Preface

This volume in the *Lecture Notes of Artificial Intelligence* represents the first book on human computing. We introduced the notion of human computing in 2006 and organized two events that were meant to explain this notion and the research conducted worldwide in the context of this notion.

The first of these events was a Special Session on Human Computing that took place during the Eighth International ACM Conference on Multimodal Interfaces (ICMI 2006), held in Banff, Canada, on November 3, 2006. The theme of the conference was multimodal collaboration and our Special Session on Human Computing was a natural extension of the discussion on this theme. We are grateful to the organizers of ICMI 2006 for supporting our efforts to organize this Special Session during the conference.

The second event in question was a Workshop on AI for Human Computing organized in conjunction with the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), held in Hyderabad (India), on January 6, 2007. The main theme of IJCAI 2007 was AI and its benefits to society. Our workshop presented a vision of the future of computing technology in which AI, in particular machine learning and agent technology, plays an essential role. We want to thank the organizers of IJCAI 2007 for their support in the organization of the Workshop on AI for Human Computing.

A large number of the contributions in this book are updated and extended versions of the papers presented during these two events. In order to obtain a more complete overview of research efforts in the field of human computing, a number of additional invited contributions are included in this book on AI for human computing.

One of the contributions in this volume starts with the observation that humans are social beings. Unfortunately, it is exceptional when we can say that a particular computer system, a computer application, or a human – computer interface has been designed from this point of view. Rather, we talk about users that have to perform tasks in a way that is prescribed by the computer. However, when we take the point of view of designing systems for social beings, we should talk rather about partners or participants instead of users, and when we do so, it is also the computer or a computer-supported environment that plays the role of a partner or a participant.

Human computing, as advocated and illustrated in this volume, aims at making computing devices and smart environments social partners of humans interacting with these devices or inhabiting these environments. These devices and environments need to understand what exactly the specifics of the current interaction flow and the surrounding environment are. This understanding allows for anticipatory and proactive feedback and real-time, unobtrusive support of human activities in the environment.

The *LNAI* volume on *AI for Human Computing* consists of three parts: a part on foundational issues of human computing, a part on sensing humans and their activities, and a part on anthropocentric interaction models.

Sensing humans and understanding their behavior is a core issue in the research on human computing. Numerous papers presented in this volume can be considered from this point of view. Human behavioral cues like facial expressions and body gestures are sensed by the computer / environment and then interpreted. Unfortunately, this interpretation is often carried out while taking into account only a limited portion of the information (if any) about the context (sensed or explicitly provided) in which the observed behavioral patterns have occurred. However, as accurate interpretation of human behavior cannot be achieved in a context-free manner, several papers in this volume argue that the realization of automatic context sensing and context-dependant analysis of human behavioral cues are the two challenging issues that need immediate attention of researchers in the field.

Sensing humans and interpreting their behavior should be followed by proactive support of their current activities. The support is to be provided by the environment and its 'inhabitants' in a suitable form. For example, the environment, as perceived by the user, can provide the support by turning on the air-conditioning when the user gets sweaty and his or her face reddens. On the other hand, human users being a part of a smart environment can be supported by artificial partners while conducting a task or a leisure-oriented activity within the environment. That is, smart digital devices within the environment including robots, virtual humans displayed within the environment, and hidden software agents of the environment can all provide support to human users within the environment and can cooperate with them and each other to provide this support. From a global point of view, realizing a smart environment able to support human activities simply means adapting the environment including its human and artificial inhabitants in such a way that the implicitly or explicitly communicated habits, preferences, desires, questions, and commands of a particular human user within the environment are anticipated, supported, and executed as well as possible.

Several papers in this book also emphasize the importance of having multimodal corpora available for research purposes. These are necessary for training and testing machine learning methods for human behavior analysis. In addition, analysis of a multimodal corpus makes us aware of properties of human behavior, correlations between behavioral signals and types (categories) of human behavior, and (causal) relations between the different categories and between the context specifics and these categories. This allows us to modify and extend existing theories (in computer science as well as in cognitive sciences) and also to refine existing algorithms to enable more accurate / suitable analysis of the data. In a research context it is not always possible to study or fully simulate realistic situations. Financial or ethical reasons will often make it impossible to have fully unobtrusive recordings of human behavior in naturalistic contexts. Hence, the large majority of currently conducted experiments relate to scripted behavior, often deliberately displayed behavior that follows specific scenarios designed by researchers. Behavioral data obtained in this way can be useful in the start-up phase of the research on human computing. However, such data and the methods trained and tested using such data are not applicable in real-life situations, where subtle changes in expression of behavior are typical rather than the exaggerated changes that typify deliberately displayed behavior. Hence, the focus of the research in the field started to shift to automatic analysis of spontaneous behavior (produced in a reflex-like manner). Several works presented in this volume address machine analysis of human spontaneous behavior and present efforts towards collecting data in realistic settings.

In what follows, we shortly summarize each of the contributions to this volume.

Foundations of Human Computing

In his contribution to this volume, Cohn surveys ways to infer emotions from expressive human behavior, in particular facial expressions. He discusses several key issues that need to be considered when designing interfaces that approach the naturalness of human face-to-face interaction. Among them are the differences between the judgment-based approach (inferring the underlying affective state) and the sign-based approach (labeling facial muscle actions) to measurement of facial behavior, the timing and overall dynamics of facial behavior, and individual differences in facial expressions. A less researched issue discussed in this paper, which is of particular importance for natural interaction, is that of synchrony of affect display in face-to-face interaction.

"Instinctive Computing" by Cai argues that for genuine intelligence and natural interaction with humans, computers must have the ability to recognize, understand, and even have primitive instincts. The message, supported by the literature, is that instincts influence how we look, feel, think, and act. Foraging, vigilance, reproduction, intuition, and learning are distinguished as the human basic instincts. These basic instincts need to be addressed and have their metaphors in the computer systems of the future. The paper discusses different case studies and observations on nontraditional physiological sensors, sensors related to reproductive esthetics, and those related to updating of instincts (learning). Instinctive computing is seen as the foundation for ambient intelligence and empathetic computing, where the latter includes the detection, understanding, and reacting to human expressions of pain, illness, depression, and anomaly.

Sensing Humans for Human Computing

Automatic sensing and understanding of human behavior in computer-supported and smart environments are discussed by the editors of this volume in their position paper on human computing. They summarize the current state of the art, challenges, and opportunities facing the researchers in intertwined research areas of context sensing, human affect sensing, and social signaling analysis. Context sensing is discussed from the W5+ (Who, Where, What, When, Why, How) perspective and it is argued that the most promising way to achieve accurate context sensing in naturalistic settings is to realize multimodal, multi-aspect context sensing. In this approach, the key is to automatically determine whether observed behavioral cues share a common cause [e.g., whether the mouth movements and audio signals complement to indicate an active known or unknown speaker (How, Who, Where) and whether his or her focus of attention is another person or a computer (What, Why)]. In addition, the paper also argues that the problem of context-constrained analysis of multimodal behavioral signals shown in temporal intervals of arbitrary length should be treated as one complex problem rather than a number of detached problems in human sensing, context sensing, and human behavior understanding.

Natural settings, in which we can infer the users' emotional state by combining information from facial expression and speech prosody, are discussed in several contributions to this volume. As remarked in the relevant papers, it is well known that posed (deliberately displayed) expressions of emotions differ in appearance and timing from those shown in a natural setting. Audiovisual recognition of spontaneous expressions of human positive and negative affective feedback in a natural human – human conversation setting has been investigated in the work of Zeng et al. Facial expressions extracted from the video signal and prosodic features (pitch and energy) extracted from the speech signal were treated separately and in a combination. It was shown that bimodal data provide more effective information for human affect recognition than single-modal data and that linear bimodal fusion is sub-optimal for the task at hand.

The paper of Karpouzis et al. discusses audiovisual emotion recognition (positive vs. negative and active vs. passive) in naturalistic data obtained from users interacting with an artificial empathetic agent. This sensitive artificial listener (SAL) agent gives the impression of sympathetic understanding. The 'understanding' component is keyword driven but suffices to induce states that are genuinely emotional and involve speech. The SAL agent can be given different personalities to elicit a range of emotions. Facial expressions and hand gestures extracted from the video signal and a range of acoustic features extracted from the speech signal were treated subsequently by means of recurrent neural networks trained for recognition of target affective states. Similar to Zeng et al., Karpouzis et al. show that multimodal data provide more effective information for human affect recognition than single-modal data.

Human – robot interaction in settings where both the human and the robot can show affective behavior has been investigated in the work of Broekens. Part of the presented setup (i.e., the robot) has been simulated. The aim of the simulated robot is to survive in a grid-world environment. It uses reinforcement learning to find its way to food. Two versions of the robot are presented: a nonsocial robot that learns its task without the affective feedback of a human observer and a social robot that uses the communicated affective feedback as the social reward. Affective feedback is obtained from the facial expressions of the human observer who monitors the robot's actions. The results show that the "social robot" learns its task significantly faster than its "nonsocial sibling" and Broekens argues that this presents strong evidence that affective communication with humans can significantly enhance the reinforcement learning loop.

The paper of Oikonomopoulos et al. presents the research on trajectory-based representation of human actions. In this work, human actions are represented as collections of short trajectories that are extracted by means of a particle filtering tracking scheme that is initialized at points that are considered salient in space and time. Improving the detection of spatio-temporal salient points and enhancing the utilized tracking scheme by means of an online background estimation algorithm is also discussed. By defining a distance metric between different sets of trajectories corresponding to different actions using a variant of the longest common subsequence algorithm and a relevance vector machine, the authors obtained promising results for recognition of human body gestures such as aerobic exercises.

Modeling the communication atmosphere is the main topic of the paper by Rutkowski and Mandic. They identify a 3D communication space for face-to-face communication, where the dimensions are environmental (i.e., related to the ambient conditions like noise and visual activity rather than to the communication itself), communicative (related to the audiovisual behavior of the communicators like feedback provision and turn taking), and emotional (related to coupled emotional states of the participants). The main focus of the paper is on the dynamics of nonverbal communication as the basis for estimating communication atmosphere. The ability to model this aspect of face-to-face communication implies the ability to manipulate (some of) the relevant conditions in order to adjust the atmosphere. Experiments using cameras and microphones to capture communicators' face-to-face situations have been conducted.

Dong and Pentland discuss the problem of combining evidence from different dynamic processes. For example, evidence about the context of a particular user can be obtained from different sensors (possibly connected in a sensor network). The proposed approach to multisensorial data fusion and interpretation introduces an 'influence model' in which experts with different knowledge can consult each other about their understanding of the data in order to decide about the classification. The authors illustrate their approach by means of two examples. The first one is a situation where data are collected from several wearable sensors (accelerometers, an audio recorder, and a video recorder) and where a team of four experts need to recognize various types of wearer context (locations, audio contexts, postures, and activities). The second situation relates to a social network where participants have mobile phones that record various data based on which the participants' social circles and individual behaviors are to be determined.

Anthropocentric Interaction Models for Human Computing

Humans have social skills. These skills allow them to manage relationships with other people in a given social structure. In ambient intelligence and virtual community environments, we need models of social intelligence in order to understand, evoke, and anticipate social behavior and to generate social intelligent behavior by the environment and its physical and synthetic agents performing in the environment. Here, generating socially intelligent behavior means performing actions, providing feedback, taking the initiative in interactions, displaying verbal and nonverbal affective and social signals, and amplifying social intelligence in such a way that there is a smooth, natural, but also effective embedding in the socio-cultural structure of the virtual, human, or augmented-reality community. Social intelligence design, as advocated by Nishida, aims at understanding and augmentation of social intelligence. Three perspectives are distinguished. The first one relates to social interaction in face-to-face and multi-party interactive settings in small groups where traditional discourse modeling and verbal and nonverbal interaction issues play important roles in displaying socially intelligent interactive behavior. The second perspective relates to social interaction in the large, possibly multimedial, chat and game environments, where sociological and socio-psychological models of multi-party interaction, large-scale collaboration, and social attitudes are of importance. The third perspective relates to the design of social artifacts that embody social intelligence and therefore facilitate social interaction. Among these artifacts are embodied agents, interactive robots, as well as collaboration technologies.

Within the human computing framework one can argue that interactive systems need to have the same communicative capabilities that humans have. In other words, human – computing technologies (i.e., interactive systems or environments) need to be based on theories and models of human – human interaction. The paper by Op den Akker and Heylen explores this view on human computing. Presently, computational models of human – human interaction are very limited and hardly take into account subtleties of verbal, let alone nonverbal, interaction. Nevertheless, there are tools such as annotation schemes that enable the researchers in the field to analyze multimodal interaction corpora and come up with new, enhanced models of human - human interplay. In the paper by Op den Akker and Heylen these issues are discussed and illustrated by analyzing conversations constituting the multimodal augmented multi-party interaction (AMI) corpus.

Human computing applications are based on displayed human behavior such as affective and social signals. Being able to understand human behavior and behavioral cues is far beyond traditional human – computer interaction research. Moreover, within human computing applications in the home, office, and public spaces, human behavior to be supported by the environment does not necessarily have to be task-oriented and efficiency may not be an issue at all. For example, how should a smart environment support leisure activities, how can it increase social interaction among inhabitants, and how can it act as a social and entertaining space for its inhabitants? Clearly, designing the environment as a social and entertaining partner is an open and complex research question where various perspectives on environmental characteristics such as efficiency, related to computing and human resources, should be considered. Poppe et al. discuss trends for emerging human computing applications. They identify where challenges remain when it comes to evaluating emerging applications of human computing technologies. The paper aims to create awareness that business-as-usual will not work for these applications, it stresses the fact that the current evaluation practices are inappropriate, and it proposes partial solutions to some of the challenges.

Maat and Pantic discuss an agent-based, adaptive interface in which context sensing and context modeling are integrated within a context-aware application. The human part of the user's context is sensed using face recognition, eye tracking, speech recognition and, as usual, key strokes and mouse movements. Contextual questions related to who the current user is, what his or her task is, how he or she feels, and when a certain user's (re)action occurred are answered in an automated manner. Case-based reasoning is used to match events against user preferences. Important parts of the paper are the reasoning about the user's preferences, the interaction adaptation in accordance to these preferences, and the conducted usability study. In the latter an assessment of the system's usability was made in terms of issues like effectiveness, usability, usefulness, affective quality, and a number of ethical issues relevant to the system.

Sonntag et al. present a study on an environment where the user employs a context-aware handheld device with which it can interact in a natural (multimodal) way using speech and gestures. More specifically, Sonntag et al. investigated development of a context-aware, multimodal mobile interface to Web-based information services. User expressions (i.e., his or her speech and pointing gestures) are processed with respect to the situational and discourse context. Question clarification, resolution of referring expressions, and resolution of elliptical expressions are handled by the dialogue manager. A media ontology guides the presentation of results to the user, using text, image, and speech as possible modalities.

We started this volume with the paper by Cohn on the human face, the recognition of facial expressions, and how to use them in helping to infer an underlying affective state. The paper of Blanz revisits the topic of the human face. Blanz presents a model-based approach to reconstruct, animate, modify, and exchange faces in images or in

3D. Human computing addresses computer-supported environments inhabited by humans and human-like agents. Faces, displaying affective and social signals, need to be understood, both for understanding and for generating purposes. Autonomous and human-controlled agents need faces, for example, to represent their ‘owners.’ We also need tools to generate and exchange faces (for example, by reconstructing them from images), to manipulate them, and to do facial animation, for example, to generate subtle facial expressions and visual speech. The model-based tools proposed by Blanz allow semantically meaningful manipulation of faces.

Humans and human-like agents (robots, virtual humans) can interact with each other in smart-, virtual-, and mixed-reality environments. While Op den Akker and Heylen take human – human conversations as the starting point for modeling natural interactive systems, Reidsma et al. require not only that the interactive system or environment interacts with its human partner in a natural and intuitive way, but that it does so by means of a human-like, embodied, virtual human. This virtual human may represent a particular functionality of the environment (e.g., a doorman, a butler, a financial adviser, a fitness trainer, or a virtual friend) or it may represent another user of the environment with which it is possible to interact. The paper emphasizes the necessity of being able to model all kinds of subtleties and ‘imperfections’ in human – human communication. For example, designing a virtual human that is reluctant to answer the user’s questions may seem a waste of time, but if this virtual human is a virtual tutor in an educational virtual environment, or a negotiation partner within a virtual auction, or an opponent in a game environment, then this is rather acceptable behavior. The paper also discusses applications and (individual) user characteristics that require modeling of preferences and peculiarities of humans and human – human interactions.

Virtual humans, but controlled by human actors, are the topic of the research presented by Zhang et al. The ‘avatars’ discussed in the paper represent human actors that have mediated interaction in a storytelling environment. One of the avatars in the environment is controlled by a human director who takes care through her avatar that the story develops appropriately. More specifically, the affective cues present in the natural language interaction of the avatars help the director to make the emerging story interesting. The aim of Zhang et al. is to automate the role of the director. This requires the detection of the affective content of the utterances. One of the issues that are researched is the metaphorical conveyance of affect. Rather than aiming at fully linguistic analyses of the utterances mediated by the avatars, Zhang et al. aim at building robust representations of the affective connotations. One obvious reason to do so is that the utterances are often ungrammatical, they borrow from language used in text-messaging and chat rooms, and often the literal meaning hardly gives a clue about the intended meaning. Obviously, being able to detect the affective state of a user-controlled virtual human makes it also possible to generate a suitable emotional animation.

March 2007

Tom Huang
Anton Nijholt
Maja Pantic
Sandy Pentland