Encyclopedia of Multimedia Technology and Networking

Second Edition

Margherita Pagani
Bocconi University, Italy

Volume I
A–Ev

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Senior Managing Editor:	Jennifer Neidig
Managing Editor:	Jamie Snavely
Assistant Managing Editor:	Carole Coulson
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of multimedia technology and networking / Margherita Pagani, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This publication offers a research compendium of human knowledge related to the emerging multimedia digital metamarket"--Provided by publisher.
ISBN 978-1-60566-014-1 (hardcover) -- ISBN 978-1-60566-015-8 (ebook)
1. Multimedia communications--Encyclopedias. I. Pagani, Margherita, 1971-
TK5105.15.E46 2009
621.38203--dc22
2008030766

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

If a library purchased a print copy of this publication, please go to http://www.igi-global.com/agreement for information on activating the library's complimentary electronic access to this publication.

Affective Computing

Maja Pantic
Imperial College London, UK
University of Twente, The Netherlands

INTRODUCTION

We seem to be entering an era of enhanced digital connectivity. Computers and Internet have become so embedded in the daily fabric of people's lives that people simply cannot live without them (Hoffman, Novak, & Venkatesh, 2004). We use this technology to work, to communicate, to shop, to seek out new information, and to entertain ourselves. With this ever-increasing diffusion of computers in society, human–computer interaction (HCI) is becoming increasingly essential to our daily lives.

HCI design was first dominated by direct manipulation and then delegation. The tacit assumption of both styles of interaction has been that the human will be explicit, unambiguous, and fully attentive while controlling the information and command flow. Boredom, preoccupation, and stress are unthinkable even though they are "very human" behaviors. This insensitivity of current HCI designs is fine for well-codified tasks. It works for making plane reservations, buying and selling stocks, and, as a matter of fact, almost everything we do with computers today. But this kind of categorical computing is inappropriate for design, debate, and deliberation. In fact, it is the major impediment to having flexible machines capable of adapting to their users and their level of attention, preferences, moods, and intentions.

The ability to detect and understand affective states of a person we are communicating with is the core of emotional intelligence. Emotional intelligence (EQ) is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life (Goleman, 1995). When it comes to computers, however, not all of them will need emotional intelligence and none will need all of the related skills that we need. Yet human–machine interactive systems capable of sensing stress, inattention, and heedfulness, and capable of adapting and responding appropriately to these affective states of the user are likely to be perceived as more natural, more efficacious, and more trustworthy. The research area of machine analysis of human affective states and employment of this information to build more natural, flexible (affective) HCI goes by a general name of affective computing, introduced first by Picard (1997).

RESEARCH MOTIVATION

Besides the research on natural, flexible HCI, various research areas and technologies would benefit from efforts to model human perception of affective feedback computationally. For instance, automatic recognition of human affective states is an important research topic for video surveillance as well. Automatic assessment of boredom, inattention, and stress will be highly valuable in situations where firm attention to a crucial, but perhaps tedious task is essential, such as aircraft control, air traffic control, nuclear power plant surveillance, or simply driving a ground vehicle like a truck, train, or car. An automated tool could provide prompts for better performance based on the sensed user's affective states.

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

Other areas where machine tools for analysis of human affective feedback could expand and enhance research and applications include specialized areas in professional and scientific sectors. Monitoring and interpreting affective behavioral cues are important to lawyers, police, and security agents who are often interested in issues concerning deception and attitude. Machine analysis of human affective states could be of considerable value in these situations where only informal interpretations are now used. It would also facile the research in areas such as behavioral science (in studies on emotion and cognition), anthropology (in studies on cross-cultural perception and production of

Table 1. The main problem areas in the research on affective computing

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

affective states), neurology (in studies on dependence between emotional abilities impairments and brain lesions), and psychiatry (in studies on schizophrenia) in which reliability, sensitivity, and precision are persisting problems. For a further discussion, see Pantic and Bartlett (2007) and Pantic, Pentland, Nijholt, and Huang (2007).

THE PROBLEM DOMAIN

While all agree that machine sensing and interpretation of human affective information would be quite beneficial for manifold research and application areas, addressing these problems is not an easy task. The main problem areas are listed in Table 1.

What is an affective state? Traditionally, the terms "affect" and "emotion" have been used synonymously. Following Darwin, discrete emotion theorists propose the existence of six or more basic emotions that are universally displayed and recognized (Lewis & Haviland-Jones, 2000). These include happiness, anger, sadness, surprise, disgust, and fear. In other words, nonverbal communicative signals (especially facial and vocal expression) involved in these basic emotions are displayed and recognized cross-culturally. In opposition to this view, Russell (1994) among others argues that emotion is best characterized in terms of a small number of latent dimensions (e.g., pleasant vs. unpleasant, strong vs. weak), rather than in terms of a small number of discrete emotion categories. Furthermore, social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations. They argue further that emotion is culturally constructed and no universals exist. Then there is lack of consensus on how affective displays should be labeled. For example, Fridlund (1997) argues that human facial expressions should not be labeled in terms of emotions but in terms of behavioral ecology interpretations, which explain the influence a certain expression has in a particular context. Thus, an "angry" face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. Yet, people still tend to use *anger* as the interpretation rather than *readiness-to-attack* interpretation. Another issue is that of culture dependency: the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent (Wierzbicka, 1993). Also, it is not only discrete emotional states like surprise or anger that are of importance for the realization of proactive human–machine interactive systems. Sensing and responding to behavioral cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are essential. However, there is even less consensus on these nonbasic affective states than there is on basic emotions. In summary, previous research literature pertaining to the nature and suitable representation of affective states provides no firm conclusions that could be safely presumed and adopted in studies on machine analysis of affective states and affective computing. Hence, we advocate that pragmatic choices (e.g., application- and user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback (Pantic & Rothkrantz, 2003).

Which human communicative signals convey information about affective state? Affective arousal modulates all verbal and nonverbal communicative signals (Ekman & Friesen, 1969). However, the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioral cues (Ambady & Rosenthal, 1992). 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. However, a large number of studies in psychology and linguistics confirm the correlation between some affective displays (especially basic emotions) and specific audio signals (e.g., Juslin & Scherer, 2005). Thus, automated human affect analyzers should at least include facial expression modality and preferably they should also include modalities for perceiving body gestures and tone of the voice.

How are various kinds of evidence to be combined to optimize inferences about affective states? Humans simultaneously employ the tightly coupled modalities of sight, sound, and touch. As a result, analysis of the perceived information is highly robust and flexible. Thus, in order to accomplish a multimodal analysis of human interactive signals acquired by multiple sensors, which resembles human processing of such information, input signals should not be considered mutually independent and should not be combined only at the end of the intended analysis as the majority of current studies do. Moreover, facial, bodily, and audio expressions of emotions should not be studied separately, as is often the case, since this precludes finding evidence of the temporal correlation between them. On the other hand, a growing body of research in cognitive sciences argues that the dynamics of human behavior are crucial for its interpretation (e.g., Ekman & Rosenberg, 2005). For example, it has been shown that temporal dynamics of facial behavior are a critical factor for distinction between spontaneous and posed facial behavior as well as for categorization of complex behaviors like pain (e.g., Pantic & Bartlett, 2007). However, when it comes to human affective feedback, temporal dynamics of each modality separately (visual and vocal) and temporal correlations between the two modalities are virtually unexplored areas of research. Another largely unexplored area of research is that of context dependency. One must know the context in which the observed interactive signals have been displayed (who the expresser is and what his current environment and task are) in order to interpret the perceived multisensory information correctly. In summary, an "ideal" automatic analyzer of human affective information should have the capabilities summarized in Table 2.

THE STATE OF THE ART

Because of the practical importance and the theoretical interest of cognitive scientists discussed above, automatic human affect analysis has attracted the interest of many researchers in the past three decades. The very first works in the field are those by Suwa, Sugie, and Fujimora (1978), who presented an early attempt to automatically analyze facial expressions, and by Williams and Stevens (1972), who reported the first study conducted on vocal emotion analysis. Since late 1990s, an increasing number of efforts toward automatic affect recognition were reported in the literature. Early efforts toward machine affect recognition from face images include those of Mase (1991) and Kobayashi and Hara (1991). Early efforts toward machine analy-

Table 2. The characteristics of an "ideal" automatic human-affect analyzer

- Multimodal (modalities: visual and audio; signals: facial, bodily,and vocal expressions)
- Robust and accurate (despite occlusions, changes in viewing and lighting conditions, and ambient noise, which occur often in naturalistic contexts)
- Generic (independent of physiognomy, sex, age, and ethnicity of the subject)
- Sensitive to the dynamics of displayed affective expressions (performing temporal analysis of the sensed data)
- Context-sensitive (realizing environment- and task-dependent data interpretation in terms of user-profiled affect-descriptive labels)

sis of basic emotions from vocal cues include studies like that of Dellaert, Polzin, and Waibel (1996). The study of Chen, Huang, Miyasato, and Nakatsu (1998) represents an early attempt toward audiovisual affect recognition. Currently, the existing body of literature in machine analysis of human affect is immense (Pantic et al., 2007; Pantic & Rothkrantz, 2003; Zeng, Pantic, Roisman, & Huang, 2007). 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 three to seven 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 perform analysis of temporal dynamics and correlations between different signals coming from one or more observation channels.
- 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; that is, sentences that are short, delimited by pauses, and carefully pronounced by nonsmoking actors.

Hence, while automatic detection of the six basic emotions in posed, controlled audio or visual displays can be done with reasonably high accuracy, detecting these expressions or any expression of human affective behavior in less constrained settings is still a very challenging problem due to the fact that deliberate behavior differs in visual appearance, audio profile, and timing, from spontaneously occurring behavior. Due to this criticism received from both cognitive and computer scientists, the focus of the research in the field started to shift to automatic analysis of spontaneously displayed affective behavior. Several studies have recently emerged on machine analysis of spontaneous facial and/or vocal expressions (Pantic et al., 2007; Zeng et al., 2007).

Also, it has been shown by several experimental studies that integrating the information from audio and video leads to an improved performance of affective behavior recognition. The improved reliability of audiovisual (multimodal) approaches in comparison to single-modal approaches can be explained as follows. Current techniques for detection and tracking of facial expressions are sensitive to head pose, clutter, and variations in lighting conditions, while current techniques for speech processing are sensitive to auditory noise. Audiovisual (multimodal) data fusion can make use of the complementary information from these two (or more) channels. In addition, many psychological studies have theoretically and empirically demonstrated the importance of integration of information from multiple modalities to yield a coherent representation and inference of emotions (e.g., Ambady & Rosenthal, 1992). As a result, an increased number of studies on audiovisual (multimodal) human affect recognition have emerged in recent years (Pantic et al., 2007; Zeng et al., 2007). Those include analysis of pain and frustration from naturalistic facial and vocal expressions (Pal, Iyer, & Yantorno, 2006), analysis of the level of interest in meetings from tone of voice, head and hand movements (Gatica-Perez, McCowan, Zhang, & Bengio, 2005), and analysis of posed vs. spontaneous smiles from facial expressions, head, and shoulder movements (Valstar, Gunes, & Pantic, 2007), to mention a few. However, most of these methods are context insensitive, do not perform analysis of temporal dynamics of the observed behavior, and are incapable of handling unconstrained environments correctly (e.g., sudden movements, occlusions, auditory noise).

CRITICAL ISSUES

The studies reviewed in the previous section indicate two new trends in the research on automatic human affect recognition: analysis of spontaneous affective behavior and multimodal analysis of human affective behavior including audiovisual analysis and multicue visual analysis based on facial expressions, head movements, and/or body gestures. Several previously recognized problems have been studied in depth including multimodal data fusion on both feature-level and decision-level. At the same time, several new challenging issues have been recognized, including the necessity of studying the temporal correlations between the different modalities (audio and visual) and between various behavioral cues (e.g., prosody, vocal outbursts like laughs, facial, head, and body gestures). Besides this critical issue, there are a number of scientific and technical challenges that are essential for advancing the state of the art in the field.

- **Fusion:** Although the problem of multimodal data fusion has been studied in great detail (Zeng et al., 2007), a number of issues require further investigation including the optimal level of integrating different streams, the optimal function for the integration, as well as inclusion of suitable estimations of reliability of each stream.
- **Fusion and context:** How to build context-dependent multimodal fusion is an open and highly relevant issue. Note that context-dependent fusion and discordance handling were never attempted.
- **Dynamics and 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 affective displays can be learned. Since the grammar of human behavior is context-dependent, should this be done in a user-centered manner or in an activity/application-centered manner?
- **Learning vs. education:** What are the relevant parameters in shown affective behavior that an anticipatory interface can use to support humans in their activities? How should this 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.
- **Robustness:** Most methods for human affect sensing and context sensing work only in (often highly) constrained environments. Noise, fast and sudden movements, changes in illumination, and so on, cause them to fail.
- **Speed:** Many of the methods in the field do not perform fast enough to support interactivity. Researchers usually choose more sophisticated processing rather than real time processing. A typical excuse is, according to Moore's Law, is that we will have faster hardware soon enough.

CONCLUSION

Multimodal context-sensitive (user-, task-, and application-profiled and affect-sensitive) HCI is likely to become the single most widespread research topic of artificial intelligence (AI) research community (Pantic et al., 2007; Picard, 1997). Breakthroughs in such HCI designs could bring about the most radical change in computing world; they could change not only how professionals practice computing, but also how mass consumers conceive and interact with the technology. However, many aspects of this "new generation" HCI technology, in particular ones concerned with the interpretation of human behavior at a deeper level and the provision of the appropriate response, are not mature yet and need many improvements.

REFERENCES

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

Chen, L., Huang, T.S., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition* (pp. 396–401).

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 1970–1973).

Ekman, P., & Friesen, W.F. (1969). The repertoire of nonverbal behavioral categories: Origins, usage, and coding. *Semiotica, 1*, 49–98.

Ekman, P., & Rosenberg, E.L., (Eds.). (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*. Oxford, UK: Oxford University Press.

Fridlund, A.J. (1997). The new ethology of human facial expression. In J.A. Russell & J.M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 103–129). Cambridge, MA: Cambridge University Press.

Gatica-Perez, D., McCowan, I., Zhang, D., & Bengio, S. (2005). Detecting group interest level in meetings. *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, 1*, 489–492.

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.

Hoffman, D.L., Novak, T.P., & Venkatesh, A. (2004). Has the Internet become indispensable? *Communications of the ACM, 47*(7), 37–42.

Juslin, P.N., & Scherer, K.R. (2005) Vocal expression of affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research*. Oxford, UK: Oxford University Press.

Kobayashi, H., & Hara, F. (1991). The recognition of basic expressions by neural network. In *Proceedings of the International Conference on Neural Networks* (pp. 460–466).

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

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions, E74*(10), 3474–3483.

Pal, P., Iyer, A.N., & Yantorno, R.E. (2006). Emotion detection from infant facial expressions and cries. *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, 2*, 721–724.

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

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. (2007). Human computing and machine understanding of human behavior: A survey. In *Proceedings of the ACM International Conference on Multimodal Interfaces* (pp. 239–248).

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390.

Picard, R.W. (1997). *Affective computing*. Cambridge, MA: The MIT Press.

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

Suwa, M., Sugie, N., & Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. In *Proceedings of the International Conference on Pattern Recognition* (pp. 408–410).

Valstar, M.F., Gunes, H., & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the International Conference on Multimodal Interfaces*.

Wierzbicka, A. (1993). Reading human faces. *Pragmatics and Cognition, 1*(1), 1–23.

Williams, C., & Stevens, K. (1972). Emotions and speech: Some acoustic correlates. *Journal of the Acoustic Society of America, 52*(4), 1238–1250.

Zeng, Z., Pantic, M., Roisman, G.I., & Huang, T.S. (2007). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In *Proceedings of the International Conference on Multimodal Interfaces*.

KEY TERMS

Affective Computing: The research area concerned with computing that relates to, arises from, or deliberately influences emotion. Affective computing expands HCI by including emotional communication together with appropriate means of handling affective information.

Anticipatory Interface: Software application that realizes human–computer interaction by means of understanding and proactively reacting (ideally, in a context-sensitive manner) to certain human behaviors such as moods and affective feedback.

Context-sensitive HCI: HCI in which the computer's context with respect to nearby humans (i.e., who the current user is, where he is, what his current task is, and how he feels) is automatically sensed, interpreted, and used to enable the computer to act or respond appropriately.

Emotional Intelligence: A facet of human intelligence that includes the ability to have, express, recognize, and regulate affective states, employ them for constructive purposes, and skillfully handle the affective arousal of others. The skills of emotional intelligence have been argued to be a better predictor than IQ for measuring aspects of success in life.

Human–Computer Interaction (HCI): The command and information flow that streams between the user and the computer. It is usually characterized in terms of speed, reliability, consistency, portability, naturalness, and users' subjective satisfaction.

Human–Computer Interface: A software application, a system that realizes human-computer interaction.

Multimodal (Natural) HCI: HCI in which command and information flow exchanges via multiple natural sensory modes of sight, sound, and touch. The user commands are issued by means of speech, hand gestures, gaze direction, facial expressions, and so forth, and the requested information or the computer's feedback is provided by means of animated characters and appropriate media.