions', in K. Delac and M. Grgic (Eds), *Face Recognition* (pp.377–416). Vienna, Austria: I-Tech Education and Publishing.

Pantic, M. and Patras, I. (2006) 'Dynamics of facial expressions – recognition of facial actions and their temporal segments from face profile image sequences', *IEEE Transaction on Systems, Man, and Cybernetics, Part B*, Vol. 36, pp.433–449.

Pantic, M., Pentland, A., Nijholt, A. and Huang, T.S. (2007) 'Human computing and machine understanding of human behaviour: a survey', *Lecture Notes in Artificial Intelligence*, Vol. 4451, pp.47–71.

Pantic, M. and Rothkrantz, L.J.M. (2003) 'Toward an affect-sensitive multimodal human–computer interaction', Paper presented in the Proceedings of the *IEEE*, Vol. 91, pp.1370–1390.

Pentland, A. (2005) 'Socially aware computation and communication', *IEEE Computer*, Vol. 38, pp.33–40.

Petridis, S. and Pantic, M. (2008) 'Audiovisual discrimination between laughter and speech', Paper presented in the Proceedings of the *IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp.5117–5120.

Poppe, R. (2007) 'Vision-based human motion analysis: an overview', *Computer Vision and Image Understanding*, Vol. 108, pp.4–18.

Russell, J.A., Bachorowski, J. and Fernandez-Dols, J. (2003) 'Facial and vocal expressions of emotion', *Annual Review Of Psychology*, Vol. 54, pp.329–349.

Russell, J.A. and Fernandez-Dols, J.M. (Eds) (1997) *The Psychology of Facial Expression*. Cambridge, UK: Cambridge University Press.

Sato, A., Imaoka, H., Suzuki, T. and Hosoi, T. (2005) 'Advances in face detection and recognition technologies', *NEC Journal of Advanced Technology*, Vol. 2, pp.28–34.

Schuller, B., Mueller, R., Hoernler, B., Hoethker, A., Konosu, H. and Rigoll, G. (2007) 'Audiovisual recognition of spontaneous interest within conversations', Paper presented in the Proceedings of the *ACM International Conference on Multimodal Interfaces*, pp.30–37.

Smith, K., Ba, S., Odobez, J-M. and Gatica-Perez, D. (2008) 'Tracking the visual focus of attention for a varying number of wandering people', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30.

Streitz, N. and Nixon, P. (2005) 'The disappearing computer', *ACM Communications*, Vol. 48, pp.33–35.

Tian, Y.L., Kanade, T. and Cohn, J.F. (2005) 'Facial expression analysis', In S.Z. Li and A. K. Jain (Eds), *Handbook of Face Recognition*, New York, NY: Springer, pp.247–276.

Tong, Y., Liao, W. and Ji, Q. (2007) 'Facial action unit recognition by exploiting their dynamics and semantic relationships', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, pp.1683–1699.

Truong, K.P. and van Leeuwen, D.A. (2007) 'Automatic discrimination between laughter and speech', *Speech Communication*, Vol. 49, pp.144–158.

Valstar, M.F. and Pantic, M. (2006) 'Fully automatic facial action unit detection and temporal analysis', Paper presented in the Proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 3, p.149.

Valstar, M.F., Gunes, H. and Pantic, M. (2007) 'How to distinguish posed from spontaneous smiles using geometric features', Paper presented in the Proceedings of the *ACM International Conference On Multimodal Interfaces*, pp.38–45.

Wang, J.J. and Singh, S. (2003) 'Video analysis of human dynamics – a survey', *Real Time Imaging*, Vol. 9, pp.321–346.

Whissell, C.M. (1989) 'The dictionary of affect in language', in R. Plutchik and H. Kellerman (Eds), *Emotion – Theory, Research and Experience, Vol 4: The Measurement of Emotions* (pp.113–131). New York, NY: Academic Press.

Yang, M-H., Kriegman, D.J. and Ahuja, N. (2002) 'Detecting faces in images: a survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp.34–58.

Zeng, Z., Hu, Y., Liu, M., Fu, Y. and Huang, T.S. (2006) 'Training combination strategy of multi-stream fused hidden Markov model for audio–visual affect recognition', Paper presented in the Proceedings of the *ACM International Conference on Multimedia*, pp.65–68.

Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S. (2008) 'A survey of affect recognition methods: audio, visual and spontaneous expressions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30.

Zhao, W., Chellappa, R., Phillips, P.J. and Rosenfeld, A. (2003) 'Face recognition: a literature survey', *ACM Computing Surveys*, Vol. 35, pp.399–458.